Improving the Neural Network Algorithm for Assessing the Quality of Facial Images*

Nikita Lisin[1] [0000-0003-4943-0733], Alexander Gromov[2] [0000-0001-9818-3770],
Vadim Konushin[2] [0000-0003-3949-0548], and Anton Konushin[1,3] [0000-0002-6152-0021]

[1] Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University, Moscow, Russia
{nikita.lisin, anton.konushin}@graphics.cs.msu.ru
[2] Video Analysis Technologies LLC, Moscow, Russia
{alexander.gromov, vadim}@tevian.ru
[3] NRU Higher School of Economics, Moscow, Russia

Abstract. The paper considers the task of obtaining a quality assessment of facial images for usage in various video surveillance systems, video analytics and biometric identification. Accuracy of person recognition and classification depends on the quality of the input images. We consider an approach to obtaining single face image quality assessment using neural network model, which is trained on pairs of images that are split into two possible classes: the quality of the first image is better or worse than the quality of the second one. Two modifications of the selected baseline algorithm are proposed. A face recognition system is applied to change the loss function and image and face quality attributes are used when training the model. Experimental studies of the proposed modifications show their effectiveness. The accuracy of selecting the best and worst frame is increased by 1.3% and 1.9%, respectively.

Keywords: Computer Vision, Face Quality Assessment, Face Recognition

1 Introduction

Computer vision algorithms such as face recognition, algorithms for determining emotions, demographic characteristics and key points of a human face, are widely used in video surveillance systems, video analytics and biometric identification. The received data in these systems is a video stream, which contains a set of several frames for each person. But most algorithms are built so that they process frames independently of each other and, as has been shown in many studies, their accuracy depends on the quality of the input images [1]. Therefore, these systems use face quality assessment algorithms

*Publication is supported by RFBR grant №19-07-00844.

to select the best frame to improve system performance [2] or reject the worst frames to improve the overall accuracy [3].

The task of face quality assessment is to obtain one scalar value for the input image that reflects the overall quality and takes into account both image quality attributes (illumination, blur, noise, etc.) and face quality attributes (head pose, face occlusion, etc.). Usually, this scalar value is enclosed in the range from 0 to 100, where the values 0 and 100 correspond to the image with the worst and best quality, respectively. Algorithms for obtaining this value are trained either on pairs of images split into two possible classes (the quality of the first image is better or worse than the quality of the second one), as in [4], or using regression to obtain a specific quality value [5]. Obtaining ground truth labels in these works is carried out with the help of experts.

Many of the recent works consider the problem of face quality assessment from a different point of view: as an indicator that reflects the usefulness of the image for the specific algorithm being used. Algorithms for obtaining this value use one of the existing face recognition systems, based on which either the training and test dataset is marked up [6] or the finished model is obtained directly [7], [8]. The main idea is that the confidence of the face recognition system for a pair of images of the same person and the difference in the quality of these images are interrelated. The lower the confidence of the face recognition system that the images in a pair belong to the same person, the more they differ from each other in quality, and vice versa.

In this paper, we use the approach from article [4] to obtain an overall quality indicator. Two modifications are proposed for the baseline algorithm. The first modification is the use of face recognition system to change the loss function. Our approach differs from the previous ones in that the developed algorithm remains universal: it can be applied together with any other algorithm, and not only with the used face recognition system. The second proposed modification is to apply image and face attributes. Algorithms based on this approach use training of the neural network model for several tasks [9], when one of the tasks is quality assessment, and the remaining tasks are image and face attributes assessments. For example, in [10], the authors use sharpness, tone and colorfulness, and their experiments show that this leads to improved algorithm accuracy. At the same time, there is a small number of works that use images of human face and take into account new properties. An example of such work is [11], which uses alignment, visibility, deflection and clarity. The disadvantage is that the markup method chosen in this article is quite subjective: the ground truth labels are the mean opinion scores in the range from 0 to 1, obtained with the help of experts. We use image attributes such as illumination and blur, which are marked up for image pairs, as well as face attributes such as head rotation angles in the range from −90° to +90° and occlusion marking for 23 areas of the face into two possible classes. We assume that the considered approach to applying attributes for face quality assessment is more reliable than previously proposed.

