# 1

# Facial Landmark Detection

**Romain BELMONTE[1], Pierre TIRILLY[1], Ioan Marius BILASCO[1], Nacim IHADDADENE[2] and Chaabane DJERABA[1]**

[1] *University of Lille, France*

[2] *Junia ISEN, Lille, France*



As stated by Jin and Tan (2017), the face corresponds to a deformable object that can vary in terms of shape and appearance. The first attempts at facial landmark detection can be traced back to the 1990s. The active shape models (ASMs) (Cootes et al. 1995) represent one of the seminal works on the subject. Such a generative approach consists of a parametric model that can be fitted to a given face by optimizing its parameters. This process of applying the model to new data is called inference. Rapid progress has been made through the development of active appearance models (AAMs) (Cootes et al. 2001), constrained local models (CLMs) (Cristinacce and Cootes 2006) and other extensions (Baltrusaitis et al. 2013; Belhumeur et al. 2013; Antonakos et al. 2015; Tzimiropoulos 2015), to the point where the problem is now considered well addressed for constrained faces (Jin and Tan 2017). The research has therefore shifted to unconstrained faces with multiple and complex challenges, for example, occlusion, variations in pose, illumination and expression. Faster and more robust discriminative methods such as cascaded shape regression (CSR) (Dollár et al. 2010) have been proposed to address these challenges. They differ from generative approaches as they directly learn a mapping function between images and facial shapes with better generalization ability. Due in particular to the phenomenon of data massification and the increase in the computational capacity of machines, a subset of these approaches, deep learning (DL), has stood out. These approaches need larger datasets than traditional ML algorithms but do not require feature engineering. Above all, they currently outperform all previous approaches and most research now focuses on them.

*Face Analysis Under Uncontrolled Conditions*,
coordinated by Romain BELMONTE and Benjamin ALLAERT.

Consequently, two main categories of approaches can be defined, generative and discriminative, not to mention the complementary solutions that can be applied to most of these approaches, for example, multi-task learning, to explicitly address some selected challenges. The proposed nomenclature is close to those proposed in (Jin and Tan 2017; Wu and Ji 2018). Considering its wide range of applications and the persistent difficulties encountered under uncontrolled conditions, facial landmark localization remains a very active area of research. Recently, there has been a trend towards video-based solutions (Shen et al. 2015). One of the main reasons to include this additional dimension to the problem is that temporal consistency provides a useful input to achieve robust detection under uncontrolled conditions. Different strategies have been proposed, from the most traditional tracking strategies, such as tracking by detection, to more complex strategies that can jointly extract features and perform tracking. The current work is mostly tracking-oriented and focuses on global head movements.

All these items are discussed in detail in the following sections. In section 1.1, the groundbreaking work and advances in the two main categories of approaches are reviewed, with an emphasis on DL approaches. This allows us to properly describe, in section 1.1.4, the complementary solutions that can be used to handle major challenges. In contrast to existing literature reviews (Jin and Tan 2017; Wu and Ji 2018), the different strategies proposed in the literature to extend facial landmark localization to video are extensively discussed in section 1.2. Finally, section 1.3 concludes with a positioning of the work presented in this part of the book.

## 1.1. Facial landmark detection in still images

Efforts to tackle the problem of facial landmark detection have long focused on still images, and much work has been published. In sections 1.1.1 and 1.1.2, the two main categories of approaches, generative and discriminative, are detailed along with their developments. Among the discriminative approaches, DL approaches have become very popular. Therefore, the latter is given close attention in section 1.1.3. Finally, complementary solutions to handle the difficulties encountered under uncontrolled conditions are reviewed in section 1.1.4.

### 1.1.1. *Generative approaches*

Generative methods typically build a statistical model for both shape and appearance. They are referred to as generative approaches since they provide a model of the joint probability distribution of the shape and appearance. These models are parametric and therefore have some degrees of freedom constrained during training; so, they are expected to match only possible facial configurations. By optimizing the model parameters, for example, by minimizing the reconstruction error for a given

image, the best possible instance of the model for that image can be generated. Initially, only shape variations were modeled (Cootes et al. 1995). However, on the same line of thought, texture variations were also added to jointly apply constraints to variations in shape and texture (Cootes et al. 2001). Facial appearance can be represented in different ways. The entire face can be considered, which constitutes a holistic representation. The face can also be decomposed into different parts, for example, patches centered at each landmark, which is known as a part-based representation. Generative approaches provide a good fitting accuracy with little training data but are sensitive to shape initialization and do not generalize well to new images. The optimization can be difficult under uncontrolled conditions due to the high dimensionality of the appearance space. They tend to get trapped in local minima as well. Note also that these methods are mainly based on principal component analysis (PCA) (Pearson 1901) and therefore the distribution is assumed to be Gaussian, which does not fit the true distribution.

The AAMs (Cootes et al. 2001) are probably the most representative method in this category and the most widely studied. In the following, AAM modeling and fitting are presented as well as some of their improvements. The modeling can be split into three parts: a shape model, an appearance model and a motion model. The shape model, also called point distribution model (PDM) (Cootes et al. 1995), is common among deformable models (Cootes et al. 1995, 2001; Cristinacce and Cootes 2006; Baltrusaitis et al. 2013; Tzimiropoulos 2015). It is built using the $N$ training facial shapes. These shapes are first normalized (aligned) using a GPA (Gower 1975) to remove affine transformations (i.e. rotation, scale and translation variations) and keep only local, non-rigid shape deformation. Next, PCA is applied to obtain a set of orthogonal bases that capture the maximal variance. Only the $n$ eigenvectors corresponding to the largest eigenvalues are kept as they summarize well the data. This reduces the dimensions while ensuring that the key information is maintained. The model can then be expressed as:

$$s_p = \overline{s} + U_s p \qquad [1.1]$$

where $\overline{s} \in \mathbb{R}^{2L,1}$, $U_s \in \mathbb{R}^{2L,n}$ and $p \in \mathbb{R}^n$ are, respectively, the mean shape, the shape eigenvectors and the vector of shape parameters. Four eigenvectors corresponding to the similarity transforms (scaling, in-plane rotation and translation) are added to $U_s$ (re-orthonormalization) to be able to fit the model on any image (Matthews and Baker 2004).

To build the appearance model, features from the $N$ training images are first extracted using the function $F$. Initially, holistic pixel-based representation was used, which is sensitive to lighting and occlusions (Cootes et al. 2001). To improve robustness, feature-based representations such as histogram of oriented gradients (HOG) (Dalal and Triggs 2005) or scale-invariant feature transform (SIFT) (Lowe 2004) can also be used. In this way, relevant facial features are extracted, with a better

ability to generalize to new faces. These features are then warped into the reference shape using the motion model. PCA is finally applied onto these vectorized warped feature-based images. The model can be expressed as:

$$a_q = \overline{a} + U_a q \quad [1.2]$$

where $\overline{a} \in \mathbb{R}^{M,1}$, $U_a \in \mathbb{R}^{M,m}$ and $q \in \mathbb{R}^m$ are, respectively, the mean appearance vector, the appearance eigenvectors and the vector of appearance parameters.

The motion model refers typically to a warp function $W$ such as a piece-wise affine warp or a thin-plate spline (Matthews and Baker 2004; Tzimiropoulos and Pantic 2013). Given a shape generated from parameters $p$, this function defines how to warp the texture into a reference shape, for example, the mean shape $\overline{s}$. Figure 1.1 shows an example of AAM instantiation.

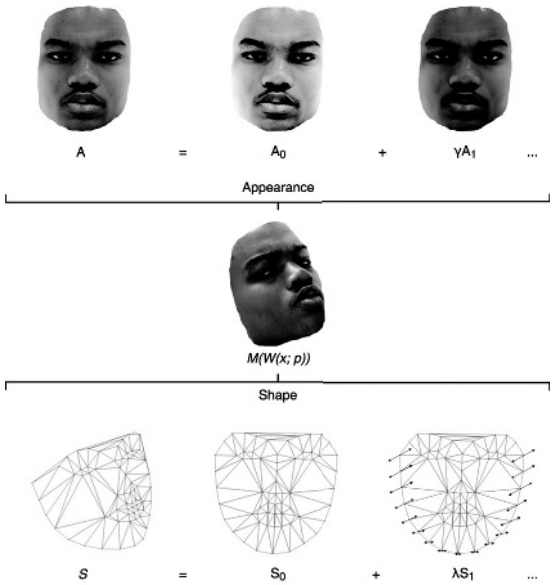


**Figure 1.1.** *Example of AAM instantiation. The appearance computed from the parameters q is warped into the shape computed from the parameters p (Matthews and Baker 2004). For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Finally, the AMMs can be fit to a previously unseen image by optimizing the following cost function, which consists of the minimization of the appearance reconstruction error with respect to the shape parameters:

$$\arg\min_{p,q} \; ||F(I)(W(p)) - \overline{a} - U_a q||^2 \quad [1.3]$$

The optimization is an iterative process by which parameters $p$ and $q$ are found so that the best instance of the model is fitted to the given image. This can be solved by two main approaches: analytically or by learning. The first one typically uses gradient descent optimization algorithms (Matthews and Baker 2004; Gross et al. 2005; Papandreou and Maragos 2008). The standard gradient descent being inefficient for this problem, other ways to update the parameters have been proposed. In the project-out inverse compositional algorithm (Matthews and Baker 2004), shape and appearance are decoupled by projecting out the appearance variations. This results in a faster fitting as well as with convergence issues that decrease robustness. With the simultaneous inverse compositional algorithm (Gross et al. 2005), shape and appearance parameters are optimized simultaneously to provide a more robust fitting but at a higher computational cost. It is, however, possible to speed it up using the alternating inverse compositional algorithm (Papandreou and Maragos 2008), which optimizes the shape and appearance parameters in an alternative manner instead of simultaneously. The second optimization approach typically uses cascaded regression to learn a mapping function between the facial appearance and the shape parameters (Cootes et al. 2001; Saragih and Goecke 2007). The original AAMs (Cootes et al. 2001) employ linear regression. Nonlinear regression has also been proposed (Saragih and Goecke 2007). This approach may be efficient but makes the invalid assumption that there is a constant linear relationship between the image features and the parameter updates (Matthews and Baker 2004). Despite various efforts in optimization strategies, the AAMs remain difficult to optimize under uncontrolled conditions due to the high dimensionality of the appearance space and the propensity of the optimizer to converge to local minima.

To overcome these problems, more robust, part-based representations can be used. In part-based generative models, the local appearance around each landmark is extracted and combined to build a model for the whole face (Tzimiropoulos and Pantic 2014; Tzimiropoulos 2015), which reduces the size of the appearance space. Active pictorial structures (Antonakos et al. 2015) go further by taking advantage of the tree structure used in pictorial structures (PSs) (Felzenszwalb and Huttenlocher 2005) to replace PCA. PSs are extended to model the appearance of the face using multiple graph-based pairwise distributions between the facial parts and prove to be more accurate than PCA. Note that ASMs are also considered as a part-based generative model, with the difference that an appearance model is used for each facial part rather than a single model for all parts. However, its evolution towards models such as CLMs (Cristinacce and Cootes 2006; Saragih et al. 2011), considered as discriminative, will be discussed in the next section. Although holistic and part-based models appear to have the same representational power, part-based models are easier to optimize and more robust to poor initializations, lighting and occlusions since local features are usually not as sensitive as global features.

### 1.1.2. *Discriminative approaches*

Unlike generative approaches that rely on statistical parametric models of appearance and shape, discriminative approaches learn a mapping from the image to the facial shape. In this section, two categories of approaches are distinguished. The first category refers to hybrid approaches such as CLMs (Cristinacce and Cootes 2006; Saragih et al. 2011). Such approaches are based on independent discriminative models of appearance, one for each facial part, constrained by a statistical parametric shape model, for example, PDM. This means that there is still an optimization step during inference for this category. Besides, the ambiguity between the local appearance models of different landmarks can strongly impact the performance under uncontrolled conditions. To address these difficulties, a second category of approaches has emerged. These approaches, the most notable of which is CSR (Dollár et al. 2010; Xiong and De la Torre 2013), infer the whole face shape by directly learning one or more regression functions which implicitly encode the shape constraint. It provides more freedom than a parametric shape model would as no explicit assumption on the distribution of the data is made. During inference, there is no optimization as well. The facial shape is directly estimated from the image features. Because of these distinctions, discriminative approaches are faster and more robust in comparison to generative ones. They generalize better to new unseen images as they can benefit from large datasets, which are omnipresent nowadays. However, it remains challenging to directly map the appearance to the facial shape when confronted to severe difficulties such as extreme poses or expressions. A third category could be added for DL approaches, for example, CNNs. Most of the work today is based on DNNs, leaving traditional methods behind[1]. These approaches are capable of learning highly discriminative and task-specific features. They no longer need feature engineering. Given the breakthrough of DL in computer vision, a separate section (section 1.1.3) is dedicated to these approaches.

#### 1.1.2.1. *Hybrid approaches*

CLMs are an extension of ASMs, with the main difference that the appearance models are discriminative rather than generative. It is a part-based hybrid approach composed of two elements: discriminative local detectors or regressors to represent the appearance and a generative shape model to regularize the deformations and ensure a valid face. In the following, the modeling, fitting and improvements related to this approach are described. Local detectors or regressors are used to compute response maps providing the probability that a given landmark is located at a specific position. To build them for each landmark, a statistical model of the gray level structure along with the Mahalanobis distance as the response was initially proposed in ASMs (Cootes et al. 1995). However, more powerful discriminative approaches – for example, binary

1. See https://jponttuset.cat/dl-lstm-gan-evolution/ and https://medium.com/@karpathy/a-peek-at-trends-in-machine-learning-ab8a1085a106.

classifiers such as logistic regression (Saragih et al. 2011) or support vector machines (SVMs) (Lucey et al. 2009) – have been used. Consider a linear SVM that determines whether or not a given patch matches the description of the region of a landmark. From the training data, positive and negative examples are extracted for each landmark to train different SVMs, also called patch experts. The response can be expressed as follows:

$$R = -I(c + \delta c)^T \sum_{v=1}^{V} \alpha_v \lambda_v(c) \qquad [1.4]$$

where $I$ is the given image, $c$ is the initial coordinates, $\delta c$ is a displacement constrained by the PDM, $\lambda_v(c)$ is the $V$ support vectors and $\alpha_v$ is the support weights. The output corresponds to the inverted classifier score, which is 1 if positive, –1 otherwise. By fitting a logistic regression function to the output, Platt scaling (Platt 1999), an approximate probabilistic output can be obtained. To refine the detection and maintain a valid shape, a shape constraint is then imposed using a statistical shape model such as PDM, already presented in section 1.1.1.

