big data and cognitive computing

MDPI

Review

A Review of Facial Landmark Extraction in 2D Images and Videos Using Deep Learning

Matteo Bodini

Department of Computer Science, Università degli Studi di Milano, 20133 Milano, Italy; matteo.bodini@unimi.it

Received: 14 January 2019; Accepted: 4 February 2019; Published: 13 February 2019

Abstract: The task of facial landmark extraction is fundamental in several applications which involve facial analysis, such as facial expression analysis, identity and face recognition, facial animation, and 3D face reconstruction. Taking into account the most recent advances resulting from deep-learning techniques, the performance of methods for facial landmark extraction have been substantially improved, even on in-the-wild datasets. Thus, this article presents an updated survey on facial landmark extraction on 2D images and video, focusing on methods that make use of deep-learning techniques. An analysis of many approaches comparing the performances is provided. In summary, an analysis of common datasets, challenges, and future research directions are provided.

Keywords: facial landmark extraction; deep learning

1. Introduction

The face represents a key component in conveying verbal and non-verbal information during interactions between humans. By looking at the face, humans are able to extract a lot of non-verbal information, like identity, intent, and emotion. In the field of computer vision, to automatically extract such information, the detection of facial landmarks (Figure 1) is usually an important step, and several facial analysis techniques are based on the precise detection of landmarks. Head pose estimation [1], and facial expression recognition [2] algorithms are strongly based on the facial shape information, given by landmark positions. For instance, facial landmarks around the eyes can provide an initial guess of the pupil center positions for eye detection, as well as eye-gaze tracking [3]. In facial recognition, landmarks on a 2D image are usually used along with a 3D head model to frontalize the face and reduce within-subject variability, which improves recognition accuracy [4]. Further, facial information obtained through facial landmarks can improve applications in the fields of human and computer interaction, entertainment, security surveillance, and medical applications [5–8].

Facial landmark detection algorithms seek to identify the locations of the facial key landmark points on facial images or videos. Such key points are the main points that describe the unique location of a facial component (e.g., eye corner), or an interpolated point that connects those dominant points around the facial components, as well as facial contour. Formally, given a facial image denoted as $\mathcal{I}$, a landmark detection algorithm predicts the locations of D landmarks $\mathbf{x} = \{x_1, y_1, x_2, y_2, \ldots, x_D, y_D\}$, where x and y represent the image coordinates of the facial landmarks.

Facial landmark extraction is challenging for several reasons: Firstly, facial appearance changes significantly across subjects under different facial expressions and head poses; secondly, facial occlusions by other objects, or self-occlusion due to extreme head poses, leads to incomplete facial appearance information; and thirdly, environmental conditions such as illumination can affect the appearance of the face on facial images.

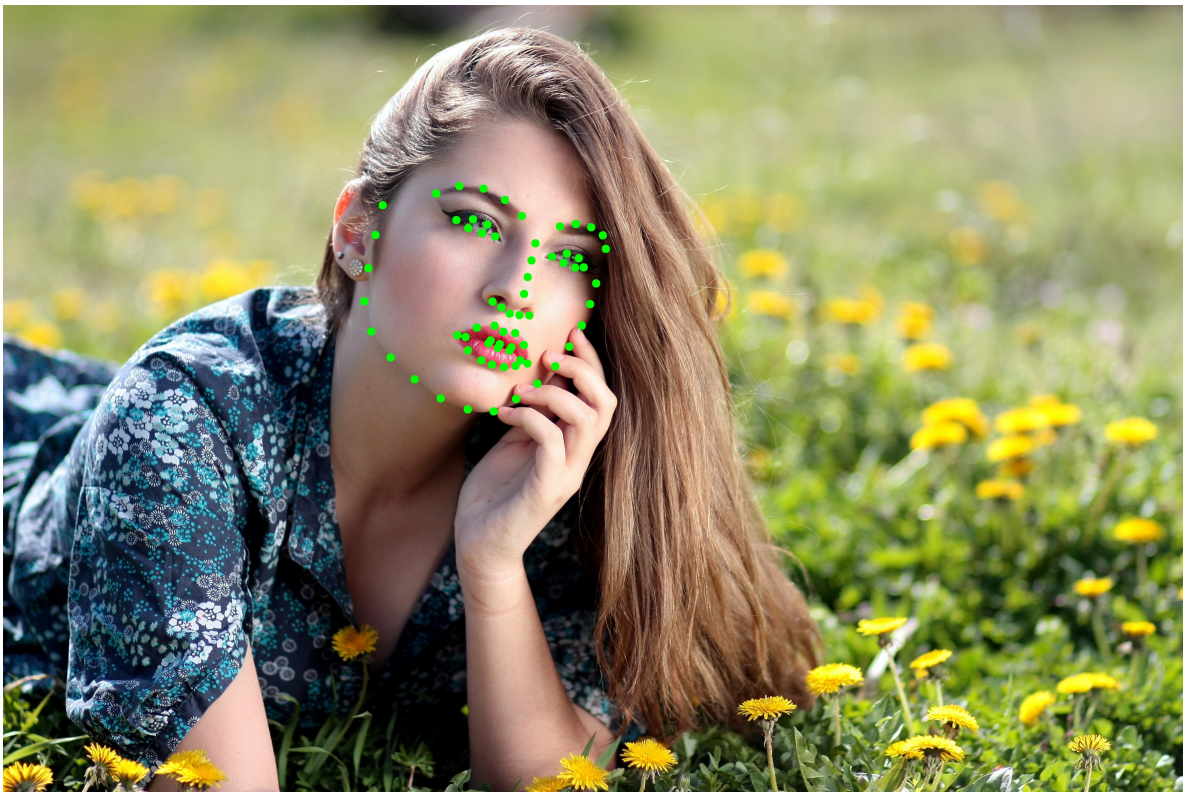

Figure 1. Sample image with predicted facial landmarks. An ensemble of randomized regression trees is used to detect 194 landmarks on a face from a single image [9].

