A survey of deep facial landmark detection

Yongzhe Yan, Xavier Naturel, Thierry Chateau, Stefan Duffner, Christophe Garcia, Christophe Blanc

A survey of deep facial landmark detection

Yongzhe Yan[1,2] Xavier Naturel[2] Thierry Chateau[1] Stefan Duffner[3]
Christophe Garcia[3] Christophe Blanc[1]

[1] Université Clermont Auvergne, France
[2] Wisimage, France
[3] Université de Lyon, CNRS, INSA Lyon, LIRIS, UMR5205, Lyon, France

yongzhe.yan@etu.uca.fr

Résumé

La détection de landmarks joue un rôle crucial dans de nombreuses applications d'analyse du visage comme la reconnaissance de l'identité, des expressions, l'animation d'avatar, la reconstruction 3D du visage, ainsi que pour les applications de réalité augmentée comme la pose de masque ou de maquillage virtuel. L'avènement de l'apprentissage profond a permis des progrès très importants dans ce domaine, y compris sur les corpus non contraints (in-the-wild). Nous présentons ici un état de l'art centré sur la détection 2D dans une image fixe, et les méthodes spécifiques pour la vidéo. Nous présentons ensuite les corpus existants pour ces trois tâches, ainsi que les métriques d'évaluations associées. Nous exposons finalement quelques résultats, ainsi que quelques pistes de recherche.

Mots Clef

Détection de landmark facial, Alignement de visage, Deep learning

Abstract

Facial landmark detection plays a very important role in many facial analysis applications such as identity recognition, facial expression analysis, facial animation, 3D face reconstruction as well as facial beautification. With the recent advance of deep learning, the performance of facial landmark detection, including on unconstrained in-the-wild dataset, has seen considerable improvement. This paper presents a survey of deep facial landmark detection for 2D images and video. A comparative analysis of different face alignment approaches is provided as well as some future research directions.

Keywords

Facial landmark detection, Face alignment, Deep learning

1 Introduction

Face alignment plays a very important role in face recognition, facial expression analysis, facial animation, 3D face reconstruction and facial photo beautification. Facial landmarks, also known as facial feature points or fiducial facial points are located at semantic locations such as eye contours, eye brows, mouth contours. Detection of facial landmarks is a subset of face alignment techniques. This survey is concerned only about landmark detection and tracking using deep learning.

Several comprehensive surveys exist for facial landmark detection [10, 65, 59, 27] and facial landmark tracking [11]. Facial landmark detection algorithms can be mainly categorized into two types, generative or discriminative. The generative ones, which include the part-based generative models such as ASM [14, 15] and holistic generative models such as AAM [13, 18, 57, 39], model the facial shape and facial texture/appearance as probabilistic distributions. Cascaded regression models [64, 9, 3, 48, 8, 76, 28] have gained a large popularity thanks to excellent performance and low complexity. Cascaded regression models usually consist of three important parts, the initial shape s_0, the cascaded regressors R and the shape-indexed feature extraction function ϕ. To fit the model, the shape is iteratively updated stage by stage from the initial shape. At each stage t, the shape s^t is updated by :

$$s^t = s^{t-1} + R^t \phi(I, s^{t-1}) \,, \tag{1}$$

where I is the input image.

Since the arrival of deep neural networks architectures [30, 34], deep learning methods have achieved state-of-art performance in face recognition, semantic segmentation and object detection. Thus, in this survey, we focus on deep learning methods for facial landmark detection.

Recently, researchers have also shown interest for closely-related tasks such as video face alignment and 3D face alignment. Video face alignment generally tackles the challenge of facial landmark tracking in consecutive *in-the-wild* video frames of the same person by exploiting temporal redundancy and identity specifics. The objective of 3D face alignment is to predict facial landmarks in arbitrary poses, aiming to recover the projected 3D locations of in-

visible facial landmarks given a 2D image. Due to space constraints, this paper will focus only on recent deep learning methods for 2D image and video face landmarks detection.

