DOI: 10.21917/ijsc.2019.0256

DETECTION OF ACCURATE FACIAL DETECTION USING HYBRID DEEP CONVOLUTIONAL RECURRENT NEURAL NETWORK

M. Sivaram, V. Porkodi, Amin Salih Mohammed and V. Manikandan
Department of Information Technology, Lebanese French University, Iraqi Kurdistan

Abstract

Facial Landmark discovery is an imperative issue in numerous PC vision applications about appearances. It is extremely testing as human faces in wild conditions regularly present expansive varieties fit as a fiddle because of various stances, impediments or demeanors. Profound neural systems have been connected to take in the guide from face pictures to confront shapes. To the best of our insight, Recurrent Neural Network (RNN) has not been utilized in this issue yet. In this paper, we propose a technique which uses RNN and Deep Neural Network (DNN) to take in the face shape. To start with, we design a system utilizing Convolutional Neural Network (CNN) to get the underlying Landmark estimation of appearances. At that point, we utilize feed-forward neural systems for neighborhood look where a segment based seeking technique is investigated. By utilizing LSTM-CNN-RNN, the underlying estimation is more dependable which makes the accompanying segment based pursuit doable and exact. Tests demonstrate that the worldwide system utilizing CNN-LSTM-RNN shows signs of improvement results than past systems in the two recordings and single picture. Our technique beats the cutting edge calculations particularly regarding fine estimation of Landmark spots.

Keywords:

Facial landmark, Deep Neural Network, Recurrent Neural Network, Convolutional Neural Network

1. INTRODUCTION

Facial points include noteworthy facial points, such as the center of the eyes, the tip of turmoil or mouth edges. In several pc view applications, face recognition, confrontation, embellishment, head present estimates and animation age are a main issue in the programmed field of facial landmarks. The question is particularly troubling when considering non-frontal faces or faces with exceptional looks. Indeed, the varieties of posture or appearance causes nonlinear disfigures that are difficult to manage in the face states. It is therefore still a testing company for strong and accurate limits in true applications.

All these techniques can be divided into two classifications if they've all been completed: 2D or 3D photography. In improving business depth and RGB cameras such as Microsoft Kinect, 3D image techniques in genuine applications have been more practical and the depth of the data helps find the focus [1-4]. However, we have to process 2D images in spite of everything with a quick increase in the number. This paper examines the limitation of marks on 2D photographs and 3D photographic techniques with depth data. As shown by how to assess the majority of pictorial arrangement strategies, they can be grouped into two different classes: fit the form by streamlining or taking the guide straight away.

In this paper, we want to manage confrontation by using fleeting data in image groupings, as well as more still photographic data. Fleeting data is used in a TSR display to manage the issue of facial mark recognition in records [51] to improve the flow between nearby cartridges. Initially, the deep learning strategies develop intermittent neural systems [36]-[38]. In two ways, it is critical to use world data in video face recognition. It can start by strengthening the forecast. The problem of recognition of landmarks is to anticipate expanding positions or extraordinary articulation. In any case, in recordings, all of these extraordinary conditions are consistently increased by unbiased behavior or face posture. The anticipated area can be more confident by using past or accompanying data by utilizing CNN-RNN in facial landmark issues in video successions. Second, expectations can be made more stable. The sound of landmark spots is critical for the visual impact in the preparation of recordings. The current expectations of Landmark are influenced by the repetitive associations of CNN-RNN by its adjacent borders to increase the output of various arcs. Again, if a man keeps its exterior appearance for a long time anomalous, at the same time the expectation of landmarks could be dependable on the contours of an adjacent (as well as the unbiased articulation). For this reason, we use Long Short Term Memory (LSTM) to determine the data needed.