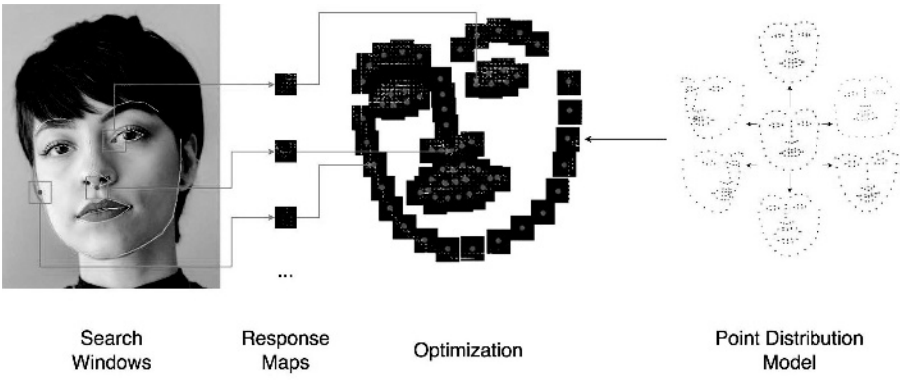


**Figure 1.2.** *Overview of CLM fitting: ELS is generally performed to get a response map for each landmark. To refine the detection and maintain a valid shape, a shape constraint is then imposed using a statistical shape model (Saragih et al. 2011). For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Hence, two steps (illustrated in Figure 1.2) can be differentiated when fitting CLM. An exhaustive local search (ELS) (Saragih et al. 2011) is generally performed first to get a response map for each landmark. This step is followed by a shape refinement over these response maps to find the shape parameters that maximize the probability

that the landmarks are accurately detected given the appearance features. The cost function can be expressed as:

$$\arg\min_{p} \sum_{l=1}^{L} R_l(I(\overline{s}_l + U_l p)) \quad [1.5]$$

where $I$ is the given image, $L$ is the number of landmarks, $\overline{s}_l$ is the coordinate of the $l$-th landmark from the mean shape, $U_l$ is the shape eigenvectors, $p$ is the shape parameters, and $R$ is the response. The Gauss–Newton algorithm, although it suffers from local minima, is generally employed to solve this problem (Saragih et al. 2011). Also, the true response maps are not directly used due to performance issues; they are replaced by approximations. There are several approximation techniques, most of them parametric, yet a non-parametric representation known as regularized landmark mean-shift has proved to be a good balance between representational power and computational complexity (Saragih et al. 2011). It takes the form of a Gaussian kernel density estimate (Silverman 2018):

$$KDE = \sum_{y_i \in \psi_i} \pi_{y_i} \mathcal{N}(x_i; y_i, pI) \quad [1.6]$$

where $y_i$ is a candidate location among $\psi_i$ locations within a given region, $\pi_{y_i}$ is the likelihood that the $i$-th landmark is aligned at location $y_i$, $x_i$ is the location of the $i$-th landmark generated by the shape model, $p$ is the variance of the noise on landmark locations and $I$ is the identity matrix. A regression-based fitting approach can also be used to learn mapping functions from response maps to the shape parameter updates. It is known as the discriminative response map fitting (Asthana et al. 2013). Given its nature and ability to benefit from large amounts of data, this approach achieves a performance gain over RLMS fitting.

One of the limitations of part-based approaches is the ambiguity between local detectors due to the small size and large appearance variations of the training patches. Feature descriptors can be used to obtain a more robust appearance representation. As an example, HOG has proved to be effective (Asthana et al. 2013). To further reduce ambiguity, more reliable patch experts have been introduced. They are capable of learning nonlinear and spatial relationships between pixels and responses, for example, the minimum output sum of squared errors (Bolme et al. 2010; Martins et al. 2014) or the local neural field (Baltrusaitis et al. 2013). A second limitation of part-based approaches may be the ELS, which can be computationally expensive. However, it can be avoided by computing response maps using regression voting (Cristinacce and Cootes 2007; Cootes et al. 2012). A third limitation is related to the use of PDM as a shape model, which has restricted degrees of freedom due to PCA. Other shape models have been proposed, from separate statistical shape models for each facial component to preserve local deformations (Huang et al. 2007)

to discriminative shape models based on restricted Boltzmann machines (Wu and Ji 2015a). The most distinctive one is the exemplar-based shape model (Belhumeur et al. 2013; Jin and Tan 2016) which, implicitly from training data, imposes shape constraints using a consensus of non-parametric models. It has the advantage of being independent of the initialization. It should be noted that there are some other variants close to the CLMs but generally considered as independent. This includes the combination of local regressors based on boosted support vector regression with graph models such as Markov random fields (Valstar et al. 2010; Martinez et al. 2013), or the use of tree-structured models (Zhu and Ramanan 2012; Uřičář et al. 2012; Hsu et al. 2015). However, globally optimizing Markov random fields appears to be difficult (Valstar et al. 2010). Tree-based models are easier to optimize thanks to dynamic programming but struggle to perform in real time (Felzenszwalb and Huttenlocher 2005; Zhu and Ramanan 2012).

#### 1.1.2.2. *Regression-based approaches*

Direct regression (DR) learns the mapping from the image appearance to the facial shape in one iteration without any initialization nor parametric appearance or shape models. The landmark locations are jointly estimated while the shape constraints are implicitly encoded. The model can be expressed as follows:

$$M : F(I) \mapsto s \qquad [1.7]$$

where $M$ is the model, $F(I)$ is a function that extract appearance features from an image $I$ and $s \in \mathbb{R}^{2L,1}$ is the predicted facial shape. The regression function being central, its choice is decisive. Regression forests (Breiman 2001) are generally used to cast votes for the facial shape from randomly sampled local patches (Dantone et al. 2012; Yang and Patras 2013a, 2013b). However, the random strategy to sample patches is sensitive to occlusions (Yang and Patras 2013b). Additionally, the information provided by local patches is not sufficient to perform global shape estimation. More recently, holistic approaches performed DR using global facial images. In particular, DL approaches (Sun et al. 2013) are discussed in section 1.1.3. This mapping is more difficult to learn due to the significant variations in appearance that can occur in the global facial image.

Since DR is challenging, CSR has been proposed to divide the regression process into stages, as shown in Figure 1.3. It usually starts with an initial shape, typically the mean shape of the training data. The early stages focus on large variations, for example, yaw, roll or scaling, while the later ones focus on subtle variations, for example, face contours and motions of the mouth, nose and eyes, to sequentially update and refine the prediction, following a coarse-to-fine strategy. Each stage corresponds to an independent regressor, which learns the mapping from the shape-indexed features to the shape update. Shape-indexed features correspond to local features extracted based on the current landmark location estimates.

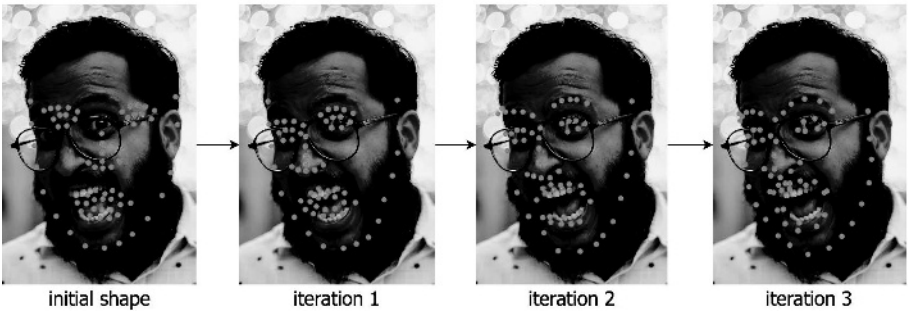


**Figure 1.3.** *Coarse-to-fine prediction using CSR. The regression process is divided into stages. The early stages focus on large variations, while the later ones focus on subtle variations, to sequentially update and refine the prediction. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

A shape increment can be expressed as follows:

$$\delta\hat{s}_t = r_t(\phi_t(I, s_{t-1})) \quad [1.8]$$

where $\delta s_t$ is the shape increment at stage $t$, $r_t$ is the stage regressor and $\phi_t$ is a function that extracts shape-indexed features from image $I$ using the landmark locations at stage $t-1$. The shape constraints are implicitly and adaptively encoded through the sequential training of the stage regressors $r_t$. During inference, the initial shape is refined by estimating shape increments $\delta s$ to reduce errors. This results in fast and accurate predictions.

Explicit shape regression (Cao et al. 2014) represents one seminal work on CSR, using random ferns as stage regressors and pixel intensity differences as shape-indexed features. To train such a model, the mean shape of the training data is first calculated and normalized, i.e. rescaled and centered at the origin. Then, each stage regressor is trained using the ground truth transformed into the mean shape coordinates through GPA (Gower 1975). This ensures invariance in scale. The objective is to minimize the mean-squared error (MSE) between the target and the estimate shape increment. It can be expressed as follows:

$$\arg\min_{r} \sum_{n=1}^{N} ||\delta s_{t,n} - r_t(I_n, s_{t-1,n})||^2 \quad [1.9]$$

where $r$ is the stage regressor, $\delta s$ is the target shape increment at stage $t$, $I$ is the $n$th training image and $s$ is the shape estimate from the previous stage. The initial shape is generally perturbated multiple times, using data augmentation, to improve

generalization. During inference, the same objective as during training is kept. The shape estimate is gradually updated using the shape increments transformed back into global coordinates. The complete process is a two-level boosted regression, which consists of a cascade of random ferns, i.e. 10 stages of 500 ferns with a depth of 5. Candidate features are proposed by generating a random set of normalized pixel coordinates indexed relatively to the nearest landmark on the mean shape. These shape-indexed features are extracted based on the current shape estimate at each stage by transforming the pixel locations back into global coordinates. Local pixel intensity differences provide discriminative features invariant to illumination. A correlation-based feature selection strategy is then applied to retain only the most discriminative and distinctive features using Pearson's correlation coefficient. The sum of the outputs of each fern constitutes the final output of the stage. With some adjustments, this approach can achieve impressive execution speeds and handle partial or uncertain labels during training (Kazemi and Sullivan 2014).

CSR approaches differ mainly in terms of the shape-indexed features and regressors they use. An approach worth mentioning is the supervised descend method (SDM) (Xiong and De la Torre 2013). Initially developed to solve general nonlinear least-squares problems, SDM uses linear regression with SIFT features to learn descent directions that minimize the objective. Random forests can also be used to learn discriminative local binary features for each landmark and to learn the mapping between the concatenated features and the facial shape (Ren et al. 2014). Gaussian process regression trees (Lee et al. 2015) have been proposed to enhance generalization. A Gaussian process regression tree corresponds to a kernel function defined by a set of trees to measure the similarity between two inputs and uses differences of Gaussians as input features.

One of the main concerns of CSR may be the initialization with the mean shape, which is sub-optimal. Poor initialization can lead to local optima and impact robustness, for example, for large head poses (Smith et al. 2014). DR can be used to generate a better initial estimate (Zhang et al. 2014a). The bounding box can be randomly shifted and rescaled to generate multiple shape hypotheses and learn to rank or combine them to get the final result (Yan et al. 2013). The coarse-to-fine strategy can also be employed to perform a progressive and adaptive shape searching (Zhu et al. 2015). Another concern related to the mapping from the face bounding box to the facial shape is the sensitivity to the face detector (Sagonas et al. 2016). A mapping learned with bounding boxes from one detector could perform poorly if a different detector is used during inference, due to the possible differences in box sizes and locations. Lastly, a minor concern could be the fixed number of cascades. Good results or locally optimal solutions can be delivered early in the process, making the rest unnecessary. There is some work that considers confidence measure to stop the process or for adjusting the number of cascades (Shapira et al. 2021).

### 1.1.3. *Deep learning approaches*

Traditional ML techniques that have been used so far are not very efficient when dealing with raw image data directly, i.e. to work directly with pixels. Facial landmark detection is too difficult due to the variations in appearance (e.g. position, orientation, illumination) that can occur, especially under uncontrolled conditions. To overcome this issue, discriminative features as invariant as possible to variations are extracted (e.g. HOG, SIFT), which require engineering skills and domain expertise. This can be avoided by using DL approaches, which are capable of learning highly discriminative task-specific features. This subset of ML methods is based on artificial neural networks (ANNs). These methods, although not new (LeCun et al. 1990; Lawrence et al. 1997), have recently led to breakthroughs in many fields, including computer vision (Krizhevsky et al. 2012). This is associated with the increase in computational power and data availability currently required to leverage such approaches.

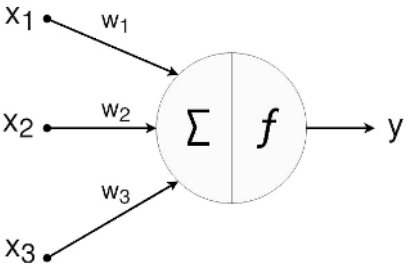


**Figure 1.4.** *Structure of an artificial neuron. A weighted sum of the inputs is performed. The result is then passed through a nonlinear activation function*

The most common type of ANN is the feedforward network, in which information is forwarded in a single direction through multiple layers of computational units, without cycles or loops. There is a variety of architecture with a potentially high number of layers, hence the use of the term DL. The multilayer perceptron (Rumelhart et al. 1985) can be considered as the groundwork of DL and is often referred to as vanilla neural networks. Nowadays, CNN is the architecture commonly used to process images (LeCun et al. 1998). ANNs are often referred to as biologically inspired models (Rosenblatt 1988). An artificial neuron, illustrated in Figure 1.4, corresponds to a very simplified version of biological neuron. It can be formulated as follows:

$$y = f(\sum_{i=1}^{I} W_i x_i + b) \qquad [1.10]$$

where $x_i$ is the $i$-th input, $W$ are the connection weights, $b$ is the neuron bias, and $f$ is called the activation function. The same applies to CNNs, which mimic the

ventral pathway of the visual cortex by considering an image as a compositional hierarchy using convolution and pooling layers inspired by simple cells, complex cells and the LGN–V1–V2–V4–IT hierarchy in the brain (Hubel and Wiesel 1962; Felleman and Van 1991). Each layer of a CNN represents the input at a specific level of abstraction. The first layers focus on low-level features such as colors, edges and orientation, while high-level ones (parts of objects, objects) are obtained in the subsequent layers by composing these low-level features. Although this is an over-simplification from the actual visual cortex, CNNs allow us to learn complex functions, which outperform traditional ML techniques in most tasks (Alom et al. 2018). Among the popular CNN architectures are AlexNet (Krizhevsky et al. 2012), VGG (Simonyan and Zisserman 2014), Inception (Szegedy et al. 2015) and ResNet (He et al. 2016), which are widely used in the literature, or at least they serve as a basis for more specific architectures. Today, much work is focused, not exclusively, on designing architectures: deep (Simonyan and Zisserman 2014), wide (Wu et al. 2019), new layer (Dai et al. 2017), new connection (He et al. 2016), new layer composition (Szegedy et al. 2015), efficiency (Iandola et al. 2016), structural diversity (Zhang et al. 2017), combination with a projection model (Zhu et al. 2016), etc. Many criteria need to be considered, including the amount, type and quality of data available for the targeted problem.