Over the past few years, there have been important developments in facial landmark detection algorithms. We noticed that the first works focused on less challenging facial images without the said facial variations. Also, facial landmark detection algorithms usually aimed at handling several variations within fixed categories, and the facial images were usually collected with *controlled* conditions. For instance, in controlled conditions, facial poses and facial expressions can only be in certain categories. More recently, researchers began focusing on challenging *in-the-wild* conditions, in which facial images may undergo arbitrary facial expressions, head poses, illumination, facial occlusions, etc. In general, there is still a lack of a robust method that can handle all of those variations [5,6]. Algorithms for facial landmark extraction can be split into two categories: generative, and discriminative algorithms. The first type builds an object representation, and then searches for the region most similar to the object without taking background information into account. Discriminative methods train an online binary classifier to adaptively separate the object from the background, which are more robust against appearance variations of an object. Generative models include part-based generative models, such as Active Shape Models (ASM) [10,11], and holistic generative models, such as Active Appearance Models (AAM) [12–16] that model the shape of the face and its appearance as probabilistic distributions. Among discriminative models, cascaded regression models (CRMs) [9,17–23] have gained widespread popularity due to its excellent performance and low complexity. Several exhaustive surveys exist for facial landmark detection and tracking [5,24–27]. This survey is focused on landmark extraction and tracking using deep-learning techniques. We focus on deep-learning methods for facial landmark detection, as since the arrival of deep-learning neural networks [28,29], deep-learning methods have achieved state-of-art performance in identity and face recognition, object detection, and other research fields mentioned above. In the last few years, there appeared even some closely related tasks to facial landmark extraction in the literature, such as video facial landmark extraction and 3D facial landmark extraction. The first consists of tracking facial landmarks in successive video frames of the same person taking advantage of temporal redundancy

and identity. The hardest challenge is represented by in-the-wild video frames. The aim of 3D facial landmark extraction is to detect facial landmarks in arbitrary poses, for recovery of the projected 3D locations of invisible facial landmarks, given a 2D image. This article will consider the newest deep-learning methods for facial landmark extraction in 2D image and video. An analysis of many methods is given, and a comparison between performance is reported based on the performance measures used in the literature. The most common datasets used for both training and assessment of performance are reported and briefly analyzed. Eventually, the main challenges and future research directions will be provided.

2. Landmarks Extraction

In brief, the aim of facial landmark extraction is to detect facial landmark coordinates in an image, inside a face-region-bounding box. Duffner et al. [30] used a Convolutional Neural Network (CNN) to extract five landmarks on a facial image of low resolution. It trained a CNN to predict facial feature likelihood maps, supervised by ground truth maps defined through Gaussian distribution on the positions of the features. It may be safe to state that this was the very first study where this was proposed, and which successfully established image-to-image mapping for facial landmark detection. However, the authors did not face applications which required dense prediction on images in high resolution. Thus, they proposed an optimized and efficient CNN only for a limited number of fiducial points. In another article, Luo et al. [31] proposed to extract the facial landmarks exploiting the face-parsing segmentation results making use of a deep belief network. Their method is composed by four parts: a face detector, a face part detector, a components detector, and components segmentators. Luo et al. designed these layers in a Bayesian framework, assuming between them a spatial consistency probability prior.

Each one of these layers was unsupervised and pre-trained through a layer-wise Restricted Boltzmann Machine (RBM), and fine-tuned using logistic regression. Eventually, the segmentators were trained as deep auto-encoders. Seeing the great success of deep-learning techniques in the task of image classification [28], researchers started to extract sparse facial landmarks with similar structures. Sun et al. [32] used a cascaded coarse-to-fine CNN to detect five facial fiducial points. A three-stage method was adopted, and many CNNs were included in each stage. The CNNs in the first stage estimate the rough positions of many different sets of landmarks. Each one was then separately refined by the CNNs in the following stages. Despite its novelty and high accuracy, the input of the next CNN depends on local patches which are extracted from the previous one. The approach of TCDCN (Tasks Constrained Deep Convolutional Network) [33] consists of multi-task learning for optimizing the performance of five-point landmark extraction. The authors showed that additional facial features, such as gender and pose, could be useful for landmark extraction, while simultaneously providing further information during the inference procedure. It is possible to see that in Kumar et al. [34], local patch features extracted by a CNN work better with a linear regressor to provide a five-point prediction. In Zhang et al. [35], a fine-tuned, pre-trained CNN was used to detect local facial patch features, followed by a cascaded regressor to extract the landmarks. Both of the articles have shown that CNNs could stand for a good feature extractor in the conventional cascaded regression framework, explained in the introduction. Now take into account *dense* facial landmarks, which are landmarks that are not necessarily semantic, but which can also be contained in a contour. In Zhang et al. [36], a coarse-to-fine encoder–decoder network was used to extract 68 fiducial landmarks. The authors designed a four-stage cascaded encoder–decoder with a growing input resolution for different stages. Positions of landmarks were updated at the end of each stage with the output of the CNN. Then, the authors improved the method by suggesting considering an occlusion-recovering auto-encoder to reconstruct the eventual occluded facial parts with the aim of avoiding estimation errors caused by occlusions [37]. The occlusion-recovering network reconstructs the original face from the occluded one by training on a synthetic and occluded dataset. Sun et al. [38] used a multi-layer perceptron (MLP) as a *graph transformer network* [39] to replace the regressors in