2. RELATED WORK

The Active Shape Model (ASM) [5] is one of the main methodologies for arranging the face. With ASM, the direction of the landmark focuses on a form vector in a face image frame and a PDM speaks about the inconsistency of these form vectors. PDM is a generative model of vector form which shows the non-existent evolution of facet shape and the relative changes in displaying inflexible ones, using Principal Component Analysis (PCA). The fit capacity depending on the ordinary profile to the limit is received at this point in order to fit in the parameters of the assessed form to the actual form. The Active Appearance Model [6]-[9] is an all-embracing ASM model, in which facial area data are taken into account. Shape and surface with another PCA are coordinated to display the appearance. Based on covetous calculations, ASM and AAM are touchy about the position of the Landmark in an image. The Constricted Local (CLM) Model was proposed in the light of ASM in [10]. It mounts the face surface in free layouts of the shape and still uses an ASM to show distortion of shape. Later techniques will manufacture CLMs or non-direct twist models that rely on neighborhood designer parts removed from every location and will use different fitting processes to get the form evaluated [11]-[14].

The second route then again is directly to the guide. To understand this objective, there are two fundamental systems. The recurrence technique is one procedure. Early shooting of the recurrence technique also depends on a parameter display and limit display errors in preparation [15]-[17]. A falling reciprocal technique has been launched later on, which learns from a rough estimation of position that regressors anticipate the face shape directly [18]-[20]. In this technique the green regressors [21] and

the supported recurrence [22], [23] are all misused to build the regressors. In these reciprocal techniques, for example, near-by highlights are the scale-invariant change in components (SIFT) [24] and paired highlights [25]-[29]. Furthermore, Artizzu et al. [30] proposed a Robust Cascaded Pose Regression (RCPR) that would reduce the occurrence of anomalies by identifying unequivocal impediments. In the other case, the deep neural systems should be used to specifically follow the nonlinear guide. As deep neural systems have shown notable results in learning from a variety of companies, they are also familiar with landmarks. By using the Deep Revolutionary Neural Network (DCNN) [31], Sun et al. proposed a decreased methodology.

The main system maps the world element's underlying shape. In two phases the progressive systems thus refine the baselines. In all cases every aspect of the strategy is freely refined, after a global evaluation, so that the global shape can be overcome and precariousness prevented. Moreover, Zhang et al. suggested the CFAN [32] approach with a variety of auto-encoding systems with stacked frontal arrangements. The focus in the near vicinity is mutually rejuvenated, and this remains the international forms learned by the main system. This results in better than DCNN results. Pose-Induced Automotive Encoder Network [33] has been proposed by Chen et al. which provides a landmark area and posture data which can be used to identify landmarks accurately in different grade levels. Moreover, a thorough approach to various learning tasks has been proposed, in which the strength of deep neural systems is shown to adjust these nonlinear capacities in the meantime [34]. While such a nonlinear ability is shown to be strongly adaptive by a deep neural system, it also can be shrunk if there is insufficient preparation.

Former management groupings, for example Graves and Schmidhuber RNNs are available in full multi-dimensional variations [59]. More recently, [19] [60] investigate the spatial conduction of a RNN on an element that is a defining location of convolutions. These papers analyze separately spatial RNNs for semantic division and image order undertakings.

In [57] the authors use spatial RNNs to handle logical highlights for the purpose of the discovery in question. RNNs without worldly arrangements are used in these strategies. The key idea is to employ repetitive associations in RNNs in order to show the links of different highlights in order to strengthen future associations in RNNs. Only indistinct images are customarily used when it comes to identification of facial marks. Since human appearance in diverse shading places is frequently noticeable under various conditions, data on essential shading spaces, especially surface data, may be lost. Furthermore, the technology can perform better once improved images with strong nearby inclination data are used. We use repeated LSTM to show different examples and in the meantime take feed forward associations in the predictive area of landmarks.

As in [19] [31]-[33], [35], [40], [41], [49], [50] our method generally follows the cascaded architectures. Contrary to the method used in the past, we use LSTM for the first cascaded module to find the initial location and in the second stage we use a component based search strategy. In the third stage, we use the network similar to [33] to further refine the results.

3. GLOBAL NETWORK USING LSTM FOR FACIAL LANDMARK DETECTION

In this study, we will discuss the architecture of the LSTM network used to find landmarks worldwide in our method. By using LSTM to search globally, we can estimate landmarks more accurately than original neural networks. There are two advantages. First, since the initial estimate is reliable, we can use a component-based strategy that can achieve fine places of reference and at the same time maintain facial shape. Secondly, the required phases of the cascaded process are reduced. In the cascading procedure our method has just three stages, while the previous methods often have five.