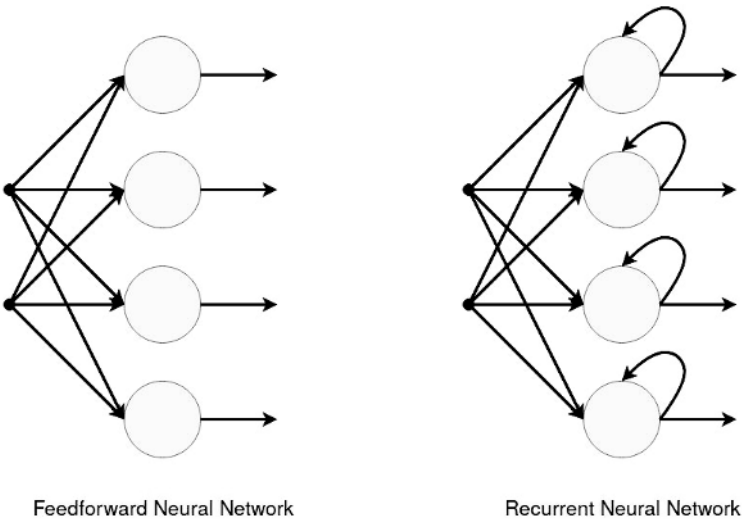


**Figure 1.5.** *Comparison between feedforward and recurrent neural networks. In feedforward networks, the information flows in one direction only, from the input layer to the output layer. In recurrent neural networks, there are cycles or loops, which make them useful for time series/sequential tasks*

Conventional feedforward networks have no persistence. To allow a reasoning about their past decisions, they can be extended by adding a feedback loop to them

and thus making them recurrent (see Figure 1.5 and Rumelhart et al. 1988). This introduces a notion of sequence where information can be passed from one step to another. A common way to represent recurrent neural networks (RNNs) is to unroll each step, which results in multiple copies of the same network interacting with each other. In the following, the different layers used to build a DNN architecture, how to train and deploy such an architecture and the specificities related to the landmark detection problem are described.

#### 1.1.3.1. *Theory*

A convolutional neural network is a composition of layers, each performing a (supposedly) differentiable function, all together resulting in an architecture that enables the learning of a nonlinear mapping between input(s) and output(s). There are several types of layers, the most common ones being fully connected (*fc*) layers, convolution (*conv*) layers, activation layers and pooling (*pool*) layers.

As illustrated in Figure 1.6, *fc* layers are composed of a fixed number of neurons that are connected to all the neurons in the adjacent layers. Each neuron output is a weighted sum of its inputs. This leads to a high-level reasoning through the extraction of global relationship between neurons/features.

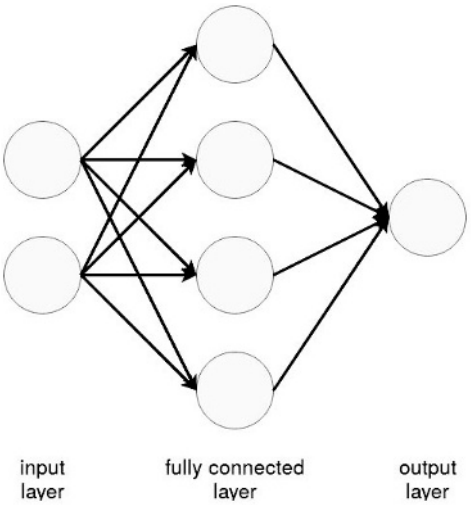


**Figure 1.6.** *Illustration of an fc layer. Each neuron in the fc layer is connected to each neuron in the previous layer*

Because using the *fc* layer for high-dimensional data is impracticable due to too many connections, *conv* layer was proposed (LeCun et al. 1998). CNNs make the explicit assumption that the inputs are images, which results in two key properties

of the *conv* layer: local connections and weight sharing. A *conv* layer produces a set of two-dimensional feature map. Each neuron of a feature map is connected to a local region, its receptive field, in the previous layer through a set of parameters called a filter or kernel (see Figure 1.7). The local group of pixels are generally highly correlated. Since a feature should be useful to compute at different positions, all neurons in a feature map share the same filter. This provides a feature extractor with translation equivariance and far less parameters than with an *fc* layer. 2D convolution can be expressed as:

$$v_{ij}^{xy} = b_{ij} + \sum_{m} \sum_{p=0}^{Pi-1} \sum_{q=0}^{Qi-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} \qquad [1.11]$$

where $v_{ij}^{xyz}$ is the value at position $(x, y)$ on the $j$-th feature map in the $i$-th layer, $b$ is the bias, $m$ is the index of the feature map in the previous layer, $P, Q$ are respectively the height and width of the kernel and $w$ is the value of the kernel. Note that a one-dimensional *conv* layer, where the filter is of the same size as the input, is equivalent to an *fc* layer.

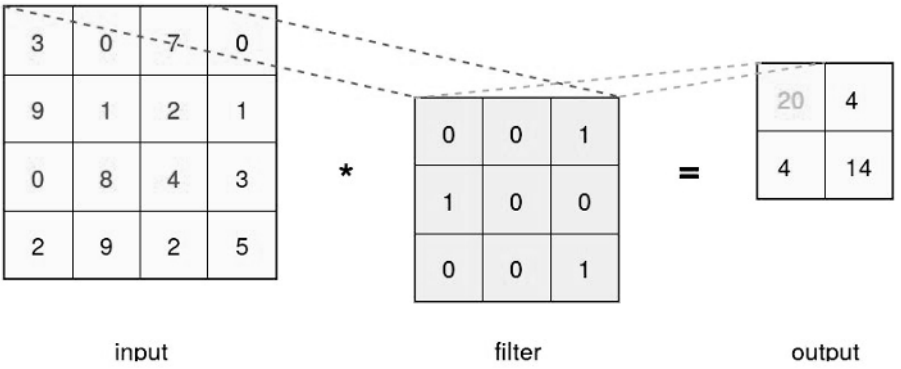


**Figure 1.7.** *Example of a convolution operation with a 2x2 filter and stride 1. The dot products between the filter and a local region of the input are illustrated. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

The activation layer is an elementwise function, which performs a nonlinear transformation, for example, rectified linear unit (ReLu) (Nair and Hinton 2010). Without this transformation, only linear operations are performed by the network. By increasing its nonlinear properties, this makes the network more suitable to solve complex problems. ReLu is one of the most commonly used activations. In

comparison to the first used activations, for example, tanh and sigmoid, it makes the training faster and alleviate optimization problems. It can be expressed as follows:

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad [1.12]$$

Finally, the *pool* layer corresponds to a downsampling operation, which generally computes the maximum or average of a local region. Max pooling is depicted in Figure 1.8. It helps further reduce the amount of parameters and computation, allowing a better control of overfitting. It also adds invariance to translation (small shift) and distortion (LeCun et al. 1998).

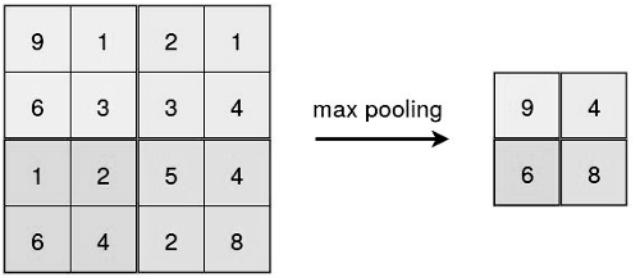


**Figure 1.8.** *Example of a max pooling with a 2x2 filter and stride 2. The maximum of local regions is computed, which leads to a downsampling of the input. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Most of these layers have hyperparameters, which are not learned but must be defined manually. It is necessary to specify the capacity in terms of the number of neurons for *fc* layers. The number, size and the displacement (stride) of filters are to be defined for *conv* layers. Similarly, the size and displacement of the window are also to be provided for *pool* layers. To control the size of the input and avoid losing information at the boundaries, *conv* and *pool* layers also require a certain amount of padding. Figure 1.9 depicts stride and padding. The *conv* and *fc* layers also have parameters that need to be learned.

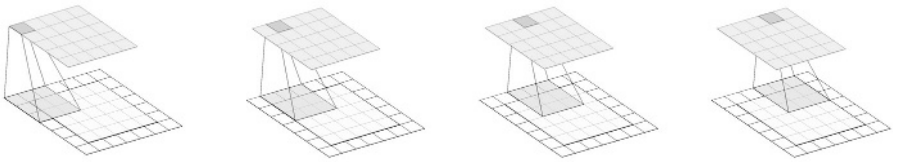

**Figure 1.9.** *Behavior of a filter of size 3x3 with padding and stride 1. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Once the architecture is defined, the training phase aims to adjust the layer parameters to obtain the best possible predictions. An objective or loss function is applied at the output of the network to measure the error between predictions and ground truth. Mean-squared error (MSE) is one of the most common regression loss functions:

$$MSE = \frac{1}{N}\sum_{n=1}^{N} ||\overline{s}_i - s_i||^2 \quad [1.13]$$

where $N$ is the number of training samples, $\overline{s}_i$ is the $i$-th ground truth shape and $s_i$ is the $i$-th predicted shape. The (mini-batch) stochastic gradient descent (SGD) algorithm (Robbins and Monro 1951) is used to find the minimum of this function. The error is computed and distributed backwards (backpropagation). A gradient vector for each parameter is computed from the last layer to the first using the chain rule. It determines the amount by which the error would increase/decrease if the parameter was changed. The parameter is then adjusted in the direction opposite to the gradient vector multiplied by an additional hyperparameter called the learning rate (LR). The latter defines by how much the parameters are changed according to the error, which helps to control the speed at which the model learns. It is called stochastic because batches of training samples are selected randomly to compute gradient at each iteration instead of the full training set. It helps prevent the algorithm from reaching local minima. A parameter update can be expressed as follows:

$$W \leftarrow W - \lambda(\frac{\partial MSE}{\partial W}) \quad [1.14]$$

where $W$ corresponds to the parameters, $\lambda$ corresponds to the LR and $\frac{\partial MSE}{\partial W}$ corresponds to the partial derivative of the lost function with respect to the parameters. To avoid gradient problems when performing gradient descent optimization, the parameters are initialized with random values in such a way that they are not too small nor too large, for example, Glorot initialization. Going through all the training data corresponds to an epoch; it is done multiple times. This iterative process leads to a convergence towards a minimum generally close to the global solution. It is crucial to carefully tune hyperparameters, some as the LR being critical to ensure a good convergence. Variants of SGD such as Adam (Kingma and Ba 2014) can be used to maintain an LR for each parameter and to adapt it dynamically during training. Once training is completed and the parameters are set, the network can then return the corresponding landmarks from an input image by passing it through the stack of layers.

RNNs rely on an extension of backpropagation called backpropagation through time (BPTT), which unrolls the network and propagates the error backward over the entire input sequence. The weights are then updated with the accumulated gradients. BPTT can be computationally expensive as the number of time steps increases. To

avoid computational burden, truncated BPTT is generally used. The sequence is split into smaller sequences that are treated as separate training cases. Vanilla RNNs have fundamental optimization limitations as the gradients tend to either vanish or explode during optimization, resulting in an unstable network unable to learn long-term dependencies (Bengio et al. 1994). Gradient explosion can be addressed through gradient clipping (Pascanu et al. 2013). Gradient vanishing can be solved using long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) illustrated in Figure 1.10. By adopting a memory cell coupled with gating functions to control the flow of information, the network is better able to learn over long sequences. LSTM can be expressed as:

$$
\begin{aligned}
i_t &= \text{sigmoid}(W_{xi}X_t + W_{hi}H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\
f_t &= \text{sigmoid}(W_{xf}X_t + W_{hf}H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc}X_t + W_{hc}H_{t-1} + b_c) \\
o_t &= \text{sigmoid}(W_{xo}X_t + W_{ho}H_{t-1} + W_{co} \circ C_t + b_o) \\
H_t &= o_t \circ \tanh(C_t)
\end{aligned}
\qquad [1.15]
$$

where $X_t$ and $H_t$ are, respectively, the input and the hidden state at time $t$. $W$ denotes the weights, $b$ denotes the bias, and $\circ$ denotes the Hadamard product. $C_t$ is the memory cell of the LSTM, which is updated by the input gate $i_t$, the forget gate $f_t$ and the output gate $o_t$. Theses gates help reduce the vanishing gradient effect. Gated recurrent unit (GRU) (Cho et al. 2014) is an increasingly popular alternative with a simplified structure, which combines the forget and input gates into a single update gate and merges the hidden state and memory cell.

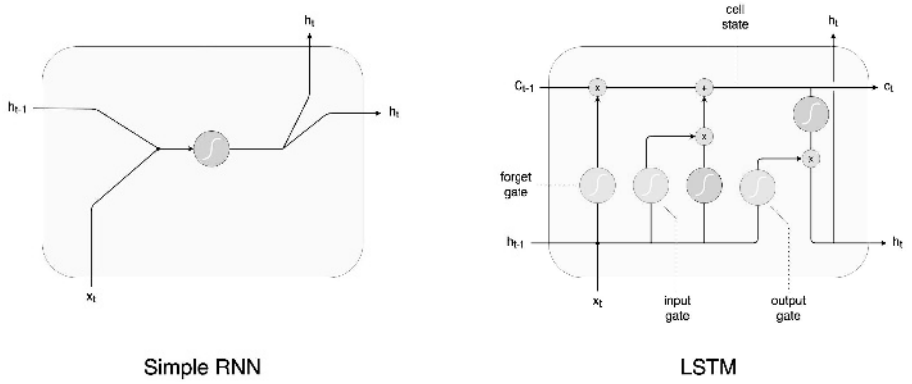


**Figure 1.10.** *Comparison of a vanilla RNN unit (left) and an LSTM block (right). Compared to vanilla rnn, LSTM has a memory cell updated by different gates to control the flow of information and reduce the vanishing gradient effect. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Beyond the standard supervised learning framework, other techniques can be used. Adversarial learning (AL) is an increasingly popular framework, which consists of replacing or supplementing the explicit loss function with one or several DNNs. This is often presented as a game theory scenario where several networks are used to compete against each other. Training is modeled as a zero-sum game. The gain of one necessarily constitutes a loss for the other. As shown in Figure 1.11, one neural network, called the generator (G), generates new data (e.g. an heatmap), while its opponent, the discriminator (D), tries to detect if the data is real or if it is the result of G. The aim of G is thus to fool D. The gradients accumulated from the loss of D are used to update G. In this way, D provides feedback to G so that it may improve. This strategy is also known as generative adversarial networks (GANs) (Goodfellow et al. 2014). The original GAN loss function can be expressed as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad [1.16]$$

where D maximizes the probability where it correctly classifies reals and fakes ($logD(x)$) and G minimizes the probability that D will predict that its outputs are fake ($log(1-D(G(z)))$).

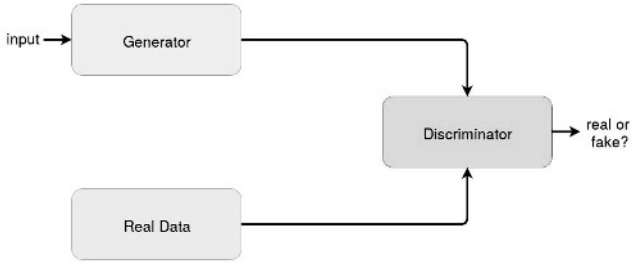


**Figure 1.11.** *Overview of GANs. The generator generates new data while its opponent, the discriminator, tries to detect if the data is real or if it is the result of the generator. The discriminator then provides feedback to the generator so that it may improve. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

#### 1.1.3.2. *Advances*

In the recent Menpo challenge (Zafeiriou et al. 2017), which is a good indicator of trends on facial landmark detection problem, only DL approaches have been proposed. Currently, two strategies to design architectures for landmark detection can be distinguished: coordinate regression and heatmap regression. They respectively regress the coordinates directly using a fully connected layer (Sun et al. 2013) or compute heatmaps, one for each landmark with the same size as the input, using an FCN (Bulat and Tzimiropoulos 2017b). The coordinates of the maximum value in a heatmap are generally considered as the coordinates of a landmark.