2 2D landmark detection

The objective of 2D face alignment is to detect 2D facial landmark coordinates given an image, along with a face region bounding box. Duffner et al. [16] proposed to use a CNN to detect five key landmarks over a low resolution facial image. They trained a "light-weight" neural network to predict facial feature likelihood maps supervised by ground truth maps defined by 2D Gaussian distribution on the feature positions. To the best of our knowledge, this is the very first work that proposed and successfully established an image-to-image mapping for facial landmark detection. However, the authors did not target applications requiring dense prediction on images in high resolution and therefore proposed an optimized and efficient CNN for a limited number of key points. Another work by Luo et al. [37] proposed to detect the facial landmarks based on the face parsing segmentation results using a deep belief network. Their hierarchical face parsing framework includes four parts : the face detector, the face part detector, the components detector and components segmentators. They model the different layers under a Bayesian framework with spatial consistency prior between the layers. Each layer is pre-trained in an unsupervised manner using a layer-wise Restricted Boltzmann Machine (RBM) and fine-tuned for classification using logistic regression. The segmentators are trained as deep auto-encoders.

2.1 Sparse facial landmark detection

Following the great success of deep learning in image classification [30], researchers started to predict sparse facial landmarks with similar structures. Sun et al. [55] proposed to use a cascaded coarse-to-fine CNN to detect five facial landmarks. A 3-stage framework is adopted and several convolutional neural networks are included in each stage. The CNNs in the first stage estimate the rough positions of several different sets of landmarks. Each landmark is then separately refined by the CNNs in the following stages. Despite its innovation and high accuracy, this method is not completely end-to-end since the input of the following CNN depends on the local patches extracted from the previous one. The TCDCN approach [75] uses multi-task learning to optimize the performance of 5-point facial landmark detection. The authors showed that auxiliary facial attributes such as gender and pose can be helpful for the detection while providing additional information during the inference at the same time. Kumar et al. [31] showed that the local patch features extracted by a CNN work well with a linear regressor to give a five-point prediction. Zhang et al. [73] fine-tune a pre-trained CNN to extract local facial patch features followed by a cascaded regressor to predict the facial landmarks. Both of them proved that CNN could act as a good feature extractor in the conventional cascaded regression framework.

2.2 Dense facial landmark detection

We now consider dense facial landmarks, i.e. landmarks that are not necessarily semantic but can also be part of a contour (e.g. the popular 68 points or the 194 Helen models). Zhang et al. [71] proposed to use a coarse-to-fine encoder-decoder network to simultaneously detect 68 facial points. They proposed a 4-stage cascaded encoder-decoder network with increasing input resolution in different stages. The landmark positions are updated at the end of each stage by the CNN output. The author consequently improved this method by proposing to add an occlusion-recovering auto-encoder to reconstruct the occluded facial parts in order to avoid errors due to occlusions [70]. The occlusion-recovering auto-encoder network is designed to reconstruct the genuine face appearance from the occluded one by training on a synthetic occluded dataset. Sun et al. [54] used a MLP as a graph transformer network to replace the regressors in a cascaded regression framework to detect the facial landmarks and proved that this combination could be fully trained by backpropagation. Wu et al. [61] used a 3-way factorized Restricted Boltzmann Machine(RBM) [24] to build a deep face shape model to predict the dense 68-point facial landmarks.

One disadvantage of using a single CNN to predict directly a dense prediction is that the network is trained to achieve its best result on the global shape, which could possibly leave some local imprecisions. A straight-forward idea is to refine different facial parts locally and independently as post-processing. Duffner et al.[17] and later [19, 26] proposed to predict dense facial landmarks by a CNN to estimate rough positions followed by several small regional CNNs to refine different parts locally. This kind of structure is more time-consuming but can significantly optimize the precision. Another work by Lv et al. [38] proposed to use two-stage re-initialization with a deep network regressor in each stage. The framework consists of a global stage, where a coarse facial landmark shape is predicted and a local stage, where landmarks of each facial part are estimated respectively. One of the innovation is that the global/local transformation parameter is estimated by a CNN to reinitialize the facial region to a canonical shape prior to the landmark prediction. This largely improves the performance on large poses. Shao et al. [52] applies adaptive weights to different landmarks during different phases of the training. They give a relatively bigger coefficient to some important points such as eye corners and mouths corners at the beginning of the training process and then reduce their weights if the result has converged. This operation enables the neural network to first learn a robust global shape, and learning locally-refined predictions afterwards.