a cascaded regression method to detect the facial landmarks. The authors have shown that such combination can be entirely trained by backpropagation. Wu et al. [40] designed a *three*-way factorized Restricted Boltzmann Machine (RBM) [41] for building a deep face-shape model for predicting the 68 point landmarks. The main drawback of using only one CNN to directly obtain a dense prediction lies in the fact that the network is trained to reach its best result only on the global shape, which could lead to several local inaccuracies. A simple and immediate idea is to refine different facial parts locally and independently with post-processing. Fan et al. [42] and Huang et al. [43] extracted dense facial landmarks by a CNN to estimate rough positions, followed by several small regional CNNs to refine different parts locally. This structure is more time-consuming, but can significantly optimize the precision. In another article, Lv et al. [44] used two-stage re-initialization with a deep network regressor in each stage. The framework is composed by a global step, where a gross landmark shape is extracted, and a local step, where fiducial points of each facial part are respectively estimated. A novelty lies in the fact that the global or local transformation parameter is predicted by a CNN to again initialize the region of the face to a standard shape before the landmark extraction. Performances on large poses are then highly improved. In Shao et al. [45], adaptive weights were applied to different landmarks during the different steps of training. The authors assigned a higher coefficient to some important landmarks, such as the eye and corners of the mouth at the starting phase of the training process, and then reduced such weights if the result converged. In this manner, the neural network first learns a robust global shape, and finally learns predictions which are locally and more refined. In addition, even if deep-learning-based techniques are not sensitive to initialization, head poses which are quite large still represent a big challenge. Wu et al. [6] proposed a tweaked structure at the end of a *Vanilla* network in TCDCN [33], where different branches were aimed at regressing shapes in several head poses. Over the last few years, Trigeorgis et al. [46] used a cascaded regression method with a Recurrent Neural Network (RNN), named the Mnemonic Descent Method (MDM). In MDM, the CNNs extract patch features, replacing well-known feature extractors such as SIFT [47] in SDM [17]. Furthermore, RNNs act as memory units, capable of sharing information between the cascaded levels. The RNN facilitates the joint optimization of the regressors, assuming that the cascades form a non-linear dynamical system. One more widely used approach consists of training the CNN to extract likelihood maps (also named response maps, probability maps, voting maps, and heat maps) as the network output, as originally proposed by [30,48]. The value of a pixel on the maps can be seen as the probability of the presence of each landmark in the pixel. Zadeh et al. [49] proposed using DCNN to produce a local response map and to fit the model as a Constrained Local Model (CLM). Since the deep encoder-decoder can provide image-to-image mapping, Lai et al. [50] made use of a full CNN to predict a starting face shape instead of a mean shape which is commonly used in cascaded regression models [17,18]. The authors introduced *Shape-Indexed Pooling* as a feature-mapping function to extract local features of each point, which was then given in input to the regressor. In the first version, the authors used a fully-connected layer to regress the final shape step-by-step, while replacing it with a RNN in the second version, inspired by MDM [46]. In Xiao et al. [51] an attention mechanism [52,53] was used where landmarks near the attention center were subject to a specific refinement procedure. Wang et al. [54] studied an approach for detecting multi-face landmarks through maps. Using an ROI pooling branch [55], face detection is not necessary and non-face activations are deleted over all of the likelihood maps. Concerning likelihood maps, another interesting method was proposed by Kowalski et al. [56]. The authors predicted the transformation to a standard pose and a feature image at the same time, with global likelihood maps in a cascaded fashion. The networks in different steps share the information by taking the transformation parameters from the previous step. Many recent studies have focused on multi-task CNN methods to get additional semantic information other than facial fiducial points. In addition to the aforementioned TCDCN [33], in Zhang et al. a method named MTCNN [57] was studied, which is a three-stage structure composed of CNNs which together performs face detection, face classification, and landmark extraction. The authors designed a fast Proposal Network (P-Net) to obtain facial region candidates and landmarks on low-resolution images in the

first step. Then, these candidates were refined in the next stage through a Refinement Network (R-Net), followed by an Output Network (O-Net) to obtain final bounding boxes and landmark positions, with higher resolution inputs. Ranjan et al. [58] designed a multi-branch CNN, named the All-in-One CNN, to simultaneously detect the face region, facial landmarks, head pose, smile probability, gender, and the age of the person by sharing a common convolutional feature extractor. In the previous paragraphs, of this survey an attempt has been made for reporting methods for facial landmark extraction, splitting them into different categories. The mentioned works can be divided into the following categories, respectively:

- Sparse facial landmarks detection;
- Dense facial landmarks detection;
- Landmarks detection with RNN;
- Landmark detection with likelihood maps;
- Landmark detection with multi-task learning.

However, some studies that make use of DCNN are harder to categorize. In a study by Belharbi et al. [59], facial landmark detection is treated as a structured-output problem. In a study by Güler et al. [60], facial landmarks were extracted in a deformation-free space, defined by two-dimensional mapping *U-V* in a 3D Morphable Model. The authors divided the regression of landmarks position into two parts, which they call quantized classification and residual regression, which respectively mean positioning the band region and the relative residual to the band region. A coarse band position is predicted through quantized regression, and a refined shape was predicted through residual regression. This algorithm gives a dense map between a three-dimensional object template and an image, given in input. Eventually, their output provides a robust initialization for CRMs. We are seeing great changes related to face landmark extraction, from directly predicting the point coordinates by fully-connected layers to predicting the landmark positions by likelihood maps using CNN and several other interesting ideas. It is noticeable that in many of the listed works, CNN is no longer simply regarded as a learnable image feature extractor, but rather as a multi-functional tool for processing different types of information for facial landmark extraction.