One of the first DNNs designed for landmark detection is based on the CSR approach presented in section 1.1.2 (Sun et al. 2013; Zhou et al. 2013). It consists of multiple levels of CNN, the first level processing the whole face to capture texture information and geometric constraints, while the subsequent levels perform a local refinement. Figure 1.12 illustrates the CNN architecture composed of four *conv* and *pool* layers and two *fc* layers allowing the DR of landmark coordinates. There is now a trend towards fully convolutional network (FCN) architectures. It has some advantages (Long et al. 2015), including no input size limit, better spatial information handling and a better computational cost/representation power ratio. The HourGlass (HG) network shown in Figure 1.13 is currently very popular and considered as the state-of-the-art facial landmark detection method (Bulat and Tzimiropoulos 2017b). Initially designed for human pose estimation (Newell et al. 2016), HG consists of a symmetrical encoder–decoder with skip connections between the two to consolidate information at different scales. It is generally stacked several times (up to 8) with intermediate supervision to follow the CSR approach. A compromise between these two approaches is also possible, by regressing coordinates and using landmark heatmaps to transfer spatial information across stages (Kowalski et al. 2017) or by using soft-argmax loss (Honari et al. 2018). HG is the most popular model, yet it struggles to run in real time due to its high computational complexity. Neural network compression such as weight binarization can be applied to improve speed and reduce the size of the model, but at the expense of accuracy (Bulat and Tzimiropoulos 2017a). Student–teacher training strategy can enable knowledge distillation from a large complex model to a lightweight one (Zhao et al. 2020). Upsampling layers can be removed without loss of accuracy by predicting an offset along with the score on low-resolution feature maps (Jin et al. 2021).

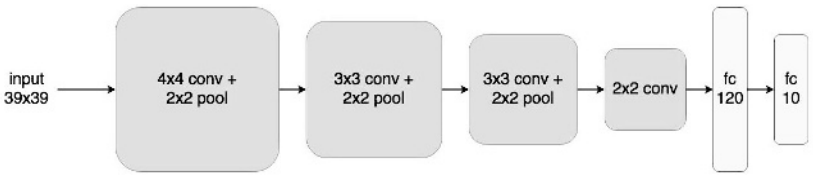


**Figure 1.12.** *Early CNN architecture for facial landmark detection (Sun et al. 2013). It consists of four conv and pool layers and two fc layers to produce landmark coordinates. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

CSR remains popular within DL approaches and can be implemented in different ways, either by stacking networks (Zhang et al. 2014a; He et al. 2017; Guo et al. 2018a), as seen previously, or by formulating it as a recurrent process by combining CNNs and RNNs (Trigeorgis et al. 2016; Lai et al. 2018; Wang et al. 2018). In the latter, each recurrent step corresponds to a shape increment; shape increments are jointly learned relying on an explicit memory shared across steps. CSR can be

improved by adding attention mechanisms to the network to enable it to focus on relevant parts of the data. Such a mechanism, combined with an RNN, can identify reliable landmarks and use them as attention centers to primarily refine landmarks close to them (Xiao et al. 2016). It can also be incorporated into *conv* layers using intermediate supervision to generate top-down attention maps (Yue et al. 2018). This allows the network to focus on regions around the landmarks and thus capture more discriminative and powerful features for facial landmark detection. This approach suggests that CSR is not a necessity and that DR remains a valid solution. Deep reinforcement learning (DRL) has also proven to be a promising alternative to CSR through the use of a shape searching policy, which maximizes a shape evaluation function (Liu et al. 2018b).

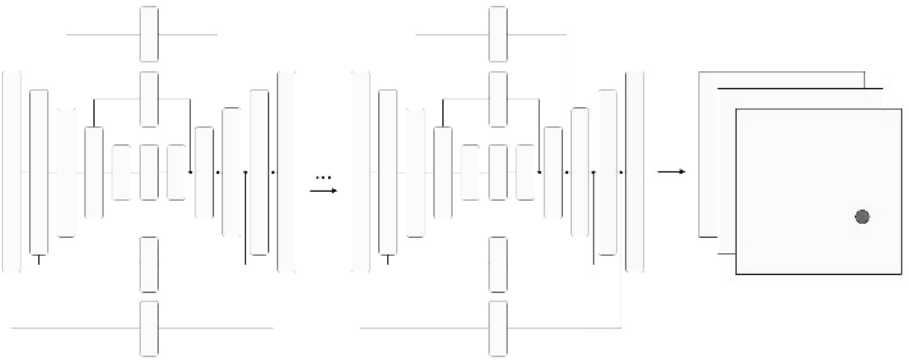

**Figure 1.13.** *Hourglass network (Bulat and Tzimiropoulos 2017b). It consists of a symmetrical encoder–decoder with skip connections between the two to consolidate information at different scales. It is generally stacked several times to produce heatmaps, one for each landmark, in a coarse-to-fine manner. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Other improvements have been proposed. Well-known work on CNN architecture has been applied to facial landmark detection. CoordConv adds control over translation invariance through an extra coordinate channels (Huang et al. 2020). Anti-aliased downsampling (e.g. max-pooling, strided-convolution) includes a low-pass filter to improve the shift invariance property of deep networks (Huang et al. 2020). Deformable convolution allows us to have adaptive receptive fields by integrating 2D offsets (Zhu et al. 2020). Continuous heatmap encoding–decoding process, siamese-based training and loss based on the Jensen–Shannon divergence (Wan et al. 2020; Bulat et al. 2021) have been proposed to remove quantization errors and achieved subpixel accuracy. A comparative analysis of different loss function (L1, L2) (Feng et al. 2018; Wang et al. 2019) suggests that more attention should be paid to small- and medium-range errors and the authors proposed a new loss function called the Wing loss to address this problem. When using soft-argmax loss, a LaplaceKL

penalty can be added to explicitly take into account the variance of heatmaps in addition to the mean (i.e. landmark location) to penalize low confidence (Robinson et al. 2019). The loss function can also be used to apply more specific constraints during training. A contextual loss has been proposed to capture the visual context of each landmark (Zeng et al. 2018). It can be combined with a structural loss to add an explicit geometric constraint in order to exploit the hierarchical structure of the face. This constraint can be incorporated into the model in different ways. It can involve the use of a second model as a discriminator to distinguish between real and fake shapes with the aim of providing a feedback to the first model (Chen et al. 2019c). It can also be achieved by combining the DNN with a conditional random field (Chen et al. 2019a). Other information related to landmarks such as the boundary of the face can be integrated and used as a representation of geometric structure (Chen et al. 2019c; Huang et al. 2020). Using the Wasserstein distance as loss can also provide additional geometric information (Yan et al. 2021).

### 1.1.4. *Handling challenges*

Under uncontrolled conditions, the approaches presented in sections 1.1.1, 1.1.2 and 1.1.3 continue to encounter difficulties, for example, head pose, occlusions, facial expressions, illumination. Some of the approaches are capable of implicitly dealing with these challenges, but they remain most often ineffective when extreme cases occur. A number of authors propose to explicitly address some selected challenges, and the proposed solutions are generally complementary to most approach. These include the use of 3D models or multiple specialized models (e.g. multi-view) to treat profile faces more consistently, multi-task learning with expression recognition as a related task to better handle facial expressions or extended dataset annotations to explicitly detect occlusions. In this section, a focus is placed on three challenges considered to be the most complex to address (i.e. head pose, occlusions, expressions) and review the proposed solutions to tackle them.

#### 1.1.4.1. *Head pose*

Head pose variations are one of the main sources of error. They lead to significant changes in appearance and shape, especially when switching from frontal to profile views, which hides an entire part of the face. The approaches presented so far assume that the landmarks to be detected are visible. This is not true with profile faces. Moreover, there is little training data for extreme poses due to the difficulty of annotating it. To address these problems, apart from the few multi-tasking or conditioning solutions, which condition the measurement of 2D coordinates on 3D pose (Xu and Kakadiaris 2017; Kumar and Chellappa 2018), two solutions currently stand out: multiple view-specific models and dense 3D model.

Multiple view-specific models can be used to achieve a better accuracy under extreme poses, one for each view, for example, left profile, frontal view and right

profile. The best model is then selected depending on the scenario. It can be done using the head pose based on a few landmarks (Yan et al. 2003; Yu et al. 2013; Deng et al. 2016) or the confidence score of the models (Cootes et al. 2000; Zhu and Ramanan 2012). To reduce selection failures and improve fault tolerance, one possibility, instead of splitting the training data by pose, is to use the whole training data for each model and more heavily weight the training samples of the specific pose domain, with overlap between domains (Feng et al. 2017). Anchor templates can also be used as reference to directly split the search space. Offsets are regressed for multiple templates, and predictions are then aggregated (Xu et al. 2021). An alternative is to merge the results of different pose-dependent models (Dantone et al. 2012). Still, model selection and model fusion are not trivial tasks as it requires view estimation or the use and ranking of different models. More recently, with the introduction of a specific landmark configuration for profile faces (Zafeiriou et al. 2017), a joint multi-view model has been proposed to avoid using distinct models, by capitalizing on the correspondences between views (see Figure 1.14) (Deng et al. 2019).

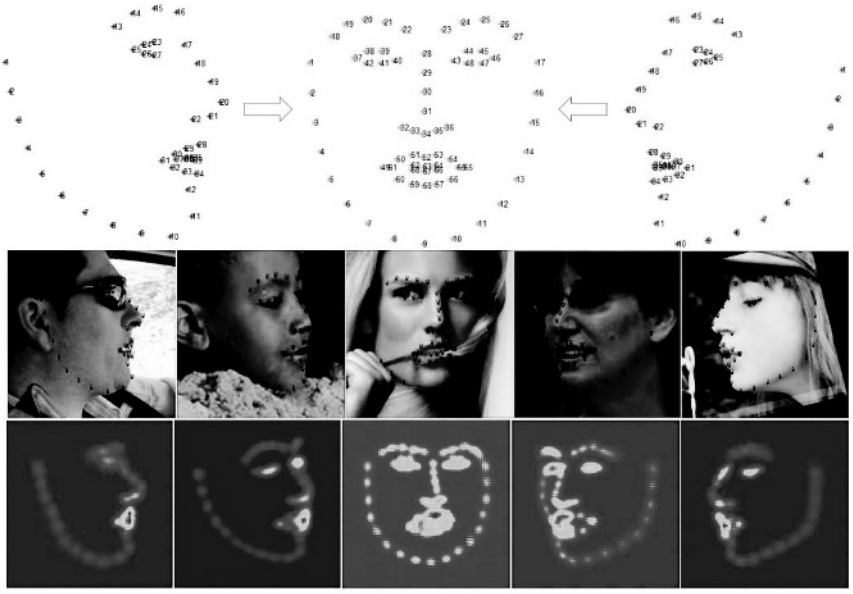

**Figure 1.14.** *Joint multi-view by capitalizing on the correspondences between views. This helps achieve better accuracy under extreme poses. Source: Deng et al. (2019), CC BY (http://creativecommons.org/licenses/by/4.0/), used without modification. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Pose variations can also be handled by fitting a dense 3D model, for example, a 3D morphable model (3DMM) (Blanz and Vetter 1999), to a 2D image. Considering the 3D structure of the face is particularly useful to reliably reason about self-occluded

landmarks. The projection matrix and model parameters are generally estimated using cascaded regression, and then a 3D-to-2D projection is performed to infer the 2D landmark locations from the 3D face (Jourabloo and Liu 2017; Zhu et al. 2017). A single model has also been proposed to improve training and inference times (Bhagavatula et al. 2017; Jourabloo et al. 2017). The main limitation with these methods is that the shape is restricted by their linear parametric 3D model, which may lead to a lack of precision in detection. This can be mitigated by jointly refining the 3D face model and 2D landmark locations through a recurrent 2D–3D double learning process to mutually reinforce each other (see Figure 1.15) (Xiao et al. 2017).

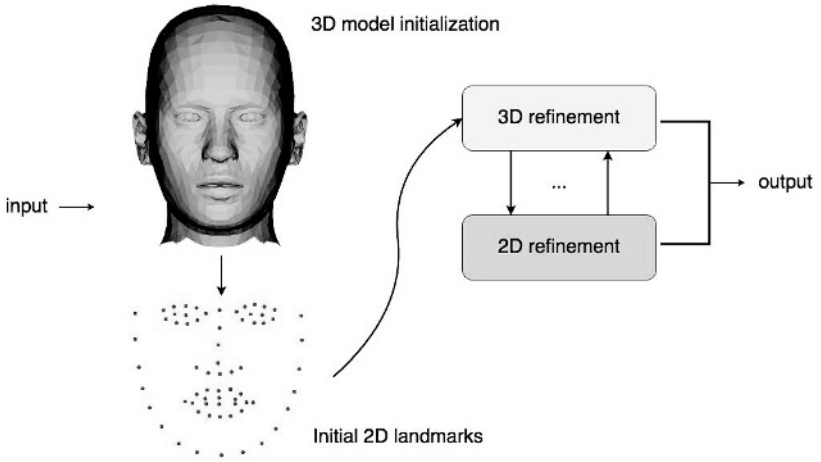


**Figure 1.15.** *2D–3D dual learning to mutually reinforce each task (Xiao et al. 2017). Considering the 3D structure of the face is particularly useful to more effectively address self-occluded landmarks. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

#### 1.1.4.2. *Occlusions*

Among other causes of error are occlusions, which lead to a loss of information since they cover a more or less important part of the face. Occlusions that are locally consistent, meaning that they do not hide all the landmarks, are considered. This challenge is particularly complex to deal with due to the variety of occlusions, which can be caused by objects with arbitrary appearances and shape, or self occlusions, for example, due to head pose. It requires relying on non-occluded parts and, therefore, determine which parts are hidden. Occlusions can be detected explicitly to improve the robustness to outliers (i.e. misdetected landmarks).

A first approach is to use occlusion-dependent models. The facial image is generally divided into several regions in which some are assumed occluded. Different models are trained with each of these regions. The resulting predictions are merged

according to certain criteria such as the amount of occlusion present in the regions (Burgos-Artizzu et al. 2013; Yu et al. 2014). The main drawback of this approach is that facial occlusions are totally arbitrary, and therefore the approach cannot cover the variety of occlusions. Moreover, relying on a single region provides insufficient appearance information. Another strategy for dividing and processing the facial image is to segment the image into non-overlapping regions and predict a heatmap for each region, which indicates how useful is the information contained in it. This heatmap is used together with the facial image to perform landmark detection (Yang et al. 2015).

A second approach helps further alleviate the mentioned drawbacks by using a unified framework to jointly model the appearance around each landmark, landmark locations and visibility, and occlusion pattern (Ghiasi and Fowlkes 2014). CSR can be used to predict landmark location and visibility with an explicit occlusion pattern constraint and more weight assigned to visible landmarks (Wu and Ji 2015b). With recent DL approaches such as FCN, occlusion information can be embedded into heatmap labels (Yuen and Trivedi 2017). Broadly speaking, these approaches require ground-truth annotations for occlusions or synthetically occluded data, which makes the annotation of datasets more laborious. However, more recent work has shown that the use of labels can be avoided. An attention mechanism can help filter features from occluded regions. Then, missing features can be recovered by leveraging inter-feature correlations of the face through low-rank learning along with geometric features (Zhu et al. 2019). To model inter-feature correlations, a capsule network has also been considered. The use of capsules provides a better ability to represent knowledge about transformations of the target object but suffers from a high computational cost (Ma et al. 2022). The uncertainty of the landmark can also be quantified with a probabilistic heatmap regression model, which can allow occlusions to be detected (Chen et al. 2019b).