Additionally, despite the fact that deep learning-based methods are not that sensitive to initialization in practice, large head poses still remain a big challenge. Wu et al. [60] pro-

posed a tweaked structure at the end of a Vanilla network in TCDCN [75], where different branches are aimed at regressing shapes in different head poses.

2.3 Recurrent neural networks

Recently, Trigeorgis et al. [56] adopted a cascaded regression-like method with a Recurrent Neural Network (RNN) called Mnemonic Descent Method (MDM). In the MDM network, the CNNs are used to extract the patch features replacing traditional hand-crafted feature extractors such as SIFT [36] in SDM [64]. In addition, they introduce RNNs as memory units to share information across the cascaded levels. The recurrent module facilitates the joint optimization of the regressors by assuming that the cascades form a non-linear dynamical system.

2.4 Likelihood maps

Another interesting approach is training the CNN to predict likelihood maps (also called response maps, probability maps, voting maps or heat maps) as network output as originally proposed by [16, 17]. The value of each pixel on the likelihood maps could be represented as the probability of the existence of each facial landmark at the pixel. Zadeh et al. [67] proposed to use deep convolutional neural networks to produce a local response map and fit the model as Constrained Local Model (CLM). Since the deep encoder-decoder can establish an image-to-image mapping, Lai et al. [32] uses a fully convolutional network to predict an initial face shape instead of a mean face shape which is commonly used in cascaded regression [64, 9]. They introduced "Shape-Indexed Pooling" as a feature mapping function to extract local patch features of each point, which is then given to the regressor. In their first version, they used a fully-connected layer to sequentially regress the final shape while replacing it with a recurrent neural network in their second version, inspired by MDM [56]. In the work of Xiao et al. [62] an attention mechanism is used, where landmarks around the attention center are subject to a specific refinement procedure. Wang et al. [58] proposed an approach to detect multi-face landmarks by likelihood maps. With the help of a ROI pooling [21] branch, face detection is not required and non-face activations are eliminated over the entire likelihood maps. Another interesting work is proposed by Kowalski et al. [29] which predicts the transformation to a canonical pose and a feature image simultaneously with global point likelihood maps in a cascaded manner. All the networks in different stages share the information by taking the transformation parameters from the preceding stage.

2.5 Multi-task learning

Other recent works focus on multi-task CNN frameworks to obtain additional semantic information other than facial landmarks. In addition to the aforementioned TCDCN [75], Zhang et al. proposed a method called MTCNN [72], which adopted a three-stage structure composed of CNNs to perform together face detection, face classification and face alignment. They proposed a fast Proposal Network (P-Net) to produce facial region candidates and facial landmarks on low-resolution images in the first stage. After that, they refined these candidates in the next stage through a Refinement Network (R-Net) followed by Output Network (O-Net) to produce final bounding boxes and facial landmarks position with higher resolution inputs. In the work proposed by Ranjan et al. [47] multi-branch CNN named All-in-One CNN was trained to detect the face region, facial landmarks with their visibility, the head pose, smile probability, gender as well as the age of the person concurrently by sharing the same convolutional feature extractor.

2.6 Miscellaneous

Some works using deep CNN are harder to categorize. Belharbi et al. [4] treated the facial landmark detection as a structured-output problem. Güler et al. [1] proposed to detect the facial landmarks in a deformation-free space, which is defined by a 2-dimensional U-V mapping in 3D Morphable Models. They decouple the landmark position regression into two parts, referred as quantized classification and residual regression, which mean respectively positioning the band region and the relative residual difference to the band region. Quantized regression aims to predict a rough band position and the residual regression aims to predict the refined shape. This algorithm provides dense correspondence between a three-dimensional object template and the input image. They show that their output can provide a robust initialization for cascaded regression methods.

We witness a great change in this domain from predicting directly the point coordinates by fully-connected layers to predicting the landmark location by likelihood maps using fully convolutional networks and various other ideas. CNN are no longer simply considered as a learnable image feature extractor but rather a multi-functional tool for processing various types of information for face alignment.

3 2D landmarks tracking

Video face alignment, sometimes referred as sequential face alignment, aims at aligning a sequence of consecutive person-specific images by leveraging continuous information in the video.