2 Baseline algorithm

In the article [4], on which our algorithm is based, a neural network model is trained in two stages. At the first stage the neural network model is a siamese network with two identical branches with shared weights. The input is a pair of images, where second image in the pair is obtained from the first using some distortion, which degrades the image quality. The first and second images are the input of the first and second branches of the network respectively and the output is the scores Q_1 and Q_2. These values are used in Hinge Loss as follows (1):

$$\text{Hinge Loss} = \max(0, \mathrm{Q}_2 - \mathrm{Q}_1 + 1). \quad (1)$$

The minimum of this function is achieved when the quality score of the second image in the pair is less than the quality score of the first image in more than 1. Proposed approach allows to obtain a ranking model by training on pairs of images without manual markup. At the second stage, only one branch of the siamese network is used to evaluate the quality of a single face image. This branch is also fine-tuned on a separate dataset using regression.

Our algorithm trains in the same way on pairs of images, but we use pairs of different images containing people's faces. For each pair, the markup into two possible classes was obtained using experts: the quality of the first image is better or worse than the quality of the second one. The best image in a pair was considered to be the one that best matches the combination of the following properties: front projection, normal illumination, absence of occlusion, noise and blur, etc. This approach for obtaining pairs was chosen because, from our point of view, it allows us to take into account more complex cases that occur between pairs of different images, which cannot be obtained by applying distortion to the original image. Different strategies for selecting pairs for markup were used to cover more cases, and pairs for which it is not possible to uniquely define a class were not used later. We don't apply fine-tuning on a separate dataset. Resnet-10 [12], shown in Fig. 1, is used as a neural network model.

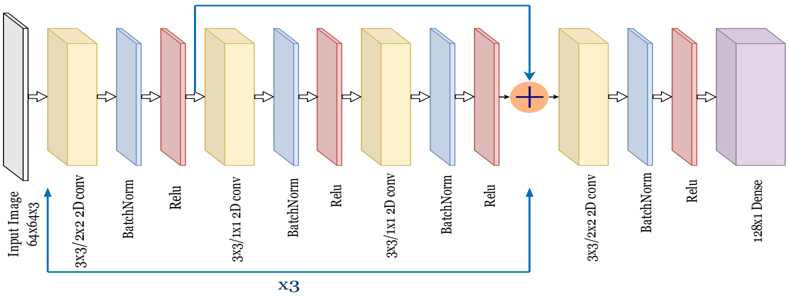

Fig. 1. Neural network architecture

3 First modification

As the first modification of the baseline algorithm, a new loss function is proposed, which uses the output of the face recognition system. In the baseline algorithm loss function for quality assessment has the following form (2):

$$\mathrm{L}_{\mathrm{FQA}} = \max(0, \mathrm{Q}_2 - \mathrm{Q}_1 + \mathrm{margin}), \tag{2}$$

where margin is a constant value equal to 1. For a part of the training dataset that consists of pairs of images of the same person, we calculate a set of probabilities using the face recognition system. A probability is in the range from 0 to 1, where the value 1 means that a pair of images belongs to the same person and the value 0 means that they belong to different people. We use this probability as an indicator of the similarity of two images in terms of quality. This set is then normalized so that the expected value would be 0 and the standard deviation would be 1. In the new loss function, margin has the following form (3):

$$\mathrm{margin} = \begin{cases} 1, & \text{a pair of images of different people} \\ \max(\alpha, 1 - \beta * \mathrm{FR}), & \text{a pair of images of a single person} \end{cases}, \tag{3}$$

where $\alpha \in (0, 1)$ – new minimum value of the margin; $\beta \in \mathbb{R}^+$ – custom parameter; $\mathrm{FR} \in \mathbb{R}$ – output of the face recognition system after normalization. In our experiments, we use the following parameter values: $\alpha = 0.4$, $\beta = 2$. Face recognition system is developed by Video Analysis Technologies [16]. Our approach differs from the previous ones in that the resulting algorithm remains universal: it can be used in the future together with any other algorithm, and not only with the face recognition system selected for training.

4 Second modification

As a second modification of the baseline algorithm, we consider multi-task learning of the neural network to evaluate the face quality and image attributes such as illumination and blur, as well as face attributes such as head pose and face occlusion. The general loss function has the following form (4):

$$\mathrm{Loss} = \mathrm{w}_1 * \mathrm{L}_{\mathrm{FQA}} + \mathrm{w}_2 * \mathrm{L}_{\mathrm{Illumination}} + \mathrm{w}_3 * \mathrm{L}_{\mathrm{Blur}} + \mathrm{w}_4 * \mathrm{L}_{\mathrm{Pose}} + \mathrm{w}_5 * \mathrm{L}_{\mathrm{Occlusion}}, \tag{4}$$

where $\{\mathrm{w}_\mathrm{i}\}_{\mathrm{i}=1}^5 -$ weight coefficients. In our experiments, we use the following parameter values: $\mathrm{w}_1 = 64,\ \mathrm{w}_2 = 48,\ \mathrm{w}_3 = 48,\ \mathrm{w}_4 = 2,\ \mathrm{w}_5 = 4$.