The main contributions of this paper are summarized as follows:

- As far as our own knowledge is concerned, this is the first job to use Recurrent Neural Network to identify visual and individual pictures.
- We assess the efficiency of CNN-LSTM-RNN in sequence images and single pictures to detect facial signs and demonstrate the efficiency and robustness of CNN-RNN in the treatment of these problems with certain previous methods.
- A component based search strategy is developed after obtaining the initial points of reference with the LSTM network, which is the initial place of reference. This helps to get more precise co - ordinates and keeps the face shape.

The Fig.1 illustrates the architecture of our method. The CNN-RNN provides a number of face inputs for the initial positioning of the landmarks. Then two local networks using the Deep Neural Networks (DNN) are used to refine the results based on the results of the global network. The three networks constitute a cascaded procedure in which the image resolution shifts from low to high.

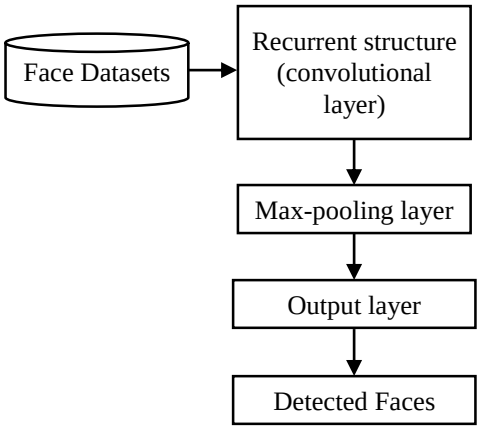

Fig.1. Architecture of Proposed Method

3.1 RNN

RNN have a long history in the neural artificial system community; however, sequential data models, like handwriting recognition and speech recognition, are among the more successful applications. The ability to use contextual information when mapping input to output sequences is an important benefit of recurring neural networks. However, when it passes through repeated network connections, the influence of a given input on

the hidden layer either decays or blows up exponentially. The fluttering gradient problem [39] is often called this and the RNN Long Short Term Memory (LSTM) architecture has been proposed to deal with this problem [34].

The so-called memory blocks in LSTM substitute the hidden units of traditional RNN. Each memory block includes an input, output and forget gates of one or more self-connected memory cells and three multiplier units. These doors control the memory block behaviour. Since these gates allow operations to be written, read and reset in a memory block, the LSTM block can be seen as a differentiable version of the digital computer's memory chips.

The RCNN and some sparse coding models have a very interesting relation, with fixed-point updates being used inferences. Implicitly recurrent neural networks are defined by the iterative optimization procedures. Please note that supervised learning techniques can be included in a sparse coding model's uncontrolled learning framework. But the sparse coding models have not been competitive with CNN for recognition of objects.

3.2 GLOBAL RCNN NETWORK

RNN is not restricted to video sequence management. The pattern along the time axis is missing for a single face image. However, RNNs are also spatially available to address various characteristics. The template can be replaced by a series of patterns from the original image. To achieve this, we use two different ways. The three RGB channels should be used separately as one simple way. The relations between the different color spaces are modeled on this network.

Weights are divided into various times in the RNNs. In addition, the same feed-forward network is like a deep network, regardless of the reciprocal connections, to implement the project functionality of the image towards the initial landmarks. Meanwhile, the recurring connections across different stages provide information to the current color space in hidden layers of other colour. This can compensate for the absence of information in some color areas and use the RGB pictures in full. Features may of course be useful in some color space and some may be useless. The gates in LSTM are therefore expected to control the data to other inputs.

In addition, with regard to facial detection, the points of reference often fall in the area of high gradient data. Thus, the network performance can be further improved by using improved images with stronger gradient information. To get the image sequences, we use several image enhancement methods. The first input to the network is the original image. The image is then used as the second input following histogram equalization.