Person-specific modeling is a direct idea for video face tracking since personal identity information remains unchanged. Chrysos et al. [12] proposed an off-line tracking pipeline to reinforce the tracking robustness in speech videos. Other than general face detection and alignment, they implemented a person-specific face detection using a Deformable Part Model [20] and a person-specific generative landmark localizer. It iteratively updates the generic/person-specific appearance variations and shape/appearance parameters in turn. The method is used to annotate the 300VW [53] dataset semi-automatically. Asthana et al. [3] reformulated the cascaded regression in

a parallel form to enable fast and efficient learning of each cascade level. The information retained by the next cascaded level is derived from the statistical distribution from the preceding cascade regressor. Another recent study on the use of cascaded regression is CCR [50] and iCCR [51]. The authors implemented a continuous regression method while reformulating it so that the algorithm does not require sampling over the perturbed shapes (e.g. flipping, rotation, scaling...). As a result, the computational cost is largely reduced compared to the traditional cascaded regression based methods.

Since Bayesian filters are popularly used in object tracking, a direct idea is to combine them (eg. Kalman filtering) with state-of-art landmark localizers. Pabhu et al. [46] used a Kalman filter to track the facial landmarks by head positions, head orientations and facial shapes in video sequences.

The 300VW workshop [53] proposed a challenging dataset specifically for the tracking of facial landmarks. Among all the methods, one of the leaders implemented a pose-specific cascaded regression method [66] and another adopted a progressive initialization [63] which aims to resolve the problem of bad initialization in extreme poses.

Inspired by incremental learning method [41], Peng et et al. [44] proposed to use a CNN with likelihood maps to evaluate the fitting results at the end of the network, which guarantee more robust results. RED-Net [42] was proposed to optimize the performance of video face alignment by disentangling the identity information and pose/expression information. The identity information can be considered invariant in a video while pose and expression information change in the time dimension. The author proposed a dual-path network using point likelihood maps which include one path to extract the identity information and the other path to learn the pose/expression information by using a RNN network.

Recently, Gu et al. [23] proposed to integrate a one-layer RNN at the end of a VGG network to track facial landmarks as well as head poses. They proved that Bayesian filters could be formulated as a linearly-activated RNN without bias. According to their results, tracking with RNN is more accurate and stable than frame-by-frame detections and state-of-art landmark localizers tracked by Kalman filter. Another algorithm using RNN called TSTN is proposed by Liu et al. [35] by adopting two network streams, spatial and temporal. The spatial stream learns to transform local facial patches to facial 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 RNN at the center to capture facial dynamics across the temporal dimension. This stream takes consecutive frames as input and then renders the temporal shape update. The final shape is determined by a weighted fusion of two streams shape updates. Hou et al. [25] uses a Long Short Term Memory (LSTM) module to guide the spatial estimation for the next stage just as MDM and simultaneously guide the estimation for the next frame.

Deep learning methods are not yet popular in video alignment because of their high complexity and size and memory constraints which are still a significant handicap for real-time detection on mobile platforms.

4 Datasets

The dataset is essential for the training stage of a facial landmark detection method. We will list here the most common 2D point annotated image datasets, 3D point annotated image datasets as well as annotated video datasets.

4.1 2D Datasets

2D point annotated image datasets can generally be categorized into two parts. Images taken under constrained conditions such as controlled lighting conditions, controlled poses etc. and images taken under unconstrained conditions, which are usually referred to as *in-the-wild* datasets.

Multi-PIE [22] is one of the largest constrained dataset with 337 subjects in 15 views, under 19 different illumination conditions and in six expressions. The facial landmarks are labeled with 68 points or 39 points.

XM2VTS [40] is a dataset which contains four recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. The dataset is annotated with 68 points and included in the **300W** challenge [49].

Among all *in-the-wild* dataset, the **300-W** [49] dataset (300 Faces in-the-Wild Challenge) is the most popular one in recent years. It combines several datasets such as **Helen** [33], **LFPW** [5], **AFW** [77] and a newly-introduced challenging dataset **iBug**. In total, it contains 3837 images and a private test set with 300 indoor and outdoor images respectively. All of the images are annotated with 68 points. It is quite common that people divide the dataset into two parts, the common subset, including the **LFPW** and **Helen** dataset, and the challenging dataset including **AFW** and **ibug**.