3. Landmarks Tracking

The task of video facial landmark extraction, also referred to as sequential facial landmark extraction, aims at aligning a sequence of person-specific images by exploiting continuous information in the video. An immediate idea for video face-tracking is represented by person-specific modeling, since personal identity information remains unchanged. In Chrysos et al. [61] a tracking pipeline was shown to improve tracking robustness in videos that contained speech. Other than face detection and landmark extraction, the authors designed a person-specific face detector by making use of a Deformable Part Model [62,63] and a person-specific generative landmark localizer. The latter iteratively updates the generic/person-specific appearance variations and shape/appearance parameters in turn. The method is used for the annotation of the 300VW [64] dataset semi-automatically. Asthana et al. [19] reformulated the cascaded regression in a parallel form to allow fast and efficient learning of each cascade level. The information retained by the following cascaded level was derived from the statistical distribution from the previous cascade regressor. More recent studies on cascaded regression is CCR [65] and iCCR [66]. The authors studied a continuous regression method while reformulating it so that the algorithm did not require sampling over the perturbed shapes (e.g., flipping, rotation, scaling). As a consequence, the computational complexity was largely reduced compared to the traditional cascaded regression-based methods. It is highly common to use Bayesian filters in object tracking. Then, an immediate idea is to combine them (e.g., Kalman filtering) with state-of-art landmark localizers. Therefore, in Pabhu et al. [67] a Kalman filter was exploited to track the facial landmarks by the head positions, head orientations, and facial shapes in video sequences. The 300VW workshop [64] proposed a challenging dataset focused on the tracking of landmarks. Considering all the methods, one of the main ones designed a pose-specific CRM [68], and another one

used a progressive initialization [69], which aims to improve the problem of initialization in extremely bad poses. Taking inspiration from the incremental learning method [70], Peng et al. [71] used a CNN with likelihood maps to evaluate the fitting results at the end of the network. This ensures more reliable results. The RED-Net [72] was introduced to improve the performance of facial landmark extraction on video by reorganizing the identity information and pose/expression information. The first can be regarded as an invariant in a video, while the pose and expression information changes over time. The author proposed a dual-path network using likelihood maps in which include one path extracted the identity information, while the other path learned the pose/expression information making use of a RNN network. Recently, Gu et al. [73] added a one-layer RNN to the final part of a VGG network for tracking facial landmarks and head poses. The authors showed that Bayesian filters could be formulated as a RNN, linearly-activated and without bias. If we analyse their results, tracking with RNN is more accurate and reliable than frame-by-frame detections and state-of-the-art landmark localizers tracked using a Kalman filter. Another algorithm that uses RNN is called TSTN, proposed by Liu et al. [39]. They adopted two network streams, spatial and temporal. The spatial one learns to transform local facial patches to shape residuals, which is then used to refine the current facial shape based on the previous shape. The temporal stream is designed as a deep encoder–decoder with a two layers of RNN for capturing facial dynamics in the temporal dimension. This stream takes consecutive frames as the input, and renders the temporal shape update. The final shape is determined by a weighted fusion of two streams shape updates. A Long Short-Term Memory (LSTM) module was used by Hou et al. [74] to guide the spatial estimation for the next step, just as in MDM, and to simultaneously guide the estimation for the next frame. Among all these methods, it is noticeable that deep-learning methods are not yet widely used in video landmark extraction due to their high complexity, size, and memory constraints which are still a significant problem for real-time detection on mobile platforms.

4. Datasets

In this section, the most frequent datasets available in the literature are listed and analyzed; mainly, they are divided into image datasets and video datasets. The first can generally be categorized into two parts: images can be taken under constrained conditions, such as controlled lighting conditions and specific poses; otherwise, images can be taken under unconstrained conditions, which are usually referred to as *in-the-wild* datasets. In the category of image datasets, the main ones are listed below:

- Multi-PIE [75] is, above all datasets, one of the largest. It is constrained and contains 337 subjects in 15 views, with 19 illumination conditions and six different expressions. The facial landmarks are labeled with 39 points or 68 points.
- XM2VTS [76] contains four registrations for 295 subjects, taken over a period of 4 months. Each video contains a speaking and a rotating head shot. The dataset is annotated with 68 points, and is included for the 300W challenge [77].
- The 300-W [77] dataset (300 Faces in-the-Wild Challenge). Among all in-the-wild datasets, this has been the most popular one in recent years and it combines several datasets, such as Helen [78], LFPW [79], AFW [80], and a newly introduced challenging dataset, iBug. Summing up, it contains 3837 images and a further test set with 300 indoor and outdoor images, respectively. All the images are annotated with 68 points. The dataset is commonly divided in two parts: the usual subset, including LFPW and Helen, and the challenging dataset with AFW and iBug.
- The Menpo [81] dataset. It is the largest in-the-wild facial landmark dataset, and contains 6679 semi-front view face images, annotated with 68 points, and 5335 profile view face images, annotated with 39 points in the training set. The test set is composed of 12006 front view images and 4253 profile view images. It was introduced for the Menpo challenge in 2017 in order to raise an even more difficult challenge to test the robustness of facial landmark extraction algorithms, since it involves a high variation of poses, light conditions, and occlusions.

Video-based annotated datasets are used for sequential facial landmark extraction. The 300-VW [64] has the largest number of facial-point annotated videos and frames. The dataset is composed by 50 training videos and 64 videos for testing, which are further divided into three scenarios, according to different light conditions, expressions, head poses, and occlusions. All of the frames are annotated in 68 points in a semi-automatic manner. The Menpo 3D tracking [82] dataset is the unique dataset in which 3D facial landmarks are annotated in video by the 3DMM algorithm [83]. The dataset contains 55 videos from the 300VW dataset, annotated again in 3D, in addition to all of the images in 300W, and Menpo re-annotated in the same way. The dataset provides us not only the landmarks in projected image space, but also the landmarks in 3D space.