We use gradient operators to enhance the image with enhanced edge information. In order to calculate the gradient intensity for each pixel, gradient operators such as the Sobel operator are used first. Each pixel value is then added to get the improved image by the corresponding intensity of the gradient.

4. EXPERIMENTS

In this section, the experimental conditions, including datasets and comparison methods, are first illustrated. Then we assess the efficacy of our method for both video and image sequences. First, we show how effective the global network is when comparing Recurrent Neural Networks with traditional networks. Then we compare our method with the latest algorithms.

4.1 DATABASE

The 2-D Face Warehouse [43] and 300-VW database [52]-[54] are used to evaluate the efficiency of our video sequence method. Indoor environments, the images in Face Warehouse are collected. There are different expressions for each person, including the neutral expression and other expressions like mouth-opening, smile and kiss.

The images can be considered as a set of images from a video sequence dispersed. While the images are collected indoor, face variations in positions and expressions make it difficult to detect landmarks. The Wild Facial Watch 300 Videos (300-VW) database includes 114 videos with IBUG landmarks in each video frame. The training program includes 50 videos and the remaining three categories of the remaining videos.

In the first category, there are videos of people in different head positions recorded in light conditions. The second category contains uncontrolled videos of people, displaying arbitrary expressions in different head poses, but without major occlusions. The third category includes video footage of people in totally unlimited conditions including lighting conditions, occlusions, make-up, expression, head pose etc.

The LFPW, HELEN, AFW and IBU G databases are used to evaluate the efficiency of our single image method in four parts: [44], [55], [56]. The wild conditions collect LFPW. It consists of 1000 training images and 200 test images with a wide varying position, expression, occlusion, etc.

The data increase is based on two main reasons. Firstly, because the number of labeled training data is limited, many images are needed for the network to be prevented from overfitting and a stronger network for various circumstances in unrest. Second, we need different images for local searches to train local networks than data for training in the world's first global network.

The training data is rotated, scaled and shifted randomly to obtain the increased data. Images are a set of time-series data in video sequences. For training we used the FaceWarehouse database [42]. We change the time of each training data so that we can make good use of the relationship between different positions and expressions and build a robust network.

4.2 EXPERIMENTAL SETTINGS

The global network and local networks have four layers with three hidden layers and a linear output layer for regression. LSTM or BLSTM units are used in hidden layers for recurring neural connections. The logistic function is used to activate the hidden units for neural feed-in networks or auto-encoder networks. The resolution of the face image in the global network is 50×50. In the successive local networks, the resolution of the face images is increased and increased. Each component patch is extracted from the initial image in the component-based search and then translated into 40×20 pixels. For the various components, the numbers of hidden units differ somewhat as the number of landmarks varies. In the 3rd phase, the face images are all reduced to 160×160 pixels, and each landmark has a resolution of 15×15

pixels for patches extracted. There are 100, 40 and 15 hidden units respectively in this network.

We compare our method with a few state-of-the-art methods, i.e. TCDCN [34], ERT [58], DRMF [45], RCPR [30], CFAN [32], DCNN [31], SDM [20], Yu et al. [46] and Zhu et al. [47]. We have to predict 74 landmarks for comparisons on the FaceWarehouse database, and the terms on this database are unique. So only RCPR is used for fair comparison here since they provide the training code. We use our approach to predict 68 points and use the 66 points to compare them to 300-W data bases, as DRMF and Yu et al predict 66 facial markpoints. 49 points, which omit the contour of the face are foreseen for SDM. For comparison, we use the 49-point prediction result. The CFAN and TCDCN codes are used directly to test 68 sites. To measure the performance, the test error and the normalized root mean NRMSE error are used. The distance between the centers of the eyes is normalized for NRMSE. Currennt [48] is a toolkit for recurring CUDA-open source neural networks, building up the code for LSTM and BLsTM.