The **Menpo** [69] dataset, to the best of our knowledge, is the largest *in-the-wild* facial landmark dataset composed of 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 is proposed for the Menpo challenge workshop in 2017 in order to raise a even bigger challenge to test the robustness of the algorithms since it involves a big variation of poses, lighting conditions and occlusions.

4.2 Video Datasets

Video-based annotated datasets are used for research on video-based face alignment (sequential face alignment) taking into account past temporal information and identity consistency [42]. **300-VW** [53] has, to the best of our knowledge, the largest number of facial-point annotated vi-

Method	Database	NME(%)	$AUC_{0.08}$(%)
DRMF [2]	300W [49]	9.22	-
RCPR [8]	300W	8.35	-
ESR [9]	300W	7.58	43.12
SDM [64]	300W	7.52	42.94
ERT [28]	300W	6.40	-
LBF [48]	300W	6.32	-
CFSS [76]	300W	5.76	49.87/55.9*
CFAN [71]	300W	7.69	-
TCDCN [75]	300W	5.54	41.7*
TSR [38]	300W	4.99	-
RAR [62]	300W	4.94	-
DRR [32]	300W	4.90	-
MDM [56]	300W	4.05	52.12
DAN [29]	300W	3.59	55.33
2DFAN [7]	300W-private	-	66.90*
DenseReg+MDM [1]	300W	-	52.19

TABLE 1 – Performance comparison of different 2D alignment methods based on the IOD normalized error. The data are obtained from [75], [29] and the original publications (*) : The measure is from [7] using a threshold of 0.07 on the 300W-private dataset. The upper part of the table lists non-deep-learning methods and the lower part lists deep-learning methods.

Video face alignment method comparison on the 300VW dataset					
Method	ESR [9]	SDM [64]	CFSS [76]	PIEFA [45]	CFAN* [71]
NME(%)	7.09	7.25	6.13	6.37	6.64
FPS	67	40	10	-	20
Method	TCDCN* [75]	RED* [42]	RED-Res* [43]	RNN* [23]	TSTN* [35]
NME(%)	7.59	6.25	4.75	6.16	5.59
FPS	59	33†	18†	-	30

TABLE 2 – Comparison of the Mean Error(%) (normlized by face size) of different video face alignment methods on 300VW. The data is obtained from [42] and original publications. * indicates deep learning-based methods. † indicates that the runtime is measured on GPU

deos and frames. The dataset consists of 50 training videos and 64 videos for testing, which are further categorized into three scenarios according to different lighting conditions, expressions, head poses and occlusions. All of the frames are annotated in 68 points in a semi-automatic way.

The **Menpo 3D tracking** [68] dataset is the only dataset which annotate the 3D facial points in video by the 3DMM fitting algorithm [6]. This dataset contains 55 videos from the **300VW** dataset re-annotated in 3D, in addition to all of the images in **300W** and **menpo** re-annotated in the same way. The dataset provides not only the points in projected image space but also the points in 3D model space.

4.3 Evaluation Metric

In order to provide comparable results regardless of image size and camera focus, researchers measure the distance between the ground truth and the detection result by Normalized Mean Error (NME) :

$$e = \frac{\|S - S^*\|_2}{d^*} \ , \tag{2}$$

where S and S^* represents the detected shape and the shape of the ground truth respectively. d^* is a normalizing distance which could be inter-ocular distance (IOD), inter-pupil distance. Many researchers use the bounding box diagonal or geometric mean of image length and height as d^* instead if the distance between two eyes is too small on 3D/large pose datasets.

Another metric based on the NME is called 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. The Area Under Curve (AUC) provides a general qualitative result of how the algorithm performs at progressive mean errors :

$$AUC_\alpha = \int_0^\alpha f(e)de \ , \tag{3}$$

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 a better performance.
Failure rate is used to measure the robustness of an algorithm. A threshold of NME is chosen to be a threshold of failure and the proportion of failed detection is calculated to represent the capacity of handling the difficult images.