5. Evaluation Metrics and Comparison

For providing comparable results regardless of the image size and camera focus, in the literature it is common to measure the distance between the ground truth and the detection result by Normalized Mean Error (NME) e, calculated as:

$$e = \frac{||S - S^*||_2}{d^*},$$

where S, and S^* represent the detected shape and the shape of the ground truth, respectively. d^* is a normalizing distance, which could be the inter-ocular distance (IOD) or inter-pupil distance. Many times, it is used as the bounding box diagonal or geometric mean of image length, with the height as d^* instead if the distance between the two eyes is too small on 3D/large pose datasets. Another metric that is based on the NME is the Cumulative Error Distribution (CED) curve. The CED generally represents the proportion of images in the test set having an error below a given threshold. This curve provides a visual result of the algorithm performance in different situations, and the Area Under the Curve (AUC) provides a qualitative result of how the algorithm performs at progressive mean errors:

$$AUC_\alpha = \int_0^\alpha f(e)de,$$

where e is the normalized error, $f(e)$ is the cumulative error distribution (CED) function, and α is the upper bound that is used to calculate the definite integration. A bigger AUC value generally means that the algorithm has better performance. The failure rate is used to measure the robustness of an algorithm. A threshold of NME was chosen to be a threshold of failure, and the proportion of failed detection was calculated to represent the capacity of handling the difficult images. Now, we are ready to provide a comparison of different facial landmark extraction methods, as well as different deep-compression models. This comparison includes traditional cascaded regression methods and deep-learning-based facial landmark extraction methods. Zhang et al. [33] and Kowalski et al. [56] both provide a good benchmark on several popular methods by measuring the normalized inter-ocular distance error. Table 1 shows the performance of seven non-deep-learning 2D facial landmark extraction methods and eight deep learning facial landmark extraction methods evaluated on 300W by the normalized inter-ocular distance error. The failure rate is not included in the table since the choice of threshold is not objective. Table 2 reports a comparison of the Mean Error (%) (normalized by face size) of different video facial landmark extraction methods on 300VW. In general, the deep-learning-based algorithms outperform the others. However, considering the performances, all of the cascaded-regression based methods can run in real-time even with a Matlab implementation [26], while some deep-learning-based methods can achieve real-time detection on a powerful GPU or CPU, but most runtimes on CPU are not satisfying.

Table 1. Performance comparison of different landmark extraction methods based on NME, AUC, and FPS (frames per second). The upper part of the table lists non-deep-learning methods, while the lower part lists deep-learning methods. Data was obtained from [33,56] and the original publications (*). The measure from [84] was obtained using a threshold of 0.07 on the 300W-private dataset.

Non Deep Learning Method	Year	Database	NME (%)	$AUC_{0.08}$ (%)	FPS on Video
DRMF [85]	2013	300W	9.22	-	0.5
RCPR [22]	2013	300W	8.35	-	80
ESR [18]	2014	300W	7.58	43.12	-
SDM [17]	2013	300W	7.52	42.94	40
ERT [9]	2014	300W	6.40	-	25
LBF [21]	2014	300W	6.32	-	3000
CFSS [23]	2015	300W	5.76	55.9 *	10
Deep Learning Method		**Database**	**NME (%)**	**$AUC_{0.08}$ (%)**	
CFAN [36]	2014	300W	7.69	-	20
TCDCN [33]	2014	300W	5.54	41.7 *	58
TSR [44]	2017	300W	4.99	-	111
RAR [51]	2016	300W	4.94	-	-
DRR [50]	2018	300W	4.90	-	-
MDM [46]	2016	300W	4.05	52.12	-
DAN [56]	2017	300W	3.59	55.33	73
2DFAN [84]	2017	300W	-	66.90 *	30
DenseReg + MDM [60]	2017	300W	-	52.19	8

Table 2. Comparison of the Mean Error (%) (normalized by face size) of different video facial landmark extraction methods on 300VW. The data was obtained from [72] and original publications. * indicates deep-learning-based methods. † indicates that the runtime is measured on GPU (graphics processing unit).

Video Facial Landmark Extraction Method Comparison on the 300VW Dataset					
Method	ESR [18]	SDM [17]	CFSS [23]	PIEFA [86]	CFAN * [36]
NME	7.09	7.25	6.13	6.37	6.64
FPS	67	40	10	-	20
Method	TCDCN * [33]	RED * [72]	RED-Res * [87]	RNN * [73]	TSTN * [39]
NME	7.59	6.25	4.75	6.16	5.59
FPS	59	33 †	18 †	-	30

6. Main Challenges

Despite the articles we have analyzed, research on improved face landmarking techniques has been continuing. Emerging applications are requiring that landmarking algorithms be executed in real-time while operating with the computational power of an embedded system, such as intelligent cameras or smart phones. Furthermore, these applications require increasingly more robust algorithms against a variety of confounding factors, such as out-of-plane poses, occlusions, illumination effects, and expressions. The details of these factors that compromise the performance of facial landmark detection are as follows:

- Variability: Landmark appearances differ due to intrinsic factors, such as face variability between individuals, but also due to extrinsic factors, such as partial occlusion, illumination, expression, pose, and camera resolution. Facial landmarks can sometimes be only partially observed due to occlusions of hair, hand movements, or self-occlusion due to extensive head rotations. The other two major variations that compromise the success of landmark detection are illumination artifacts and facial expressions. A face landmarking algorithm that works well under and across all intrinsic variations of faces, and that delivers the target points in a time-efficient manner has not yet been feasible.