#### 1.1.4.3. *Expressions*

Facial expressions also cause important changes of facial appearance and shape. There are seven universal expressions (Ekman and Keltner 1970): happiness, sadness, fear, disgust, anger, contempt and surprise. All over the world, they are used to convey the same emotions. The various state transitions and intensities are source of localized and complex motions. Beyond these universal expressions, more complex expressions are likely to be encountered in natural environments (Russell 1980; Plutchik 2001). Recent approaches (i.e. DL) deal reasonably well with expressions in terms of the Euclidean distance. However, it is unclear at this time whether this is suitable to applications since the Euclidean distance is a specific evaluation metric for landmark detection and does not reflect some cases such as instability on videos. A few methods have been specifically designed to be robust to expressions by using multi-state local shape models or face shape prior models to deal with the shape variations of facial components (Tong et al. 2007; Wu and Ji 2015a). Considering the correlation between shape and expression, a more common solution is not to treat facial landmark detection

as an independent problem but to jointly learn various related tasks in order to achieve individual performance gains. Among the related task is facial action unit detection (Li et al. 2013; Shao et al. 2018). The relationship between facial action units and landmarks can serve as a constraint when iteratively updating landmarks during CSR (Wu and Ji 2016). Facial expression recognition is also widely used (Wu et al. 2014; Zhao et al. 2014; Zhang et al. 2016; Ranjan et al. 2017). An example is depicted in Figure 1.16. While such a solution may be straightforward to implement, by adding as many outputs as auxiliary tasks, it can make the training stage much more complex because the optimal convergence rates may vary from one task to another (Zhang et al. 2016).

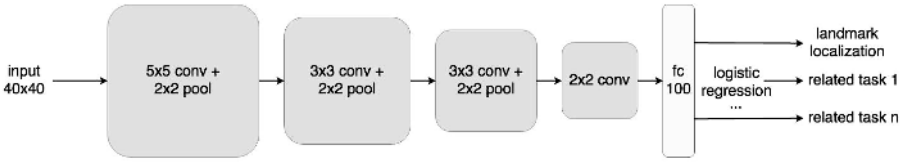


**Figure 1.16.** *Landmark detection with auxiliary tasks based on multi-task learning (Zhang et al. 2014). Facial landmark detection is not treated as an independent problem but jointly learned with various related tasks in order to achieve individual performance gains. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

#### 1.1.4.4. *Other*

Some challenges, mostly related to the quality of the image, for example, illumination, resolution, blur, have attracted less interest, either because they are easier to handle or because they do not have major impacts in terms of occurrence frequency or error. Recently, work based on GANs has been carried out to detect landmarks more accurately on low-resolution images (Bulat and Tzimiropoulos 2018). As shown in Figure 1.17, it involves three models to improve both super-resolution and facial landmark detection: a generator for synthesizing high-resolution images from low-resolution ones, a first discriminator to distinguish between synthesized and original high-resolution images, and a second discriminator for the detection of landmarks on synthesized high-resolution images.

On the other hand, tackling very high-resolution images can also be challenging, especially in terms of memory. To avoid working with the entire image, regions of interest can be automatically defined via attention-driven cropping based on a global network working on low-resolution images (Chandran et al. 2020). Another solution is to predict 1D heatmaps instead of 2D heatmaps to represent x and y coordinates and use an attention mechanism to implicitly capture the joint distribution (Yin et al. 2020).

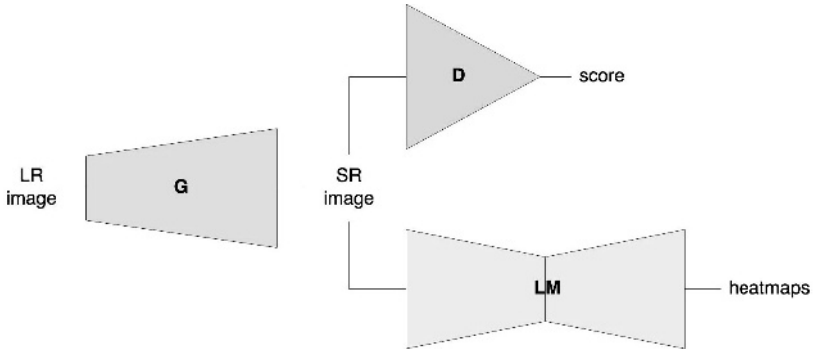


**Figure 1.17.** *Adversarial learning involving super-resolution to improve landmark detection on low-resolution images (Bulat and Tzimiropoulos 2018). A generator synthesizes high-resolution images from low-resolution ones, while a first discriminator distinguishes between synthesized and original high-resolution images, and a second discriminator detects landmarks on synthesized high-resolution images. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

The datasets used to solve the landmark detection problem present significant intra/inter-dataset variations, which can lead to poor generalization. By coupling two models, it is possible to leverage these variations (Wu and Yang 2017). The first model takes advantage of intra-dataset variations by using an implicit feature sharing mechanism. It takes an image and its flipped version as input. The second model, which focuses on inter-dataset variations, decides which candidate should be considered. Some authors have rather focused their work on simplifying the landmark detection task by normalizing the input to a canonical pose. The most advanced solutions employ supervised face transformations such as spatial transformer networks to remove the translation, scale and rotation variations of the face (Fan and Zhou 2016; Chen et al. 2017; Kowalski et al. 2017; Lv et al. 2017; Yang et al. 2017). While most authors focus on the variance of faces, the intrinsic variance of image styles can also be handled to improve performance using a style-aggregated network, which takes as input two complementary images of the same face, the original style image and an aggregated style image generated by GANs (Dong et al. 2018a). The definition of style can be pushed further by considering various environment factors. A variational auto-encoder can be trained to disentangle face images into a style and structure space, and then be used to generate new synthetic data through style translation (Qian et al. 2019).

Beyond the challenges encountered in images, annotation is also considered challenging and often proves to be imprecise, degrading the performance of current models. To address this issue, a bivariate label distribution can be associated with each landmark to cover the neighboring pixels (Su and Geng 2019). A latent variable that

represents a semantically consistent ground truth can also be introduced and optimized (Liu et al. 2019). Error-bias is found to be particularly strong in tangent direction. More constraints can be imposed on normal direction than on tangent direction using anisotropic direction loss and anisotropic attention module (Huang et al. 2021).

### 1.1.5. *Summary*

Generative approaches can achieve interesting performance under uncontrolled conditions using only few data, with suitable representation and optimization strategy (Tzimiropoulos et al. 2012; Tzimiropoulos and Pantic 2013). However, they can be computationally intensive, and due to the difference between the model and the true distribution of data, their ability to generalize is often restricted (Bernardo et al. 2007). Besides, a small error in the prediction of the model coefficients may lead to a large error regarding the landmarks. This is not true discriminative approaches, which are becoming increasingly popular. Instead of predicting the model coefficients, landmarks are directly predicted, resulting in a better accuracy.

Part-based approaches such as CLMs, although more robust to partial occlusions and lighting, suffer from an ambiguity problem along with performance issues due to their local appearance representation coupled with a global shape optimization. To overcome the limitations of these approaches, another popular approach is to estimate the whole facial shape directly from the image features while implicitly exploiting spatial constraints. This is less restrictive than modeling them explicitly with a shape model. Two solutions can be found: DR and CSR, the latter being the most commonly used and a state-of-the-art approach. By providing a mapping from the image to the facial shape without the use of any parametric appearance or shape model, faster and more accurate predictions can be obtained. The coarse-to-fine strategy of CSR offers a progressive and flexible way to encode the shape constraints while increasing robustness. The availability of large datasets further improves the generalization capacity of discriminative approaches. However, it is still not easy to perform this mapping in some scenarios, such as extreme expression and pose variations. Recently, a shift from traditional approaches to DL-based approaches has occurred. The latter can more effectively model the nonlinear relationship between images and shapes and has already shown impressive performance.

Challenges in facial landmark detection have benefited from advances in DL. With such approaches, the need for feature engineering is no longer necessary. The research has shifted to architecture design where domain expertise can still be useful but not as crucial as it used to be. Although CSR continues to predominate, DR appears to be practicable with DNNs, as some authors have shown (Merget et al. 2018; Miao et al. 2018). However, DNNs require a significant amount of computational power and annotated data, which is not easily affordable. It is also worth noting that landmark detection is a structural prediction problem that can hardly benefit from

models pre-trained on massive amounts of data such as ImageNet (Deng et al. 2009) for object detection and localization. On the other hand, overfitting can be an issue when building a deep model from scratch due to the lack of data available. Despite these challenges, DL approaches have helped take a further step towards solving the problem of facial landmark detection, but there is still much to be done.

Complementary solutions have been developed, which can be integrated into both the DL and the more traditional approaches. This include the use of 3D model, multi-view-specific models, multi-task learning, GANs and explicit occlusion modeling. They can help solve the remaining challenges and further increase generalization. By integrating such solutions, it is possible to significantly improve the robustness to specific challenges. However, it is often a combination of these challenges that are encountered under uncontrolled conditions. Although these solutions are not mutually exclusive, it may be too complex and costly to apply them together. Another kind of solution based on temporal consistency and allowing several challenges to be addressed at the same time is possible.

All the approaches presented so far, when applied to a video stream, are not able to use temporal information. Yet, with the ubiquity of video sensors, the vast majority of applications rely on videos. Moreover, there is a need for an approach that can deal with all challenges at once. Recent work has proved that taking into account video consistency helps to deal with the variability in facial appearance and ambient environment encountered under uncontrolled conditions. These works are reviewed in the following section.

## 1.2. Extending facial landmark detection to videos

Despite the quantity of facial landmark detection approaches proposed in the literature and the recent major advances, the difficulties encountered under uncontrolled conditions are still far from being solved (see section 1.1). As already seen in section I1.3, variations in pose and expression along with occlusions are among the most difficult challenges due to their significant influence on facial appearance. Moreover, applications of face analysis are mainly based on image sequences. So far, image-based methods cannot use the motion information from image sequences nor adapt their models online to make it person-specific. In the following sections, it is shown how temporal information can be beneficial to landmark detection, especially under uncontrolled conditions, by addressing many of its challenges. To this end, temporal approaches are reviewed from the most naive but popular tracking by detection to the most recent and more complex ones based on DNNs.

### 1.2.1. *Tracking by detection*

Recently, a comparative analysis of video-based facial landmark detection approaches showed that the most popular strategy for this problem is tracking by

detection (Shen et al. 2015). As illustrated in Figure 1.18 tracking by detection consists of applying face detection followed by facial landmark detection, independently on each frame, without taking into account the coherence of adjacent frames.

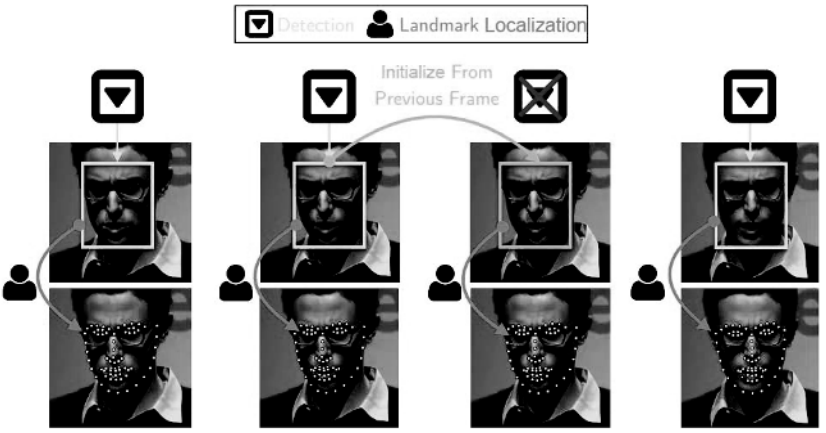


**Figure 1.18.** *Tracking by detection over a few frames. It consists of applying face detection followed by facial landmark detection, independently on each frame, without taking into account the coherence of adjacent frames. Detection from previous frame is used in case of failure. Source: Chrysos et al. (2018a), CC BY (http://creativecommons.org/licenses/by/4.0/), used without modification. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Face detection can possibly return false positives. To prevent these, predictions with the highest confidence are usually retained. The detector may also fail to detect any faces on an image, making facial landmark detection no longer possible. In this case, the bounding boxes from the previous image can possibly be used. This is a reasonable assumption if the frame rate is high enough to ensure that there is no significant variation between the current and the previous images.

Although tracking by detection is naive, a recent study has demonstrated its viability for deformable face tracking (Chrysos et al. 2018a) since current detectors are effective enough to operate under uncontrolled conditions (Zafeiriou et al. 2015). However, this strategy has severe limitations. It relies on two distinct static models that are individually trained, which is time consuming. Neither of these two models are able to integrate motion information and personalize the models online to the targeted person. Additionally, facial landmark detection being sensitive to initialization, it is very likely to obtain poor predictions due to bad initializations far from the ground truth, especially during large variations, for example, pose or expression.

### 1.2.2. *Box, landmark and pose tracking*

To avoid using face detection at each frame with the related initialization issues, an alternative is to use a substitute for the detection in order to adapt to the temporal domain. Three main solutions have been proposed to establish a correlation between the previous and current frames (Yang et al. 2015b): box, landmark and pose tracking. In each of these solutions, the aim is generally to take advantage of the previous prediction to provide a better initialization for the current frame.