4.3 PERFORMANCE EVALUATION

First, we will assess CNN-RNN effectiveness for both the video sequence and the individual image on a global network. The network can be separately trained via the SAE, the FNN, and CNN-RNN Stacked Auto-Encoder Networks (SAE). For these three types of networks, the same training data is used. Four types of recurrent connections are used for LSTM and BLSTM networks in order to fully explore the function of recurrent connections. Three hidden layers of LSTM (1), LSTM (2) and LSTM (3) are added to the current connections. In the three hidden layers marked as LSTM (1-3), recurring connections are then used. For BLSTM, the marks are the same. If an image's NRMSE is below a given number, we define that image prediction to succeed.

SAE and NN are trained directly with grayscale images for FaceWarehouse research. The same grayscale images are used to train LSTM and BLSTM. The Table.1 shows the results. The performance can be seen in comparison with traditional networks significantly improved by using CNN-RNN. When the demand for exactness is high (NRMSE<0.1), the performance is improved by CNN-RNN by more than 10%. This demonstrates CNN-RNN effectiveness in handling such time series tasks. BLSTM always has better than LSTM, indicating that two temporal contexts are both useful to predict current image landmarks. Moreover, BLSTM (1-3) and LSTM (1-3) are always more successful than other CNN-RNN types. This shows that the use of contextual information by recurring links is also important here. In order to more clearly show the benefit of using CNN-RNN, we display the results as a global network using CNN-RNN and NN. It is clear that in this issue, CNN-RNN predicts the sites more precisely than NN.

We conduct experiments on the LFPW database to investigate the performance of CNN-RNN in the single image. For networking, train systems from LFPW, HELEN and the entire range of AFW databases are used. For testing the test set is used in the LFPW database. Two experimental groups are carried out. The 3 RGB channels are used as inputs to CNN-RNN in the first experiment. We use the pictures of three RGB channels and original pictures to train the neural supply networks respectively to find out if CNN-RNN is useful in this situation.

The NN (Red), NN (Green), NN (Blue), and NN (RGB) have been marked. Three pictures of RGB channels (without pictures) will be used for LSTM and BLSTM. The Table.2 shows the results. The results. It is clear that the network can be trained by using a red, green or blue channel compared to the original images. The signals in various images can be strong in different color channels. This is understandable. One can weaken the gradient information separately, which is important for the recognition of a landmark.

Interestingly, in CNN-RNN the performance is lower than in original images when recurring connections are used separately in 3 hidden layers to form a feedforward network. However, when we use the entire recurring connections, the efficiency is greatly improved and NN (RGB) is even better. This shows that the relationship between three color channels is sufficiently utilized by CNN-RNN. The combination of three channels is fixed in the original images, while recurrent connections are excavated in CNN-RNN.

The Table.3 shows the classification performance of various models and the results shows that the proposed CNN-RNN is efficient in handling the detection of faces than the existing methods.

Table.1. Performance of different CNN-RNN types in the FaceWarehouse database compared to traditional deep neural networks

Network Type	Testing Error	Success Rate (%) (NRMSE<0.1)	Success Rate (%) (NRMSE<0.2)
SAE	955.3	63.7	89.2
NN	935.3	63.7	89.9
LSTM(1)	810.8	77.1	95.3
LSTM(2)	801.2	77.2	95.5
LSTM(3)	818.9	76	95.2
LSTM(1-3)	780.4	78.5	98.5
BLSTM(1)	795.1	77.5	95.9
BLSTM(2)	802	77.4	95.9
BLSTM(3)	796.1	77.1	96
BLSTM(1-3)	750.4	79.9	99

Table.2. Performance of various RNN types compared to traditional deep - network images from RGB channels in the LFPW database

Network Type	Testing Error	Success Rate (%) (NRMSE<0.1)	Success Rate (%) (NRMSE<0.2)
NN (Red)	175.3	62.5	92
NN (Green)	177.8	62.5	92.4
NN (Blue)	189.6	61.2	91.5

NN (RGB)	129.6	75.9	98.2
LSTM(1)	161	68.8	95.6
LSTM(2)	158.9	68.3	95.6
LSTM(3)	155.6	69.2	96
LSTM(1-3)	111.3	75.5	98.2
BLSTM(1)	141.4	69.2	96.4
BLSTM(2)	139.2	68.8	96
BLSTM(3)	135.7	71.9	96.9
BLSTM(1-3)	111.1	77.2	98.7