- Acquisition conditions: Much as in the case of face recognition, acquisition conditions, such as illumination, resolution, and background clutter, can affect the landmark localization performance. This is attested by the fact that landmark localizers trained in one database usually have inferior performance when tested on another database.
- The number of landmarks and their accuracy requirements: The accuracy requirements and the number of landmark points vary based on the intended application. For example, coarser detection of only the primary landmarks, e.g., nose tip, four eye and two mouth corners, or even the bounding box enclosing these landmarks, may be adequate for face detection or face recognition tasks. On the other hand, higher-level tasks, such as facial expression understanding or facial animation, require a greater number for landmarks that is from 20–30 to 60–80, as well as greater spatial accuracy. As for the accuracy requirement, fiducial landmarks, such as on the eyes and nose, need to be determined more accurately as they often guide the search for secondary landmarks with less prominent or reliable image evidence. It has been observed, however, that landmarks on the rim of the face, such as the chin, cannot be accurately localized in either manual annotation and automatic detection. Shape guide algorithms can benefit from the richer information coming from a larger set of landmarks. For example, Milborrow and Nicolls [88] have shown that the accuracy of landmark localization increases proportionally to the number of landmarks considered, and have recorded a 50% improvement as the ensemble increases from 3 to 68 landmarks.

In the final analysis, accurate and precise landmarking remains a difficult problem since, except for a few, landmarks do not necessarily correspond to high-gradient or other salient points. Hence, low-level image processing tools remain inadequate to detect them, and recourse has to be made for higher-order face shape information. This probably explains the tens of algorithms presented and the hundreds of articles published in the last two decades in the quest to develop a landmarking scheme on par with human annotators [5,26,27].

7. Conclusions

In this survey, recent deep learning-based 2D facial landmark extraction methods were reviewed. After analyzing face landmarking techniques, comparing performances, and seeing the main challenges, we can draw the conclusion that deep-learning-based algorithms outperform others in terms of precision. However, computation efficiency remains a major constraint, especially for video facial landmark extraction. Despite the fact that deep-learning methods achieve excellent performance in many datasets, facial landmark extraction on a limited resource platform has not been solved. One future research direction is to investigate compression methods, such as Shuffle-Net [89]. Another direction is to focus on precision [90], where specific applications like animation demand high precision to result in perfect rendering. Other promising research paths in landmarking techniques are listed as the following: Sparse dictionaries, that is, the paradigm of recognition under sparsity constraint and building of discriminatory dictionaries seems to be one viable method. The discriminative sparse dictionary can be constructed per landmark [91,92] or collectively, as in [93]. Adaboost selected features for multiview landmarking: Gabor or Haar wavelet features selected via the modified Adaboost scheme, where commonality and a geometric configuration of landmark appearances is exploited [94]. Finally, multiframe landmarking: the determination of landmark positions exploits the information in subsequent frames of a video, using, for example, spatio-temporal representations [95].

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Murphy-Chutorian, E.; Trivedi, M.M. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 607–626. [CrossRef] [PubMed]
2. Pantic, M.; Rothkrantz, L.J.M. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]
3. Hansen, D.W.; Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 478–500. [CrossRef] [PubMed]
4. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
5. Chrysos, G.G.; Antonakos, E.; Snape, P.; Asthana, A.; Zafeiriou, S. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *Int. J. Comput. Vis.* **2018**, *126*, 198–232. [CrossRef]
6. Wu, Y.; Hassner, T.; Kim, K.; Medioni, G.; Natarajan, P. Facial landmark detection with tweaked convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 3067–3074. [CrossRef] [PubMed]
7. Zhang, H.; Li, Q.; Sun, Z.; Liu, Y. Combining data-driven and model-driven methods for robust facial landmark detection. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2409–2422. [CrossRef]
8. Labati, R.D.; Genovese, A.; Muñoz, E.; Piuri, V.; Scotti, F.; Sforza, G. Computational intelligence for biometric applications: A survey. *Int. J. Comput.* **2016**, *15*, 40–49.
9. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
10. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
11. Cristinacce, D.; Cootes, T.F. Boosted regression active shape models. In Proceedings of the British Machine Vision Conference, Warwick, UK, 10–13 September 2007; Rajpoot, N.M., Bhalerao, A.H., Eds.; BMVA Press: San Francisco, CA, USA, 2007; pp. 1–10. [CrossRef]
12. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
13. Edwards, G.J.; Taylor, C.J.; Cootes, T.F. Interpreting face images using active appearance models. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 300–305.
14. Tzimiropoulos, G.; Pantic, M. Optimization problems for fast aam fitting in-the-wild. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 593–600.
15. Alabort-i Medina, J.; Zafeiriou, S. Bayesian active appearance models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3438–3445.
16. Boccignone, G.; Bodini, M.; Cuculo, V.; Grossi, G. Predictive sampling of facial expression dynamics driven by a latent action space. In Proceedings of the 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS), Las Palmas de Gran Canaria, Spain, 26–29 November 2018; pp. 143–150. [CrossRef]
17. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
18. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. *Int. J. Comput. Vis.* **2014**, *107*, 177–190. [CrossRef]
19. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Incremental face alignment in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1859–1866.
20. Bodini, M. Can We Automatically Assess the Aesthetic value of an Image? In Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2019 (ISMAC-CVB), Elayampalayam, India, 13–14 March 2019; Springer International Publishing: Cham, Switzerland, 2019; in press.