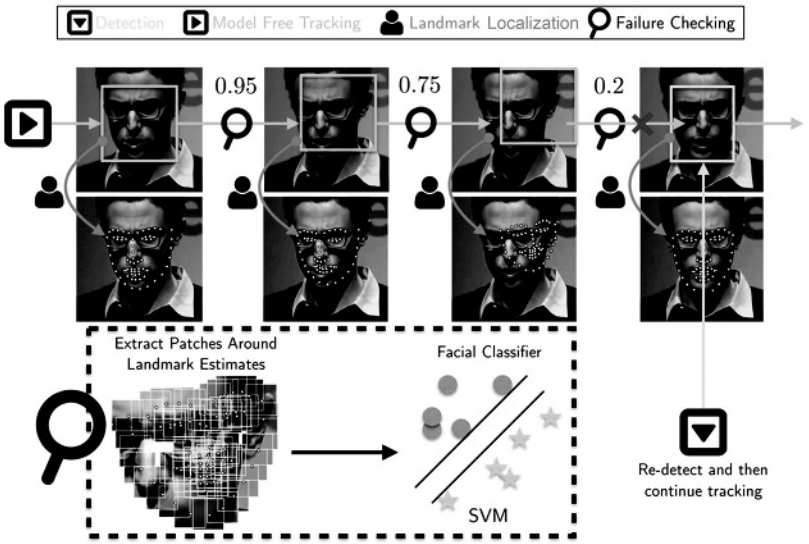


**Figure 1.19.** *Box tracking over a few frames based on a rigid tracking algorithm instead of a face detector in each frame. A reset mechanism is used in case of drift. Source: Chrysos et al. (2018a), CC BY (http://creativecommons.org/licenses/by/4.0/), used without modification. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Box tracking consists of using a rigid tracking algorithm instead of a face detector in each frame (see Figure 1.19). Such algorithms are able to be robust to some variations in the facial appearance during tracking (Yang et al. 2015b; Chrysos et al. 2018a). However, it can be as time consuming as detection, and more importantly, it can easily drift over time, especially under uncontrolled conditions. This solution is found to be equivalent to tracking by detection with the same drawbacks (Chrysos et al. 2018a). The second solution, landmark tracking, involves the use of the previous shape as an initialization for the current frame (Yang et al. 2015b). It provides the ability to inherit information about rotation, translation and scale, and thus provide an initialization closer to the ground truth and potentially a better convergence and

more accurate predictions. It might, however, lead to a poor convergence due to cumulative errors in video, when the initialization is not the same as in the training set. A third solution is to keep only the similarity transformation parameters from the previous prediction, i.e. pose tracking (Yang et al. 2015b; Shin and Kim 2016) as shown in Figure 1.20. These parameters are used to adjust the mean shape, which then serves as an initialization. By tracking the pose and not the full shape, the noise from the previous prediction is smoothed, making the following predictions more stable over time. It can greatly reduce the failure rate while improving overall performance (Shen et al. 2015). In the same vein but less popular, ensemble learning algorithms can be used to generate multiple initializations for an image, based on the shape parameters from previous predictions (Peng et al. 2015; Li et al. 2017). This can be summarized as applying facial landmark detection using the different initializations and then recovering the best result. The obvious disadvantage of this approach is its efficiency; a balance between computational efficiency and accuracy has to be found. Beyond initialization, the shape parameters of the previous prediction can also be used to select a pose-dependent model to obtain a better and more efficient convergence, as explained in section 1.1.4 (Xiong and De la Torre 2015; Yang et al. 2015b).

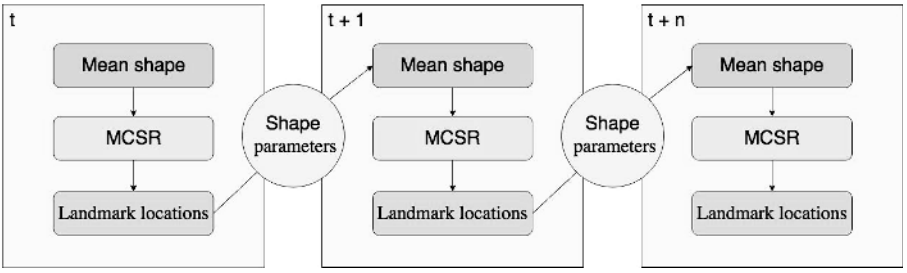


**Figure 1.20.** *Overview of pose tracking. The mean shape is adjusted based on the similarity transformation parameters from the previous frame (Yang et al. 2015b). For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Regardless of the tracking strategy, it is necessary to incorporate a reinitialization mechanism in order to avoid any drift over time. Such a mechanism is illustrated in Figure 1.19. It is generally based on the quality of the prediction, using SVMs (Yang et al. 2015b; Chrysos et al. 2018a) or more recently DNNs (Peng et al. 2016b) to decide if the shape and appearance still match those of a face. The classifier is trained with patches extracted from the ground truth for positive examples, and negative cases are randomly sampled from the region around the ground truth. It can also be done in a holistic way by concatenating the face image and the landmark heatmap. The score of the classifier is then used as a fitting score to judge the quality of the predictions. If the score is below a certain threshold, face detection is applied instead of tracking.

Non-rigid tracking produces a more accurate fitting than tracking by detection. It takes advantage of adjacent frames to improve initialization. Facial landmark detection is therefore no longer dependent on the detection window, but rather takes advantage of the previous prediction to improve initialization. These solutions remain trivial and still do not benefit from person-specific information.

### 1.2.3. *Adaptive approaches*

In contrast to the work presented in section 1.2.2, which retains one generic model after learning, a more advanced approach consists of integrating an online model update, also called incremental learning, in order to make it person-specific and thus more accurate (Asthana et al. 2014; Peng et al. 2015; Sánchez-Lozano et al. 2016; Chrysos et al. 2018b). This is close to rigid tracking algorithms but, instead of considering the object as a single bounding box, its shape and deformations are taken into account along with its appearance. As shown in Figure 1.21, the pre-trained model is updated over time to learn a person-specific representation without re-training from scratch. Incremental learning can be achieved using generative approaches as proposed with the incremental PSs (Chrysos and Zafeiriou 2018; Chrysos et al. 2018b) as well as apply to discriminative approaches as with incremental SDM (Asthana et al. 2014). These approaches are generally costly. Therefore, efficient methods based on continuous regression, as opposed to standard discrete regression, have been proposed and are able to reach real-time performance (Sánchez-Lozano et al. 2016). To avoid compromising the model during drifting, an update criterion similar to the reinitialization mechanisms presented in section 1.2.2 is generally used.

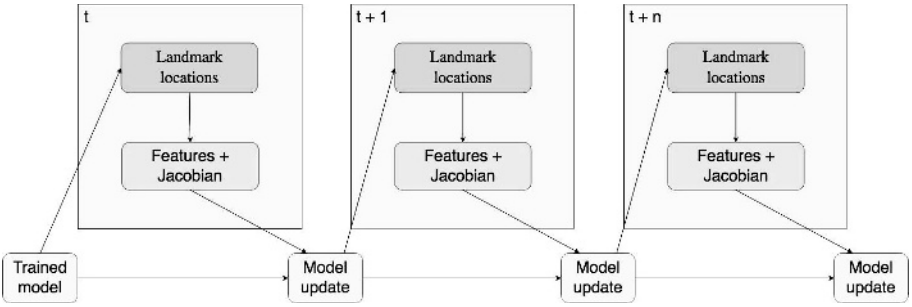


**Figure 1.21.** *CSR with on-line model update (Sánchez-Lozano et al. 2016). The pre-trained model is updated over time to learn a person-specific representation without re-training from scratch. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Adaptive approaches provide more accurate predictions but, as with generic tracking, they are prone to drifting, while tracking by detection is drift-free but less

accurate. It is possible to establish a synergy between both detection and tracking in order to limit the weaknesses of these two approaches. This is known as joint detection and tracking.

### 1.2.4. *Joint approaches*

To avoid drift issues and provide accurate predictions, both the tracking-by-detection approach and the conventional tracking approach may be combined to complement each other (Khan et al. 2017; Guo et al. 2018b). One of the first solutions to be proposed relies on the use of different detection and tracking initializations (Khan et al. 2017). Each of these initializations define different optimization sub-problems. Independent shape updates for each sub-problem are computed and then coupled using global variable consensus optimization (Boyd et al. 2011).

Using equation [1.3] of the generative model (section 1.1.1), global variable consensus can be formulated as follows:

$$\underset{p_1 \dots p_M, q_1 \dots q_M}{\arg\min} \sum_{m=1}^{M} f_m(p_m, q_m) \qquad [1.17]$$

where $M$, $f$, $p$ and $q$ correspond respectively to the different initializations, sub-problems, shape and appearance parameters. This strategy can be integrated within any state-of-the-art approach. However, the synergy between both tasks is limited due to the use of separate models, one for landmark detection and another for tracking. The two tasks are not optimized together and are actually applied subsequently.

DRL (Littman 2015) has been proposed to explicitly exploit the synergy between bounding box generation and landmark detection (Guo et al. 2018b). Two agents, a tracking agent and an alignment agent, are trained interactively (see Figure 1.22). The aim of the tracking agent is to adjust the bounding box based on a set of actions corresponding to displacements or changes in scale. The role of the alignment agent is essentially to determine whether or not it is necessary to continue the search process using continue/stop actions. Both tasks are jointly conducted through the sequence of action defined by the two agents with an emphasis on the quality of the alignment.

Joint approaches are able to benefit from both detection and tracking while avoiding their downsides. However, they do not take advantage of the dynamic nature of the face. Other information, such as the trajectories of the landmarks through the image sequence, seems relevant to consider.

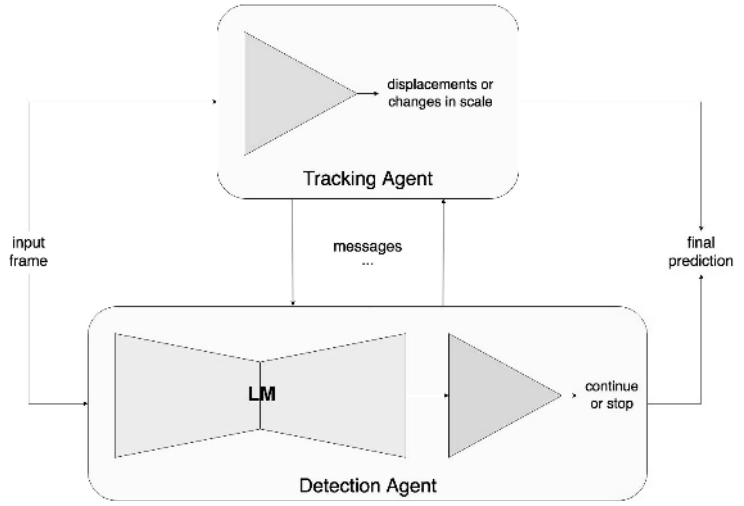


**Figure 1.22.** *Joint detection and tracking using deep reinforcement learning (Guo et al. 2018b). A tracking agent adjusts the bounding box based on a set of actions corresponding to displacements or changes in scale. An alignment agent determines whether or not it is necessary to continue the search process using continue/stop actions. The two agents are trained interactively. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

### 1.2.5. *Temporal constrained approaches*

Whether explicit or implicit, the shape constraints present in most facial landmark detection methods are crucial to obtaining good performance under uncontrolled conditions. In image sequences, an additional constraint may be applied to the trajectories of the landmarks. One simple way to achieve this is to smooth predictions between frames to avoid jittering and increase stability. Bayesian filters such as Kalman filters and their extensions (e.g. particle filters) can be used for this purpose. However, they require a complex design and tuning phase and are generally restricted to global rigid motion tracking, i.e. tracking the detected bounding box (Prabhu et al. 2010; Uřičář and Franc 2015) and ignoring local non-rigid facial deformations. It has also been demonstrated that these filters only provide a marginal gain for video-based facial landmark detection (Gu et al. 2017; Chrysos et al. 2018a).

To consider the whole shape, a stabilization model based on a Gaussian mixture has been proposed (Tai et al. 2019). The model is trained on a set of videos to learn the statistics of different kinds of movements. A loss function with two terms to address time delay and non-smooth issues is used to estimate its parameters. The model is then applied after landmark detection; it takes as an input all predictions from previous

images and produces a more accurate and stable result for the current image. A balance has to be found between accuracy and stability.

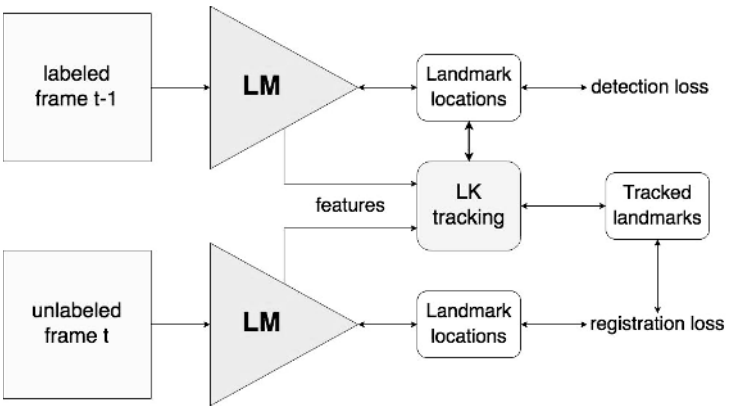


**Figure 1.23.** *Semi-supervised training using the coherency of optical flow as a source of supervision (Dong et al. 2018). Lucas–Kanade tracker tracks landmarks on future frames based on past predictions from the detector. The registration loss computes the distance between these predictions and those from the detector. This allows us to train a detector with unlabeled videos to provide temporally consistent predictions. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

Another way to reduce jittering in video is to take advantage of the coherency of optical flow by using it as a source of supervision (Dong et al. 2018). The main advantage is that no manual labeling is required. The Lucas–Kanade tracker coupled with a registration loss function can be used to train a detector with unlabeled videos to provide temporally consistent predictions. The latter tracks landmarks on future frames based on past predictions from the detector. The registration loss computes the distance between these predictions and those from the detector. The whole training procedure is illustrated in Figure 1.23. Besides jittering, optical flow can also be used to improve robustness in motion-blurred videos (Sun et al. 2019). Although optical flow registration can encourage temporal consistency to improve existing facial landmark detectors, they still remain image-based detectors with limited motion modeling capability.

RNNs are particularly suitable for time series analysis. They have proved to be effective for facial landmark detection using a CNN as a backbone to jointly estimate and track visual features over time (Peng et al. 2016a; Gu et al. 2017; Hou et al. 2018). When training is done in conjunction with a CNN, optimal features can be extracted and temporal dependencies at different scales can be captured. As a result, RNNs help stabilize predictions over time and improve robustness under uncontrolled conditions (Gu et al. 2017). It is possible to start from a pre-trained CNN and transform any

pre-trained feed-forward layer into a recurrent layer (Yang et al. 2018). As depicted in Figure 1.24, spatial and temporal information processing can also be decoupled in order to explicitly exploit their complementarity (Liu et al. 2018a). The spatial stream aims to preserve the holistic facial shape structure, while the temporal stream focuses on temporal consistency to improve shape refinements. A weighted fusion of both streams produces the final prediction.

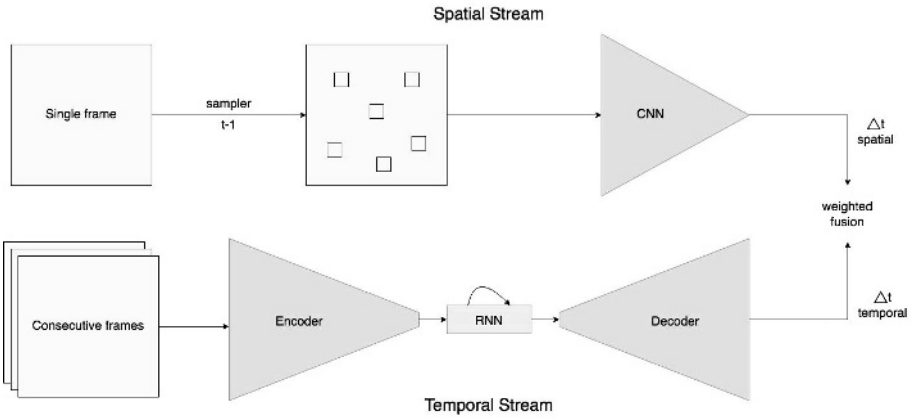


**Figure 1.24.** *Two-stream network which decomposes the video input to the spatial and temporal streams (Liu et al. 2018a). The spatial stream aims to preserve the holistic facial shape structure, while the temporal stream focuses on temporal consistency to improve shape refinements. A weighted fusion of both streams produces the final prediction. For a color version of this figure, see www.iste.co.uk/belmonte/face.zip*

However, these approaches only provide late temporal connectivity on small feature maps with a possibly high level of abstraction at the level of the recurrent layer. This layer is then able to compute global motion features (e.g. head motion) but may not be capable of accurately detecting local motion (e.g. eye and lip movements) due to the fact that CNNs cannot model any motion.