Table.3. Classification performance of various models

Methods	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**AUC or ROC**
Logistic Regression	0.94	0.93	0.93	0.9066	0.8891
SGD Classifier	0.87	0.87	0.87	0.8726	0.868
Random Forest Classifier	0.98	0.98	0.98	0.9839	0.9845
AdaBoost Classifier	0.98	0.98	0.98	0.9823	0.9823
2-layer NN RelU + Adam	0.95	0.95	0.95	0.9496	0.9475
Logistic Regression	0.99	0.99	0.99	0.9859	0.9862
SGD Classifier	0.95	0.94	0.94	0.9433	0.9443
Random Forest Classifier	0.99	0.99	0.99	0.9937	0.9938
AdaBoost Classifier	1	1	1	0.9981	0.9981
2-layer NN RelU + Adam	0.99	0.99	0.99	0.9878	0.9879
Logistic Regression	0.92	0.91	0.91	0.9094	0.9098
SGD Classifier	0.9	0.9	0.9	0.903	0.9031
Random Forest Classifier	0.99	0.99	0.99	0.9859	0.9859
Proposed Model	0.99	0.99	0.99	0.9865	0.9865

5. CONCLUSIONS

We use a sequence of recurring neuronal networks and feed-forward neural nets in an over-to-finish architecture to identify the problem to deal with nonlinear deformations in frontal shapes of face imagery. At first CNN-RNN network maps the face image's input directly to the estimated face form in low resolution. The subsequent local network then takes the components of faces as inputs based on the initial network output to obtain more accurate, higher resolution landmarks. Finally, by removing a small patch of each landmark and refining all the sites together, the local network adjusts results. Our method obtains impressive results when dealing with face images in videos through the full use of the time series. Moreover, our method also achieves good results on still images by using more images from RGB channels or enhanced images. In comparison with state-of - the-art methods, such as CFAN, SDM, TCDN, RCPR, the effectiveness of our method is validated on three databases in wild as well as interior conditions.

REFERENCES

[1] T. Weise, S. Bouaziz, H. Li and M. Pauly, "Realtime Performance-based Facial Animation", *ACM Transactions on Graphics*, Vol.30, No. 4, pp. 71-77, 2011.

[2] Q. Cai, D. Gallup, C. Zhang and Z. Zhang, "3D Deformable Face Tracking with a Commodity Depth Camera", *Proceedings of 11th International Conference on European Conference on Computer Vision*, pp. 229-242, 2010.

[3] G. Fanelli, M. Dantone, and L.V. Gool, "Real Time 3D Face Alignment with Random Forests-based Active Appearance Models", *Proceedings of 10th International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1-8, 2013.

[4] Z.Z. Zhang, W. Zhang, J.Z. Liu and X.O. Tang, "Multiview Facial Landmark Localization in RGB-D Images via Hierarchical Regression With Binary Patterns", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 9, pp. 1475-1485, 2014.

[5] T. Cootes, C.J. Taylor, D.H. Cooper and J. Graham, "Active Shape Models-Their Training and Application", *Computer Vision and Image Understanding*, Vol. 61, No. 1, pp. 38-59, 1995.

[6] T. Cootes, G.J. Edwards and C.J. Taylor, "Active Appearance Models", Available at: https://www.cs.cmu.edu/~efros/courses/AP06/Papers/cootes-eccv-98.pdf.

[7] I. Matthews and S. Baker, "Active Appearance Models Revisited", *International Journal of Computer Vision*, Vol. 6, No. 2, pp. 135-164, 2004.

[8] T. Cootes, G.J. Edwards and C.J. Taylor. "Active Appearance Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, pp. 681-685, 2001.

[9] X. Liu, "Generic Face Alignment using Boosted Appearance Model", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[10] D. Cristinacce and T. Cootes, "Automatic Feature Localization with Constrained Local Models", *Pattern Recognition*, Vol. 41, No. 10, pp. 3054-3067, 2008.

[11] D. Cristinacce and T. Cootes, "Feature Detection and Tracking with Constrained Local Models", *Proceedings of International Conference on British Machine Vision Conference*, pp. 929-938, 2006.