5 Comparison

In this section, we provide a comparison of different face alignment methods as well as different deep compression models. This comparison includes traditional cascaded regression methods and deep learning based face alignment methods.
Zhang et al. [75] and Kowalski et al. [29] both provide a good benchmark on several popular methods by measuring the normalized inter-ocular distance error. Table 1 shows the performance of 7 non-deep-learning 2D face alignment methods and 8 deep learning face alignment methods evaluated on 300W by the normalized inter-ocular distance error. We do not include the failure rate in the table since the choice of threshold is not objective. In general, the deep learning based algorithms outperform the others. However, all of the cascaded-regression based methods can run in real-time even with a Matlab implementation [59]. Some deep-learning based methods can achieve real-time detection on a powerful GPU or CPU but most runtimes on CPU are not satisfying.

6 Conclusion

Our paper reviewed recent deep learning-based 2D face alignment methods. Deep learning based algorithms outperform others in terms of precision. However, computation efficiency remains a big constraint especially for video face alignment.
Despite the fact that deep learning methods achieve excellent performance on many datasets, face alignment on limited resource platform has not been solved. One future research direction is to investigate compression methods such as Shuffle-Net [74]. Another direction is to focus on precision. Specific applications like beautification or animation demand high precision to give a perfect rendering.

Références

[1] Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I. : Densereg : Fully convolutional dense shape regression in-the-wild (2017)

[2] Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M. : Robust discriminative response map fitting with constrained local models. In : CVPR, pp. 3444–3451 (2013)

[3] Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M. : Incremental face alignment in the wild. In : CVPR, pp. 1859–1866 (2014)

[4] Belharbi, S., Chatelain, C., Hérault, R., Adam, S. : Facial landmark detection using structured output deep neural networks. arXiv preprint (2015)

[5] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N. : Localizing parts of faces using a consensus of exemplars. IEEE TPAMI **35**(12), 2930–2940 (2013)

[6] Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S. : 3d face morphable models" in-the-wild". arXiv preprint arXiv :1701.05360 (2017)

[7] 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) (2017)

[8] Burgos-Artizzu, X.P., Perona, P., Dollár, P. : Robust face landmark estimation under occlusion. In : ICCV, pp. 1513–1520 (2013)

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

[10] Çeliktutan, O., Ulukaya, S., Sankur, B. : A comparative study of face landmarking techniques. EURASIP Journal on Image and Video Processing **2013**(1), 13 (2013)

[11] Chrysos, G.G., Antonakos, E., Snape, P., Asthana, A., Zafeiriou, S. : A comprehensive performance evaluation of deformable face tracking “in-the-wild”. International Journal of Computer Vision pp. 1–35 (2014)

[12] Chrysos, G.G., Antonakos, E., Zafeiriou, S., Snape, P. : Offline Deformable Face Tracking in Arbitrary Videos. ICCV pp. 954–962 (2016). DOI 10.1109/ICCVW.2015.126

[13] Cootes, T.F., Edwards, G.J., Taylor, C.J. : Active appearance models. IEEE TPAMI **23**(6), 681–685 (2001)

[14] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J. : Active shape models-their training and application. Computer vision and image understanding **61**(1), 38–59 (1995)

[15] Cristinacce, D., Cootes, T.F. : Boosted regression active shape models. In : BMVC, vol. 2, pp. 880–889 (2007)

[16] Duffner, S., Garcia, C. : A connexionist approach for robust and precise facial feature detection in complex scenes. In : Fourth International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 316–321. Zagreb, Croatia (2005)

[17] Duffner, S., Garcia, C. : A hierarchical approach for precise facial feature detection. In : Compression et Représentation des Signaux Audiovisuels (CORESA), pp. 29–34. Rennes, France (2005)

[18] Edwards, G.J., Taylor, C.J., Cootes, T.F. : Interpreting face images using active appearance models. In : Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pp. 300–305. IEEE (1998)

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

[20] Felzenszwalb, P., McAllester, D., Ramanan, D. : A discriminatively trained, multiscale, deformable part model. In : CVPR, pp. 1–8. IEEE (2008)

[21] Girshick, R. : Fast r-cnn. In : The IEEE International Conference on Computer Vision (ICCV) (2015)

[22] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S. : Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)

[23] Gu, J., Yang, X., De Mello, S., Kautz, J. : Dynamic facial analysis : From bayesian filtering to recurrent neural network (2017)

[24] Hinton, G.E. : Training products of experts by minimizing contrastive divergence. Neural computation pp. 1771–1800 (2002)