21. Ren, S.; Cao, X.; Wei, Y.; Sun, J. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1685–1692.
22. Burgos-Artizzu, X.P.; Perona, P.; Dollár, P. Robust face landmark estimation under occlusion. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1513–1520.
23. Zhu, S.; Li, C.; Change Loy, C.; Tang, X. Face alignment by coarse-to-fine shape searching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4998–5006.
24. Celiktutan, O.; Ulukaya, S.; Sankur, B. A comparative study of face landmarking techniques. *EURASIP J. Image Video Process.* **2013**, *2013*, 13. [CrossRef]
25. Yang, H.; Jia, X.; Loy, C.C.; Robinson, P. An empirical study of recent face alignment methods. *arXiv* **2015**, arXiv:1511.05049.
26. Wang, N.; Gao, X.; Tao, D.; Yang, H.; Li, X. Facial feature point detection: A comprehensive survey. *Neurocomputing* **2018**, *275*, 50–65. [CrossRef]
27. Jin, X.; Tan, X. Face alignment in-the-wild: A survey. *Comput. Vis. Image Underst.* **2017**, *162*, 1–22. [CrossRef]
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, CA, USA, 3–6 December 2012; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates Inc.: New York, NY, USA, 2012; pp. 1097–1105.
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
30. Duffner, S.; Garcia, C. A connexionist approach for robust and precise facial feature detection in complex scenes. In Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, ISPA 2005, Zagreb, Croatia, 15–17 September 2005; pp. 316–321.
31. Luo, P.; Wang, X.; Tang, X. Hierarchical face parsing via deep learning. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2480–2487.
32. Sun, Y.; Wang, X.; Tang, X. Deep convolutional network cascade for facial point detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3476–3483.
33. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial landmark detection by deep multi-task learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 94–108.
34. Kumar, A.; Ranjan, R.; Patel, V.; Chellappa, R. Face alignment by local deep descriptor regression. *arXiv* **2016**, arXiv:1601.07950.
35. Zhang, S.; Yang, H.; Yin, Z.P. Transferred deep convolutional neural network features for extensive facial landmark localization. *IEEE Signal Process. Lett.* **2016**, *23*, 478–482. [CrossRef]
36. Zhang, J.; Shan, S.; Kan, M.; Chen, X. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 1–16.
37. Zhang, J.; Kan, M.; Shan, S.; Chen, X. Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3428–3437.
38. Sun, P.; Min, J.K.; Xiong, G. Globally tuned cascade pose regression via back propagation with application in 2D face pose estimation and heart segmentation in 3D CT images. *arXiv* **2015**, arXiv:1503.08843.
39. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Two-stream transformer networks for video-based face alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2546–2554. [CrossRef] [PubMed]
40. Wu, Y.; Ji, Q. Discriminative deep face shape model for facial point detection. *Int. J. Comput. Vis.* **2015**, *113*, 37–53. [CrossRef]
41. Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **2002**, *14*, 1771–1800. [CrossRef] [PubMed]
42. Fan, H.; Zhou, E. Approaching human level facial landmark localization by deep learning. *Image Vis. Comput.* **2016**, *47*, 27–35. [CrossRef]

43. Huang, Z.; Zhou, E.; Cao, Z. Coarse-to-fine Face Alignment with Multi-Scale Local Patch Regression. *arXiv* **2015**, arXiv:1511.04901.
44. Lv, J.J.; Shao, X.; Xing, J.; Cheng, C.; Zhou, X. A Deep Regression Architecture with Two-Stage Re-initialization for High Performance Facial Landmark Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 4.
45. Shao, Z.; Ding, S.; Zhao, Y.; Zhang, Q.; Ma, L. Learning deep representation from coarse to fine for face alignment. *arXiv* **2016**, arXiv:1608.00207.
46. Trigeorgis, G.; Snape, P.; Nicolaou, M.A.; Antonakos, E.; Zafeiriou, S. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4177–4187.
47. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
48. Duffner, S.; Garcia, C. A hierarchical approach for precise facial feature detection. In *Compression et Représentation des Signaux Audiovisuels*; CORESA: Rennes, France, 2005; pp. 29–34.
49. Zadeh, A.; Baltrušaitis, T.; Morency, L.P. Deep constrained local models for facial landmark detection. *arXiv* **2016**, *3*, 6, arXiv:1611.08657.
50. Lai, H.; Xiao, S.; Pan, Y.; Cui, Z.; Feng, J.; Xu, C.; Yin, J.; Yan, S. Deep recurrent regression for facial landmark detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1144–1157. [CrossRef]
51. Xiao, S.; Feng, J.; Xing, J.; Lai, H.; Yan, S.; Kassim, A. Robust facial landmark detection via recurrent attentive-refinement networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin, Germany, 2016; pp. 57–72.
52. Bodini, M. Sound Classification and Localization in Service Robots with Attention Mechanisms. In Proceedings of the First Annual Conference on Computer-aided Developments in Electronics and Communication (CADEC-2019), Amaravati, India, 2–3 March 2019; in press.
53. Bodini, M. Probabilistic Nonlinear Dimensionality Reduction through Gaussian Process Latent Variable Models: An Overview. In Proceedings of the First Annual Conference on Computer-aided Developments in Electronics and Communication (CADEC-2019), Amaravati, India, 2–3 March 2019; in press.
54. Wang, L.; Yu, X.; Bourlai, T.; Metaxas, D.N. A coupled encoder-decoder network for joint face detection and landmark localization. *Image Vis. Comput.* **2018**. [CrossRef]
55. Gkioxari, G.; Girshick, R.; Malik, J. Contextual action recognition with r* cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1080–1088.
56. Kowalski, M.; Naruniec, J.; Trzcinski, T. Deep alignment network: A convolutional neural network for robust face alignment. In Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR), Faces-in-the-wild Workshop/Challenge, Honolulu, HI, 21–27 July 2017, pp. 2034–2043. [CrossRef]
57. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
58. Ranjan, R.; Sankaranarayanan, S.; Castillo, C.D.; Chellappa, R. An all-in-one convolutional neural network for face analysis. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 17–24.
59. Belharbi, S.; Chatelain, C.; Herault, R.; Adam, S. Facial landmark detection using structured output deep neural networks. *arXiv* **2015**, arXiv:1504.07550v3.
60. Güler, R.A.; Trigeorgis, G.; Antonakos, E.; Snape, P.; Zafeiriou, S.; Kokkinos, I. DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 2, p. 5.
61. Chrysos, G.G. Epameinondas Antonakos Patrick Snape Akshay Asthana and Stefanos Zafeiriou. A Comprehensive Performance Evaluation of Deformable Face Tracking In-the-Wild. *Int. J. Comput. Vis.* **2016**, *126*, 198–232. [CrossRef]
62. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
63. Chrysos, G.G.; Zafeiriou, S. PD 2 T: Person-Specific Detection, Deformable Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2555–2568. [CrossRef] [PubMed]