[12] J. Saragih, S. Lucey and J. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift", *International Journal of Computer Vision*, Vol. 91, No. 2, pp. 200-215, 2011.

[13] Y. Wang, S. Lucey and J. Cohn, "Enforcing Convexity for Improved Alignment with Constrained Local Model", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[14] L. Gu, and T. Kanade, "A Generative Shape Regularization Model for Robust Face Alignment", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 413-426, 2008.

[15] P. Dollar, P. Welinder and P. Perona, "Cascaded Pose Regression", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1078-1085, 2010.

[16] T.C. Patrick Sauer and C. Taylor, "Accurate Regression Procedures for Active Appearance Models", *Proceedings of International Conference on European Conference on Computer Vision*, 2011, pp. 1-30, 2011.

[17] J. Saragih and R. Goecke, "A Nonlinear Discriminative Approach to AAM Fitting", *Proceedings of 7th IEEE International Conference on Computer Vision*, pp. 1-8, 2007.

[18] X.D. Cao, Y.C. Wei, F. Wen and J. Sun, "Face Alignment by Explicit Shape Regression", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2887-2894, 2012.

[19] F. Visin et al., "ReNet: A Recurrent Neural Network based Alternative to Convolutional Networks", Available at: https://arxiv.org/pdf/1505.00393.pdf.

[20] Y. Sun, Q. Liu, H. Lu, "Low Rank Driven Robust Facial Landmark Regression", *Neurocomputing*, Vol. 151, pp. 196-206, 2015.

[21] M. Ozuysal, M. Calonder, V. Lepetit and P. Fua, "Fast Key-Point Recognition using Random Ferns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 3, pp. 448-461, 2009.

[22] N. Duffy and D.P. Helmbold, "Boosting methods for Regression", *Machine Learning*, Vol. 47, No. 2, pp. 153-200, 2002.

[23] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", Available at: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf.

[24] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.

[25] M. Calonder, V. Lepetit, C. Strecha and P. Fua, "Brief: Binary Robust Independent Elementary Features", *Proceedings of 10th IEEE International Conference on Computer Vision*, pp. 778-792, 2010.

[26] V. Lepetit, P. Lagger and P. Fua, "Randomized Trees for Real-Time Keypoint Recognition", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 775-781, 2005.

[27] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF", *Proceedings of International Conference on Computer Vision*, pp. 2564-2571, 2011.

[28] S. Leutenegger, M. Chli and R. Siegwart, "Brisk: Binary Robust Invariant Scalable Keypoints", *Proceedings of International Conference on Computer Vision*, pp. 2548-2555, 2011.

[29] A. Alahi, R. Ortiz and P. Vandergheynst, "FREAK: Fast Retina Keypoint", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 510-517, 2012.

[30] X.P. Burgos-Artizzu, P. Perona and P. Dollar. "Robust Face Landmark Estimation Under Occlusion", *Proceedings of International Conference on Computer Vision*, pp. 1513-1520, 2013.

[31] Y. Sun, X.G. Wang and X.O. Tang, "Deep Convolutional Network Cascade for Facial Point Detection", *Proceedings of International Conference on Computer Vision*, pp. 3476-3483, 2013.

[32] J. Zhang, S.G. Shan, M.N. Kan and X.L. Chen, "Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment, *Proceedings of International Conference on European Conference on Computer Vision*, pp. 1-16, 2014.

[33] Y. Chen, W. Luo and J. Yang. "Facial Landmark Detection via Pose-Induced Auto-Encoder Networks", *Proceedings of International Conference on Image Processing*, pp. 27-30, 2015.

[34] Z.P. Zhang, P. Luo, C.C Loy and X.O. Tang, "Facial Landmark Detection by Deep Multi-Task Learning", *Proceedings of International Conference on European Conference on Computer Vision*, pp. 94-108, 2014.

[35] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computing*, Vol. 9, No. 8, pp. 1735-1780, 1997.

[36] A. Graves et al., "A Novel Connectionist System for Unconstrained Handwriting Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 5, pp. 855-868, 2009.