### 1.2.6. *Summary*

Existing solutions for video-based facial landmark detection have been reviewed in this section. The first observation is that even today, one of the most popular strategies for this problem is tracking by detection. Despite being considered viable, this strategy remains limited since it does not leverage temporal information at all.

More sophisticated solutions have been proposed. They take advantage of conventional tracking and adaptive learning to improve accuracy. Although these solutions are more accurate, they are likely to drift over time and thus require a reset

mechanism. Moreover, they do not sufficiently make use of the temporal information as they generally only use the prediction on the adjacent frames.

Since tracking is more accurate than detection but prone to drift and detection is drift-free, they have been coupled in a joint approach. This solution allows the advantages of both approaches to be exploited more effectively but it still makes limited use of temporal information.

Similar to the shape constraint present in most image-based approaches, several solutions have also been proposed to add temporal constraint in order to increase stability over time, using either 2D models with specific training strategy or RNNs. Due to their nature, temporal-constrained 2D models are reduced to smoothing only. Regarding RNNs, the most recent and advanced methods suffer especially from limited temporal connectivity since it relies on the use of a CNN as a backbone. However, the CNN is unable to capture temporal information, which only allows us to track high-level features (i.e. global motion). Fine motion cannot be easily modeled while they are just as relevant for landmark detection. Most of the proposed approaches are based on tracking or temporal smoothing. The main weakness of all these methods is that they do not fully exploit the dynamic nature of the face. Taking this dynamic into account appears crucial to ensure high accuracy and robustness under uncontrolled conditions, especially when the appearance is no longer reliable.

There are many open questions. How can motion be modeled with more detail and integrated into current approaches based on DL? How to combine local and global motion while taking advantage of their complementarity? The same question also arises regarding spatial and temporal information. Finally, which level of time dependency is necessary for a problem like facial landmark detection?

## 1.3. Discussion

Throughout the years, research has always been focused on static images. The approaches proposed so far can be grouped into two major categories: generative and discriminative. The fundamental differences between these two categories of approach include:

– model fitting based on coefficient prediction or mapping between the image and the shape;

– explicit shape model or implicitly embedded shape constraints;

– independent landmark prediction or joint prediction;

– holistic or part-based representation;

– one-step prediction or coarse-to-fine strategy;

– the need of a shape initialization.

Many solutions have also been developed to address specific challenges and can be integrated into any type of approach. They are based on techniques including 3D model (Zhu et al. 2017), multi-view model(s) (Deng et al. 2019), multi-task learning (Zhang et al. 2016), GANs (Bulat and Tzimiropoulos 2018) and explicit occlusions modeling (Yuen and Trivedi 2017). However, these solutions can make dataset annotation and model training much more complex.

Discriminative approaches perform landmark prediction through a mapping between the image and the shape. This makes them faster and more accurate than the approaches based on generative models (Kazemi and Sullivan 2014; Ren et al. 2014). Landmarks are jointly predicted, which implicitly integrates the shape constraints. The mapping is generally implemented using a cascade of regressors, coarse-to-fine strategy, since direct mapping is hard to achieve. Among the discriminative approaches, DL ones have gained increasing popularity (Bulat and Tzimiropoulos 2017b). They allow discriminative and problem-specific feature learning, as opposed to handcrafted features used in traditional approaches. The latter are generic and likely to be suboptimal for facial landmark detection (LeCun et al. 2015). With DNNs, feature extraction and regression are trained jointly. The relationship between images and shapes is modeled more effectively. DL has notably been successful with direct mapping, showing that DR remains a relevant solution (Merget et al. 2018; Miao et al. 2018).

Current approaches still encounter difficulties under uncontrolled conditions (Sagonas et al. 2016). Although most of the challenges have been addressed, this is often achieved with solutions specific to one or a few challenges. There is a lack of approaches that can address all the difficulties at once. With regard to the literature, temporal approaches appear to be the most promising direction. Today, due to the increase in video data and their potential benefits, the interest in temporal approaches is growing (Shen et al. 2015). Currently, most landmark detection approaches are unable to take advantage of the temporal information. They are often applied in tracking-by-detection manner. More advanced strategies have recently been proposed (Chrysos et al. 2018a). They mainly rely on tracking (e.g. box, landmark, pose) along with incremental learning and for a smaller part on temporal smoothing (Dong et al. 2018; Tai et al. 2019). However, these approaches are subject to drift and so far do not fully exploit facial dynamics. The most advanced ones based on RNNs appear to be limited to the global movements of the head (Gu et al. 2017; Hou et al. 2018; Liu et al. 2018a). The CNNs currently used as a backbone cannot model any motion.

Moreover, given the overall progress accomplished, more attention should be paid to evaluation protocols. There is a lack of comprehensive analysis to quantify the performance of current approaches according to the different difficulties (Zafeiriou et al. 2017). Regarding temporal approaches, the contribution of the temporal information according to the difficulties, and compared to static approaches, is not well quantified. This would help us to better identify the remaining efforts to be made

but could also be useful for anyone who uses landmarks as a pre-processing in an application (e.g. expression recognition).

Consequently, the issues addressed in this part of the book are related to the evaluation protocols and temporal approaches. The first goal is to quantify the impact of major challenges of landmark detection of the one hand and, on the other, to quantify the contribution of temporal approaches. The recent availability of datasets such as SNaP-2DFe (Allaert et al. 2018) presents an opportunity to address these issues. The second goal is to better exploit the dynamic nature of the face in order to obtain more stable predictions over time and more robustness to the full set of variations that considerably impact facial appearance. One hypothesis is that early temporal connectivity could help to model motion more finely as well as complement current approaches. Among the open questions that are also of interest are how to effectively combine facial motion information with the facial appearance. To this end, CNNs appear to be an effective and versatile baseline tool.

## 1.4. References

Allaert, B., Mennesson, J., Bilasco, I.M., Djeraba, C. (2018). Impact of the face registration techniques on facial expressions recognition. *Signal Processing: Image Communication*, 61, 44–53.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S., Asari, V.K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint, arXiv:1803.01164.

Antonakos, E., Alabort-i Medina, J., Zafeiriou, S. (2015). Active pictorial structures. *IEEE Conference on Computer Vision and Pattern Recognition*, 5435–5444.

Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. *IEEE Conference on Computer Vision and Pattern Recognition*, 3444–3451.

Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M. (2014). Incremental face alignment in the wild. *IEEE Conference on Computer Vision and Pattern Recognition*, 1859–1866.

Baltrusaitis, T., Robinson, P., Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. *IEEE International Conference on Computer Vision Workshops*, 354–361.

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2930–2940.

Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *Transactions on Neural Networks*, 5(2), 157–166.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M. (2007). Generative or discriminative? Getting the best of both worlds. *Bayesian Statistics*, 8(3), 3–24.

Bhagavatula, C., Zhu, C., Luu, K., Savvides, M. (2017). Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses. *IEEE International Conference on Computer Vision*, 3980–3989.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 187–194.

Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M. (2010). Visual object tracking using adaptive correlation filters. *IEEE Conference on Computer Vision and Pattern Recognition*, 2544–2550.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1–122.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Bulat, A. and Tzimiropoulos, G. (2017a). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *IEEE International Conference on Computer Vision*, arXiv:1703.00862.

Bulat, A. and Tzimiropoulos, G. (2017b). How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). *IEEE International Conference on Computer Vision*, 1021–1030.

Bulat, A. and Tzimiropoulos, G. (2018). Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. *IEEE Conference on Computer Vision and Pattern Recognition*, 109–117.

Bulat, A., Sanchez, E., Tzimiropoulos, G. (2021). Subpixel heatmap regression for facial landmark localization. *British Machine Vision Conference*, arXiv:2111.02360.

Burgos-Artizzu, X.P., Perona, P., Dollár, P. (2013). Robust face landmark estimation under occlusion. *IEEE International Conference on Computer Vision*, 1513–1520.

Cao, X., Wei, Y., Wen, F., Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2), 177–190.

Chandran, P., Bradley, D., Gross, M., Beeler, T. (2020). Attention-driven cropping for very high resolution facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 5861–5870.

Chen, X., Zhou, E., Mo, Y., Liu, J., Cao, Z. (2017). Delving deep into coarse-to-fine framework for facial landmark localization. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 142–149.

Chen, L., Su, H., Ji, Q. (2019a). Deep structured prediction for facial landmark detection. *Neural Information Processing Systems*, 32, 2450–2460.

Chen, L., Su, H., Ji, Q. (2019b). Face alignment with kernel density deep neural network. *IEEE International Conference on Computer Vision*, 6992–7002.

Chen, Y., Shen, C., Wei, X.-S., Liu, L., Yang, J. (2019c). Adversarial learning of structure-aware fully convolutional networks for landmark localization. *Transactions on Pattern Analysis and Machine Intelligence*, 42(7), 1654–1669.

Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint, arXiv: 1409.1259.

Chrysos, G.G. and Zafeiriou, S. (2018). PD2T: Person-specific detection, deformable tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11), 2555–2568.

Chrysos, G.G., Antonakos, E., Snape, P., Asthana, A., Zafeiriou, S. (2018a). A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, 126(2–4), 198–232.

Chrysos, G.G., Antonakos, E., Zafeiriou, S. (2018b). IPST: Incremental pictorial structures for model-free tracking of deformable objects. *Transactions on Image Processing*, 27(7), 3529–3540.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J. (1995). Active shape models – Their training and application. *Computer Vision and Image Understanding*, 61(1), 38–59.

Cootes, T.F., Walker, K., Taylor, C.J. (2000). View-based active appearance models. *IEEE Conference on Automatic Face and Gesture Recognition*, 227–232.

Cootes, T.F., Edwards, G.J., Taylor, C.J. (2001). Active appearance models. *Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 681–685.

Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P. (2012). Robust and accurate shape model fitting using random forest regression voting. *European Conference on Computer Vision*, 278–291.

Cristinacce, D. and Cootes, T.F. (2006). Feature detection and tracking with constrained local models. *British Machine Vision Conference*, September 4–7.

Cristinacce, D. and Cootes, T.F. (2007). Boosted regression active shape models. *British Machine Vision Conference*, 2, 880–889.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. (2017). Deformable convolutional networks. *IEEE International Conference on Computer Vision*, 764–773.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 886–893.

Dantone, M., Gall, J., Fanelli, G., Van Gool, L. (2012). Real-time facial feature detection using conditional regression forests. *IEEE Conference on Computer Vision and Pattern Recognition*, 2578–2585.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, J., Liu, Q., Yang, J., Tao, D. (2016). M3 CSR: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47, 19–26.

Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., Zafeiriou, S. (2019). The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6), 599–624.

Dollár, P., Welinder, P., Perona, P. (2010). Cascaded pose regression. *IEEE Conference on Computer Vision and Pattern Recognition*, 1078–1085.

Dong, X., Yan, Y., Ouyang, W., Yang, Y. (2018a). Style aggregated network for facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 379–388.

Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., Sheikh, Y. (2018b). Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. *IEEE Conference on Computer Vision and Pattern Recognition*, 360–368.

Ekman, P. and Keltner, D. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, 8(4), 151–158.

Fan, H. and Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47, 27–35.

Felleman, D.J. and Van, D.E. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.

Felzenszwalb, P.F. and Huttenlocher, D.P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 55–79.

Feng, Z.-H., Kittler, J., Christmas, W., Huber, P., Wu, X.-J. (2017). Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. *IEEE Conference on Computer Vision and Pattern Recognition*, 3681–3690.

Feng, Z.-H., Kittler, J., Awais, M., Huber, P., Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, arXiv:1711.06753.

Ghiasi, G. and Fowlkes, C.C. (2014). Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. *IEEE Conference on Computer Vision and Pattern Recognition*, 2385–2392.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *Neural Information Processing Systems*, 2672–2680.

Gower, J.C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1), 33–51.

Gross, R., Matthews, I., Baker, S. (2005). Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12), 1080–1093.

Gu, J., Yang, X., De Mello, S., Kautz, J. (2017). Dynamic facial analysis: From Bayesian filtering to recurrent neural network. *IEEE Conference on Computer Vision and Pattern Recognition*, 1548–1557.

Guo, J., Deng, J., Xue, N., Zafeiriou, S. (2018a). Stacked dense U-nets with dual transformers for robust face alignment. *BMVC*, 50.

Guo, M., Lu, J., Zhou, J. (2018b). Dual-agent deep reinforcement learning for deformable face tracking. *European Conference on Computer Vision*, 768–783.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, Z., Kan, M., Zhang, J., Chen, X., Shan, S. (2017). A fully end-to-end cascaded cnn for facial landmark detection. *IEEE Conference on Automatic Face and Gesture Recognition*, 200–207.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J. (2018). Improving landmark localization with semi-supervised learning. *Computer Vision and Pattern Recognition*, 1546–1555.

Hou, Q., Wang, J., Bai, R., Zhou, S., Gong, Y. (2018). Face alignment recurrent network. *Pattern Recognition*, 74, 448–458.

Hsu, G.-S., Chang, K.-H., Huang, S.-C. (2015). Regressive tree structured model for facial landmark localization. *IEEE International Conference on Computer Vision*, 3855–3861.

Huang, Y., Liu, Q., Metaxas, D. (2007). A component based deformable model for generalized face alignment. *IEEE International Conference on Computer Vision*, 1–8.

Huang, X., Deng, W., Shen, H., Zhang, X., Ye, J. (2020). Propagationnet: Propagate points to curve to learn structure information. *IEEE Conference on Computer Vision and Pattern Recognition*, 7265–7274.

Huang, Y., Yang, H., Li, C., Kim, J., Wei, F. (2021). ADNet: Leveraging error-bias towards normal direction in face alignment. *International Conference on Computer Vision*, 3080–3090.

Hubel, D.H. and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. arXiv preprint arXiv:1602.07360.

Jin, X. and Tan, X. (2016). Face alignment by robust discriminative Hough voting. *Pattern Recognition*, 60, 318–333.

Jin, X. and Tan, X. (2017). Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding*, 162, 1–22.

Jin, H., Liao, S., Shao, L. (2021). Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 1–21.

Jourabloo, A. and Liu, X. (2017). Pose-invariant face alignment via CNN-based dense 3D model fitting. *International Journal of Computer Vision*, 124(2), 187–203.

Jourabloo, A., Ye, M., Liu, X., Ren, L. (2017). Pose-invariant face alignment with a single CNN. *International Conference on Computer Vision*, 3200–3209.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874.

Khan, M.H., McDonagh, J., Tzimiropoulos, G. (2017). Synergy between face alignment and tracking via discriminative global consensus optimization. *IEEE International Conference on Computer Vision*, 3811–3819.

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kowalski, M., Naruniec, J., Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. *Computer Vision and Pattern Recognition Workshops*, arXiv:1706.01789.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 1097–1105.

Kumar, A. and Chellappa, R. (2018). Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 430–439.

Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., Yan, S. (2018). Deep recurrent regression for facial landmark detection. *Transactions on Circuits and Systems for Video Technology*, 28(5), 1144–1157.

Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D. (1997). Face recognition: A convolutional neural-network approach. *Transactions on Neural Networks*, 8(1), 98–113.

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D. (1990). Handwritten digit recognition with a back-propagation network. *Neural Information Processing Systems*, 396–404.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436.

Lee, D., Park, H., Yoo, C.D. (2015). Face alignment using cascade Gaussian process regression trees. *IEEE Conference on Computer Vision and Pattern Recognition*, 4204–4212.

Li, Y., Wang, S., Zhao, Y., Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *Transactions on Image Processing*, 22(7), 2559–2573.

Li, C.Y., Baltrušaitis, T., Morency, L.-P. (2017). Constrained ensemble initialization for facial landmark tracking in video. *Automatic Face and Gesture Recognition*, 697–704.

Littman, M.L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553), 445.

Liu, H., Lu, J., Feng, J., Zhou, J. (2018a). Two-stream transformer networks for video-based face alignment. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11), 2546–2554.

Liu, H., Lu, J., Guo, M., Wu, S., Zhou, J. (2018b). Learning reasoning-decision networks for robust face alignment. *Transactions on Pattern Analysis and Machine Intelligence*, 42(3), 679–693.

Liu, Z., Zhu, X., Hu, G., Guo, H., Tang, M., Lei, Z., Robertson, N.M., Wang, J. (2019). Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 3467–3476.

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Lucey, S., Wang, Y., Cox, M., Sridharan, S., Cohn, J.F. (2009). Efficient constrained local model fitting for non-rigid face alignment. *Image and Vision Computing*, 27(12), 1804–1813.

Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X. (2017). A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 3691–3700.

Ma, J., Li, J., Du, B., Wu, J., Wan, J., Xiao, Y. (2022). Robust face alignment by dual-attentional spatial-aware capsule networks. *Pattern Recognition*, 122, 108297.

Martinez, B., Valstar, M.F., Binefa, X., Pantic, M. (2013). Local evidence aggregation for regression-based facial point detection. *Transactions on Pattern Analysis and Machine Intelligence*, 35(5), 1149–1163.

Martins, P., Caseiro, R., Batista, J. (2014). Non-parametric Bayesian constrained local models. *IEEE Conference on Computer Vision and Pattern Recognition*, 1797–1804.

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, 60(2), 135–164.

Merget, D., Rock, M., Rigoll, G. (2018). Robust facial landmark detection via a fully-convolutional local-global context network. *IEEE Conference on Computer Vision and Pattern Recognition*, 781–790.

Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V., Huang, H. (2018). Direct shape regression networks for end-to-end face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 5040–5049.

Nair, V. and Hinton, G.E. (2010). Rectified linear units improve restricted Boltzmann machines. *International Conference on Machine Learning*, 807–814.

Newell, A., Yang, K., Deng, J. (2016). Stacked hourglass networks for human pose estimation. *European Conference on Computer Vision*, 483–499.

Papandreou, G. and Maragos, P. (2008). Adaptive and constrained algorithms for inverse compositional active appearance model fitting. *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Pascanu, R., Mikolov, T., Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, 1310–1318.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.

Peng, X., Zhang, S., Yang, Y., Metaxas, D.N. (2015). PIEFA: Personalized incremental and ensemble face alignment. *IEEE International Conference on Computer Vision*, 3880–3888.

Peng, X., Feris, R.S., Wang, X., Metaxas, D.N. (2016a). A recurrent encoder-decoder network for sequential face alignment. *European Conference on Computer Vision*, 38–56.

Peng, X., Hu, Q., Huang, J., Metaxas, D.N. (2016b). Track facial points in unconstrained videos. arXiv preprint arXiv:1609.02825.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large-Margin Classifiers*, 10(3), 61–74.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4), 344–350.

Prabhu, U., Seshadri, K., Savvides, M. (2010). Automatic facial landmark tracking in video sequences using Kalman filter assisted active shape models. *European Conference on Computer Vision*, 86–99.

Qian, S., Sun, K., Wu, W., Qian, C., Jia, J. (2019). Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. *International Conference on Computer Vision*, 10153–10163.

Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R. (2017). An all-in-one convolutional neural network for face analysis. *Automatic Face and Gesture Recognition*, 17–24.

Ren, S., Cao, X., Wei, Y., Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. *IEEE Conference on Computer Vision and Pattern*, 1685–1692.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Robinson, J.P., Li, Y., Zhang, N., Fu, Y., Tulyakov, S. (2019). Laplace landmark localization. *International Conference on Computer Vision*, 10103–10112.

Rosenblatt, F. (1988). The perception: A probabilistic model for information storage and organization in the brain. In *Neurocomputing: Foundations of Research*, Anderson, J.A. and Rosenfeld, E. (eds). MIT Press, Cambridge.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1985). Learning internal representations by error propagation. Technical report, University of California San Diego, La Jolla Institute for Cognitive Science.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3), 1.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47, 3–18.

Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G., Valstar, M. (2016). Cascaded continuous regression for real-time incremental face tracking. *European Conference on Computer Vision*, 645–661.

Saragih, J. and Goecke, R. (2007). A nonlinear discriminative approach to AAM fitting. *International Conference on Computer Vision*, 1–8.

Saragih, J.M., Lucey, S., Cohn, J.F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2), 200–215.

Shao, Z., Liu, Z., Cai, J., Ma, L. (2018). Deep adaptive attention for joint facial action unit detection and face alignment. *European Conference on Computer Vision*, 705–720.

Shapira, G., Levy, N., Goldin, I., Jevnisek, R.J. (2021). Knowing when to quit: Selective cascaded regression with patch attention for real-time face alignment. *International Conference on Multimedia*, 2372–2380.

Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. *International Conference on Computer Vision Workshops*, 50–58.

Shin, J. and Kim, D. (2016). Robust face alignment and tracking by combining local search and global fitting. *Image and Vision Computing*, 51, 69–83.

Silverman, B.W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge, Boca Raton.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smith, B.M., Brandt, J., Lin, Z., Zhang, L. (2014). Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization. *IEEE Conference on Computer Vision and Pattern Recognition*, 1741–1748.

Su, K. and Geng, X. (2019). Soft facial landmark detection by label distribution learning. *AAAI Conference on Artificial Intelligence*, 5008–5015.

Sun, Y., Wang, X., Tang, X. (2013). Deep convolutional network cascade for facial point detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 3476–3483.

Sun, K., Wu, W., Liu, T., Yang, S., Wang, Q., Zhou, Q., Ye, Z., Qian, C. (2019). Fab: A robust facial landmark detection framework for motion-blurred videos. *International Conference on Computer Vision*, 5462–5471.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Tai, Y., Liang, Y., Liu, X., Duan, L., Li, J., Wang, C., Huang, F., Chen, Y. (2019). Towards highly accurate and stable face alignment for high-resolution videos. *AAAI Conference on Artificial Intelligence*, arXiv:1811.00342.

Tong, Y., Wang, Y., Zhu, Z., Ji, Q. (2007). Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11), 3195–3208.

Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 4177–4187.

Tzimiropoulos, G. (2015). Project-out cascaded regression with an application to face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 3659–3667.

Tzimiropoulos, G. and Pantic, M. (2013). Optimization problems for fast AAM fitting in-the-wild. *International Conference on Computer Vision*, 593–600.

Tzimiropoulos, G. and Pantic, M. (2014). Gauss–Newton deformable part models for face alignment in-the-wild. *IEEE Conference on Computer Vision and Pattern Recognition*, 1851–1858.

Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., Pantic, M. (2012). Generic active appearance models revisited. *Asian Conference on Computer Vision*, 650–663.

Uřičář, M. and Franc, V. (2015). Real-time facial landmark tracking by tree-based deformable part model based detector. *IEEE International Conference on Computer Vision Workshops*, December.

Uřičář, M., Franc, V., Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. *Computer Vision Theory and Applications*, 12, 547–556.

Valstar, M., Martinez, B., Binefa, X., Pantic, M. (2010). Facial point detection using boosted regression and graph models. *IEEE Conference on Computer Vision and Pattern Recognition*, 2729–2736.

Wan, J., Lai, Z., Liu, J., Zhou, J., Gao, C. (2020). Robust face alignment by multi-order high-precision hourglass network. *IEEE Transactions on Image Processing*, 30, 121–133.

Wang, W., Tulyakov, S., Sebe, N. (2018). Recurrent convolutional shape regression. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11), 2569–2582.

Wang, X., Bo, L., Fuxin, L. (2019). Adaptive wing loss for robust face alignment via heatmap regression. *IEEE International Conference on Computer Vision*, 6971–6981.

Wu, Y. and Ji, Q. (2015a). Discriminative deep face shape model for facial point detection. *International Journal of Computer Vision*, 113(1), 37–53.

Wu, Y. and Ji, Q. (2015b). Robust facial landmark detection under significant head poses and occlusion. *IEEE International Conference on Computer Vision*, 3658–3666.

Wu, Y. and Ji, Q. (2016). Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 3400–3408.

Wu, Y. and Ji, Q. (2018). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2), 115–142.

Wu, W. and Yang, S. (2017). Leveraging intra and inter-dataset variations for robust face alignment. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 150–159.

Wu, Y., Wang, Z., Ji, Q. (2014). A hierarchical probabilistic model for facial feature detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1781–1788.

Wu, Z., Shen, C., Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90(June), 119–133.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. *European Conference on Computer Vision*, 57–72.

Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S., Kassim, A. (2017). Recurrent 3D-2D dual learning for large-pose facial landmark detection. *IEEE International Conference on Computer Vision*, 1633–1642.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 532–539.

Xiong, X. and De la Torre, F. (2015). Global supervised descent method. *IEEE Conference on Computer Vision and Pattern Recognition*, 2664–2673.

Xu, X. and Kakadiaris, I.A. (2017). Joint head pose estimation and face alignment framework using global and local CNN features. *IEEE Conference on Automatic Face and Gesture Recognition*, 642–649.

Xu, Z., Li, B., Yuan, Y., Geng, M. (2021). Anchorface: An anchor-based facial landmark detector across large poses. *AAAI Conference on Artificial Intelligence*, 3092–3100.

Yan, S., Hou, X., Li, S.Z., Zhang, H., Cheng, Q. (2003). Face alignment using view-based direct appearance models. *International Journal of Imaging Systems and Technology*, 13(1), 106–112.

Yan, J., Lei, Z., Yi, D., Li, S. (2013). Learn to combine multiple hypotheses for accurate face alignment. *IEEE International Conference on Computer Vision Workshops*, 392–396.

Yan, Y., Duffner, S., Phutane, P., Berthelier, A., Blanc, C., Garcia, C., Chateau, T. (2021). 2D Wasserstein loss for robust facial landmark detection. *Pattern Recognition*, 116, 107945.

Yang, H. and Patras, I. (2013a). Privileged information-based conditional regression forest for facial feature detection. *IEEE Conference on Automatic Face and Gesture Recognition*, 1–6.

Yang, H. and Patras, I. (2013b). Sieving regression forest votes for facial feature detection in the wild. *IEEE International Conference on Computer Vision*, 1936–1943.

Yang, H., He, X., Jia, X., Patras, I. (2015a). Robust face alignment under occlusion via regional predictive power estimation. *Transactions on Image Processing*, 24(8), 2393–2403.

Yang, J., Deng, J., Zhang, K., Liu, Q. (2015b). Facial shape tracking via spatio-temporal cascade shape regression. *IEEE International Conference on Computer Vision Workshops*, 41–49.

Yang, J., Liu, Q., Zhang, K. (2017). Stacked hourglass network for robust facial landmark localisation. *IEEE Computer Vision and Pattern Recognition Workshops*, 79–87.

Yang, X., Molchanov, P., Kautz, J. (2018). Making convolutional networks recurrent for visual sequence learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 6469–6478.

Yin, S., Wang, S., Chen, X., Chen, E., Liang, C. (2020). Attentive one-dimensional heatmap regression for facial landmark detection and tracking. *ACM International Conference on Multimedia*, 538–546.

Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N. (2013). Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. *IEEE International Conference on Computer Vision*, 1944–1951.

Yu, X., Lin, Z., Brandt, J., Metaxas, D.N. (2014). Consensus of regression for occlusion-robust facial feature localization. *European Conference on Computer Vision*, 105–118.

Yue, L., Miao, X., Wang, P., Zhang, B., Zhen, X., Cao, X. (2018). Attentional alignment networks. *British Machine Vision Conference*, 208.

Yuen, K. and Trivedi, M.M. (2017). An occluded stacked hourglass approach to facial landmark localization and occlusion estimation. *Transactions on Intelligent Vehicles*, 2(4), 321–331.

Zafeiriou, S., Zhang, C., Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138, 1–24.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J. (2017). The Menpo facial landmark localisation challenge: A step towards the solution. *IEEE Computer Vision and Pattern Recognition Workshops*, 2116–2125.

Zeng, J., Liu, S., Li, X., Mahdi, D.A., Wu, F., Wang, G. (2018). Deep context-sensitive facial landmark detection with tree-structured modeling. *Transactions on Image Processing*, 27(5), 2096–2107.

Zhang, J., Shan, S., Kan, M., Chen, X. (2014a). Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. *European Conference on Computer Vision*, 1–16.

Zhang, Z., Luo, P., Loy, C.C., Tang, X. (2014b). Facial landmark detection by deep multi-task learning. *European Conference on Computer Vision*, 94–108.

Zhang, Z., Luo, P., Loy, C.C., Tang, X. (2016). Learning deep representation for face alignment with auxiliary attributes. *Transactions on Pattern Analysis and Machine Intelligence*, 38(5), 918–930.

Zhang, X., Li, Z., Change Loy, C., Lin, D. (2017). Polynet: A pursuit of structural diversity in very deep networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 718–726.

Zhao, X., Kim, T.-K., Luo, W. (2014). Unified face analysis by iterative multi-output random forests. *IEEE Conference on Computer Vision and Pattern Recognition*, 1765–1772.

Zhao, Y., Liu, Y., Shen, C., Gao, Y., Xiong, S. (2020). Mobilefan: Transferring deep hidden representation for face alignment. *Pattern Recognition*, 100, 107114.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. *IEEE International Conference on Computer Vision Workshops*, 386–391.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. *IEEE Conference on Computer Vision and Pattern Recognition*, 2879–2886.

Zhu, S., Li, C., Change Loy, C., Tang, X. (2015). Face alignment by coarse-to-fine shape searching. *IEEE Conference on Computer Vision and Pattern Recognition*, 4998–5006.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z. (2016). Face alignment across large poses: A 3D solution. *IEEE Conference on Computer Vision and Pattern Recognition*, 146–155.

Zhu, X., Liu, X., Lei, Z., Li, S.Z. (2017). Face alignment in full pose range: A 3D total solution. *Transactions on Pattern Analysis and Machine Intelligence*, 41(1), 78–92.

Zhu, M., Shi, D., Zheng, M., Sadiq, M. (2019). Robust facial landmark detection via occlusion-adaptive deep networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 3486–3496.

Zhu, H., Liu, H., Zhu, C., Deng, Z., Sun, X. (2020). Learning spatial-temporal deformable networks for unconstrained face alignment and tracking in videos. *Pattern Recognition*, 107, 107354.