64. Shen, J.; Zafeiriou, S.; Chrysos, G.G.; Kossaifi, J.; Tzimiropoulos, G.; Pantic, M. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 50–58.
65. Sánchez-Lozano, E.; Martinez, B.; Tzimiropoulos, G.; Valstar, M. Cascaded continuous regression for real-time incremental face tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin, Germany, 2016; pp. 645–661.
66. Sánchez-Lozano, E.; Tzimiropoulos, G.; Martinez, B.; De la Torre, F.; Valstar, M. A functional regression approach to facial landmark tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2037–2050. [CrossRef] [PubMed]
67. Prabhu, U.; Seshadri, K.; Savvides, M. Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; Springer: Berlin, Germany, 2010; pp. 86–99.
68. Yang, J.; Deng, J.; Zhang, K.; Liu, Q. Facial shape tracking via spatio-temporal cascade shape regression. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 41–49.
69. Xiao, S.; Yan, S.; Kassim, A.A. Facial landmark detection via progressive initialization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 33–40.
70. Peng, X.; Huang, J.; Metaxas, D.N. Sequential Face Alignment via Person-Specific Modeling in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 107–116.
71. Peng, X.; Hu, Q.; Huang, J.; Metaxas, D.N. Track facial points in unconstrained videos. *arXiv* **2016**, arXiv:1609.02825.
72. Peng, X.; Feris, R.S.; Wang, X.; Metaxas, D.N. A recurrent encoder-decoder network for sequential face alignment. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin, Germany, 2016; pp. 38–56.
73. De, J.G.X.Y.S.; Kautz, M.J. Dynamic facial analysis: From Bayesian filtering to recurrent neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
74. Hou, Q.; Wang, J.; Bai, R.; Zhou, S.; Gong, Y. Face alignment recurrent network. *Pattern Recognit.* **2018**, *74*, 448–458. [CrossRef]
75. Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. Multi-pie. *Image Vis. Comput.* **2010**, *28*, 807–813. [CrossRef] [PubMed]
76. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, DC, USA, 22–24 March 1999; Volume 964, pp. 965–966.
77. Sagonas, C.; Antonakos, E.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge. *Image Vis. Comput.* **2016**, *47*, 3–18. [CrossRef]
78. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin, Germany, 2012; pp. 679–692.
79. Belhumeur, P.N.; Jacobs, D.W.; Kriegman, D.J.; Kumar, N. Localizing parts of faces using a consensus of exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2930–2940. [CrossRef]
80. Zhu, X.; Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2879–2886.
81. Zafeiriou, S.; Trigeorgis, G.; Chrysos, G.; Deng, J.; Shen, J. The menpo facial landmark localisation challenge: A step towards the solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 2.
82. Zafeiriou, S.; Chrysos, G.; Roussos, A.; Ververas, E.; Deng, J.; Trigeorgis, G. The 3d Menpo Facial Landmark Tracking Challenge. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2503-2511. [CrossRef]

83. Booth, J.; Antonakos, E.; Ploumpis, S.; Trigeorgis, G.; Panagakis, Y.; Zafeiriou, S. 3d face morphable models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 48–57.
84. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030. [CrossRef]
85. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Robust discriminative response map fitting with constrained local models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3444–3451.
86. Peng, X.; Zhang, S.; Yang, Y.; Metaxas, D.N. Piefa: Personalized incremental and ensemble face alignment. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3880–3888.
87. Peng, X.; Feris, R.S.; Wang, X.; Metaxas, D.N. RED-Net: A Recurrent Encoder–Decoder Network for Video-Based Face Alignment. *Int. J. Comput. Vis.* **2018**, *126*, 1–17. [CrossRef]
88. Milborrow, S.; Nicolls, F. Locating facial features with an extended active shape model. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Springer: Berlin, Germany, 2008; pp. 504–513.
89. Zhang, X.; Zhou.; X..; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv* **2017**, arXiv:1707.01083.
90. Dong, X.; Yu, S.I.; Weng, X.; Wei, S.E.; Yang, Y.; Sheikh, Y. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 18–22 June 2018; pp. 360–368.
91. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. *Discriminative Learned Dictionaries for Local Image Analysis*; Technical Report; Minnesota Univ Minneapolis Inst for Mathematics and Its Applications: Minneapolis, MN, USA, 2008.
92. Bodini, M.; D'Amelio, A.; Grossi, G.; Lanzarotti, R.; Lin, J. Single Sample Face Recognition by Sparse Recovery of Deep-Learned LDA Features. In *Advanced Concepts for Intelligent Vision Systems*; Blanc-Talon, J., Helbert, D., Philips, W., Popescu, D., Scheunders, P., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 297–308.
93. Salakhutdinov, R.; Torralba, A.; Tenenbaum, J. Learning to share visual appearance for multiclass object detection. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1481–1488.
94. Torralba, A.; Murphy, K.P.; Freeman, W.T. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 854–869. [CrossRef] [PubMed]
95. Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of local spatio-temporal features for action recognition. In Proceedings of the British Machine Conference, London, UK, 7–10 September 2009; Cavallaro, A., Prince, S., Alexander, D., Eds.; BMVA Press: San Francisco, CA, USA, 2009; pp. 1–11. [CrossRef]