[37] C. Plahl, M. Kozielski, R. Schluter and H. Ney, "Feature Combination and Stacking of Recurrent and Non-Recurrent Neural Networks for LVCSR", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 6714-6718, 2013.

[38] M. Wollmer et al., "Online Driver Distraction Detection using Long Short-Term Memory", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 12, No. 2, pp. 574-582, 2011.

[39] S. Hochreiter, Y. Bengio, P. Frasconi and J. Schmidhuber, "*Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*", IEEE Press, 2001.

[40] M. Schuster and K.K. Paliwal, "Bidirectional Recurrent Neural Networks", *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[41] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures", *Neural Networks*, Vol. 18, No. 5-6, pp. 602-610, 2005.

[42] A. Maas et al., "Recurrent Neural Networks for Noise Reduction in Robust ASR", Available at: http://www1.icsi.berkeley.edu/~vinyals/Files/rnn_denoise_2012.pdf.

[43] C. Cao, Q. Hou and K. Zhou, "Displaced Dynamic Expression regression for Real-Time Facial Tracking and Animation", *Proceedings of ACM Conference on Special Interest Group on Computer Graphics*, Vol. 33, No. 4, pp. 142-147, 2014.

[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "The First Facial Landmark Localization Challenge", *Proceedings of IEEE International Conference on Computer Vision*, pp. 41-52, 2013.

[45] A. Asthana, S. Zafeiriou, S. Cheng and M. Pantic, "Robust Discriminative Response Map Fitting with Constrained Local Models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444-3451, 2013.

[46] X. Yu, J. Huang, S. Zhang, W. Yan, D. N. Metaxas, "Pose-Free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model", *Proceedings of*

International Conference on Computer Vision and Pattern Recognition, pp. 1944-1951, 2013.

[47] X. Zhu and D. Ramanan, "Face Detection, Pose Estimation, and Landmark Localization in the Wild", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 2879-2886, 2012.

[48] F. Weninger, J. Bergmann and B. Schuller, "Introducing Currennt-The Munich Open-Source CUDA Recurrent Neural Network Toolkit", *Journal of Machine Learning Research*, Vol. 16, pp. 547-551, 2015.

[49] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman and N. Kumar, "Localizing Parts of Faces using a Consensus of Exemplars", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 545-552, 2013.

[50] V. Le, J. Brandt, Z. Lin, L. Bourdev and T.S. Huang, "Interactive Facial Feature Localization", *Proceedings of International Conference on European Conference on Computer Vision*, pp. 679-692, 2012.

[51] J. Yang, J.K. Deng, K.H. Zhang and Q.S. Liu, "Facial Shape Tracking Via Spatio-Temporal Cascade Shape Regression", *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pp. 41-49, 2015.

[52] G.S. Chrysos, E. Antonakos, S. Zafeiriou and P. Snape. "Offline Deformable Face Tracking in Arbitrary Videos, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 32-35, 2015.

[53] J. Shen, S. Zafeiriou, G.S. Chrysos, J. Kossaifi, G. Tzimiropoulos and M. Pantic, "The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results", *Proceedings of IEEE International Conference and Workshop on Computer Vision*, pp. 11-17, 2015.

[54] G. Tzimiropoulos, "Project-Out Cascaded Regression with an Application to Face Alignment", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3659-3667, 2015.

[55] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou and M. Pantic, "300 faces In-the-wild challenge: Database and Results", *Image and Vision Computing*, Vol. 47, pp. 3-18, 2016.

[56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "A Semi-Automatic Methodology for Facial Landmark Annotation", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2111-2117, 2013.

[57] S. Bell, K. Bala, L. Zitnick and R. Girshick, "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2874-2883, 2016.

[58] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867-1874, 2014.

[59] A. Graves and J. Schmidhuber, "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks", *Proceedings of the 21st International Conference on Neutral Information Processing Systems*, pp. 545-552, 2008.

[60] W. Byeon, T.M. Breuel, F. Raue and M. Liwicki, "Scene Labeling with LSTM Recurrent Neural Networks", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 111-117, 2015.