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

[26] Huang, Z., Zhou, E., Cao, Z. : Coarse-to-fine face alignment with multi-scale local patch regression. arXiv preprint arXiv :1511.04901 (2015)

[27] Jin, X., Tan, X. : Face alignment in-the-wild : a survey. arXiv preprint arXiv :1608.04188 (2016)

[28] Kazemi, V., Sullivan, J. : One millisecond face alignment with an ensemble of regression trees. In : CVPR, pp. 1867–1874 (2014)

[29] Kowalski, M., Naruniec, J., Trzcinski, T. : Deep alignment network : A convolutional neural network for robust face alignment. arXiv preprint arXiv :1706.01789 (2017)

[30] Krizhevsky, A., Sutskever, I., Hinton, G.E. : Imagenet classification with deep convolutional neural networks. In : NIPS, pp. 1097–1105 (2012)

[31] Kumar, A., Ranjan, R., Patel, V., Chellappa, R. : Face alignment by local deep descriptor regression. arXiv preprint arXiv :1601.07950 (2016)

[32] Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., Yan, S. : Deep Recurrent Regression for Facial Landmark Detection pp. 1–13

[33] Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S. : Interactive facial feature localization. In : ECCV, pp. 679–692. Springer (2012)

[34] LeCun, Y., Bengio, Y., Hinton, G. : Deep learning. Nature pp. 436–444 (2015)

[35] Liu, H., Lu, J., Feng, J., Zhou, J. : Two-stream transformer networks for video-based face alignment. IEEE TPAMI (2017)

[36] Lowe, D.G. : Distinctive image features from scale-invariant keypoints. International journal of computer vision pp. 91–110 (2004)

[37] Luo, P., Wang, X., Tang, X. : Hierarchical face parsing via deep learning. In : CVPR, pp. 2480–2487. IEEE (2012)

[38] Lv, 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 : CVPR, pp. 3317–3326 (2017)

[39] Alabort-i Medina, J., Zafeiriou, S. : Bayesian active appearance models. In : CVPR, pp. 3438–3445 (2014)

[40] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G. : Xm2vtsdb : The extended m2vts database. In : Second international conference on audio and video-based biometric person authentication, vol. 964, pp. 965–966 (1999)

[41] Metaxas, D.N. : Sequential Face Alignment via Person-Specific Modeling in the Wild pp. 107–116 (2016). DOI 10.1109/CVPRW.2016.194

[42] Peng, X., Feris, R.S., Wang, X., Metaxas, D.N. : A recurrent encoder-decoder network for sequential face alignment. In : ECCV, pp. 38–56. Springer (2016)

[43] Peng, X., Feris, R.S., Wang, X., Metaxas, D.N. : Rednet : A recurrent encoder-decoder network for video-based face alignment (2018)

[44] Peng, X., Hu, Q., Huang, J., Metaxas, D.N. : Track Facial Points in Unconstrained Videos pp. 1–13 (2016). URL http ://arxiv.org/abs/1609.02825

[45] Peng, X., Zhang, S., Yang, Y., Metaxas, D.N. : Piefa : Personalized incremental and ensemble face alignment. In : ICCV, pp. 3880–3888 (2015)

[46] Prabhu, U., Seshadri, K., Savvides, M. : Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models. Trends and Topics in Computer Vision, Springer Berlin Heidelberg pp. 86–99 (2012)

[47] Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R. : An all-in-one convolutional neural network for face analysis. CoRR (2016)

[48] Ren, S., Cao, X., Wei, Y., Sun, J. : Face alignment at 3000 fps via regressing local binary features. In : CVPR, pp. 1685–1692 (2014)

[49] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M. : 300 faces in-the-wild challenge : The first facial landmark localization challenge. In : ICCV Workshops, pp. 397–403 (2013)

[50] Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G., Valstar, M. : Cascaded continuous regression for real-time incremental face tracking. In : ECCV, pp. 645–661. Springer International Publishing (2016)

[51] Sánchez-Lozano, E., Tzimiropoulos, G., Martinez, B., De la Torre, F., Valstar, M. : A functional regression approach to facial landmark tracking. arXiv preprint arXiv :1612.02203 (2016)

[52] Shao, Z., Ding, S., Zhao, Y., Zhang, Q., Ma, L. : Learning deep representation from coarse to fine for face alignment. arXiv preprint arXiv :1608.00207 (2016)

[53] 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 : ICCV Workshops, pp. 50–58 (2015)

[54] 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 preprint arXiv :1503.08843 (2015)

[55] Sun, Y., Wang, X., Tang, X. : Deep convolutional network cascade for facial point detection pp. 3476–3483 (2013)

[56] Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S. : Mnemonic descent method : A recurrent process applied for end-to-end face alignment pp. 4177–4187 (2016)

[57] Tzimiropoulos, G., Pantic, M. : Optimization problems for fast aam fitting in-the-wild. In : Proceedings of the IEEE international conference on computer vision, pp. 593–600 (2013)

[58] Wang, L., Yu, X., Metaxas, D. : A coupled encoder-decoder network for joint face detection and landmark localization. In : Automatic Face and Gesture Recognition (FG), IEEE 12th International Conference on (2017)

[59] Wang, N., Gao, X., Tao, D., Yang, H., Li, X. : Facial feature point detection : A comprehensive survey. Neurocomputing (2017)

[60] Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P. : Facial landmark detection with tweaked convolutional neural networks. IEEE TPAMI (2017)

[61] Wu, Y., Ji, Q. : Discriminative deep face shape model for facial point detection. International Journal of Computer Vision pp. 37–53 (2015)

[62] Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A. : Robust facial landmark detection via recurrent attentive-refinement networks. In : ECCV, pp. 57–72. Springer (2016)

[63] Xiao, S., Yan, S., Kassim, A.A. : Facial Landmark Detection via Progressive Initialization. ICCV pp. 986–993 (2016). DOI 10.1109/ICCVW.2015.130

[64] Xiong, X., De la Torre, F. : Supervised descent method and its applications to face alignment. In : CVPR, pp. 532–539 (2013)

[65] Yang, H., Jia, X., Loy, C.C., Robinson, P. : An empirical study of recent face alignment methods. arXiv preprint arXiv :1511.05049 (2015)

[66] Yang, J., Deng, J., Zhang, K., Liu, Q. : Facial Shape Tracking via Spatio-Temporal Cascade Shape Regression. ICCV pp. 994–1002 (2016). DOI 10.1109/ICCVW.2015.131

[67] Zadeh, A., Morency, L.p. : Deep Constrained Local Models for Facial Landmark Detection

[68] Zafeiriou, S., Chrysos, G.G., Roussos, A., Ververas, E., Deng, J., Trigeorgis, G., Crispell, D., Bazik, M., Xiong, P., Li, G., et al. : The 3d menpo facial landmark tracking challenge. In : ICCV Workshop, vol. 5 (2017)

[69] Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J. : The menpo facial landmark localisation challenge : A step towards the solution. In : CVPR Workshops, pp. 2116–2125. IEEE (2017)

[70] Zhang, J., Kan, M., Shan, S., Chen, X. : Occlusion-free face alignment : deep regression networks coupled with de-corrupt autoencoders pp. 3428–3437 (2016)

[71] Zhang, J., Shan, S., Kan, M., Chen, X. : Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment pp. 1–16 (2014)

[72] Zhang, K., Zhang, Z., Li, Z., Qiao, Y. : Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters pp. 1499–1503 (2016)

[73] Zhang, S., Yang, H., Yin, Z.P. : Transferred deep convolutional neural network features for extensive facial landmark localization. IEEE Signal Processing Letters **23**(4), 478–482 (2016)

[74] Zhang, X., Zhou, X., Lin, M., Sun, J. : Shufflenet : An extremely efficient convolutional neural network for mobile devices. arXiv preprint arXiv :1707.01083 (2017)

[75] Zhang, Z., Luo, P., Loy, C.C., Tang, X. : Facial landmark detection by deep multi-task learning pp. 94–108 (2014)

[76] Zhu, S., Li, C., Change Loy, C., Tang, X. : Face alignment by coarse-to-fine shape searching. In : CVPR, pp. 4998–5006 (2015)

[77] Zhu, X., Ramanan, D. : Face detection, pose estimation, and landmark localization in the wild. In : CVPR, pp. 2879–2886. IEEE (2012)