4.1 Illumination and Blur

Evaluation of illumination and blur occurs on pairs of images similar to the face quality assessment in the baseline algorithm, with the use of Hinge Loss as $\mathrm{L}_{\mathrm{Illumination}}$ and

$\mathrm{L}_{\mathrm{Blur}}$. During training, the value of the loss function for pairs without markup is assumed to be zero.

4.2 Head Pose

The head pose estimation consists of determining three angles for each image in a pair: pitch, roll and yaw. Each angle is enclosed in the range from −90° to +90°. Training is performed using regression and Weighted Mean Absolute Error is used as the loss function (5):

$$\mathrm{L}_{\mathrm{Pose}} = \frac{\alpha * |\mathrm{P}-\overline{\mathrm{P}}| + \beta * |\mathrm{R}-\overline{\mathrm{R}}| + \gamma * |\mathrm{Y}-\overline{\mathrm{Y}}|}{\alpha+\beta+\gamma}, \tag{5}$$

where $\overline{\mathrm{P}}, \overline{\mathrm{R}}, \overline{\mathrm{Y}} \in \mathbb{N}$ — ground truth labels; $\mathrm{P}, \mathrm{R}, \mathrm{Y} \in \mathbb{N}$ — neural network output; α, β, $\gamma \in \mathbb{R}$ — weight coefficients for pitch, roll and yaw respectively. In this work, we use the following parameter values: $\alpha = 1,\ \beta = 1,\ \gamma = 0.05$. The low value of γ is explained by the fact that the marking for yaw is less accurate than for pitch and roll. Fig. 2 shows an example of head rotation angles using three guide vectors.

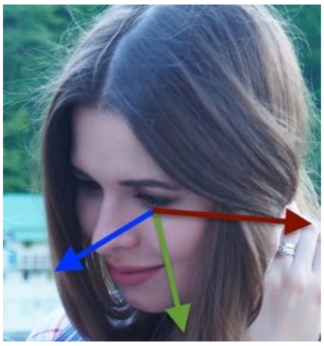

Fig. 2. Example of head rotation angles

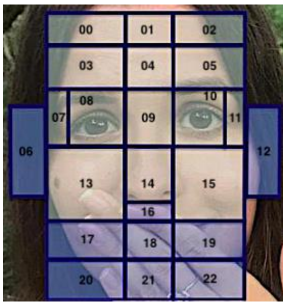

Fig. 3. An example of areas markup. Blue indicates the invisible class

4.3 Face Occlusion

Face occlusion task is to determine one of two types of visibility for 23 areas of the face: the visible region and the invisible region due to the rotation of the head, overlapping by an external object, or going beyond the boundaries of the image. The scheme of dividing the face into regions is based on the approach proposed in [13], while some regions were divided into subdomains, and new ones were added. Fig. 3 shows an example of areas markup. The loss function has the following form (6):

$$\mathrm{L_{Occlusion}} = \sum_{\mathrm{i}=1}^{23} \mathrm{CE\ Loss_i} \tag{6}$$

where $\mathrm{CE\ Loss_i}$ is the Cross Entropy Loss for the i-th region.

5 Training Dataset

Since the necessary dataset are not publicly available, we created our own training set. Images for constructing pairs were provided by Video Analysis Technologies [16]. Table 1 describes the methods of obtaining labels for quality assessment and attributes. Expert markup for face quality assessment and blur was performed with the help of five people and each pair was labeled only by one. The neural network models used for markup of illumination, head pose, and face occlusion were trained on separate datasets:

1. the dataset for training face occlusion classifier is marked up with the help of experts;
2. the dataset for training head poses classifier is marked up by determining the rotation angles based on 68 key points;
3. the dataset for training illumination classifier is marked up into 13 levels of illumination as follows:
   a. the autoencoder was trained;
   b. outputs correlating with the degree of illumination were found in the intermediate representation of the autoencoder;
   c. based on the found outputs, all data was divided into 13 classes;
   d. additional expert markup was made to remove false cases.

It should be noted that in the case of illumination, training is carried out on pairs of images, the markup for which is obtained automatically based on their illumination levels, because this approach turned out to be more stable. General characteristics of the obtained dataset, taking into account the transitive closure and the number of pairs used together with the face recognition system, are given in Table 2.

Table 1. Methods used to get ground truth labels.

Task	Type of markup	Markup method
Face Quality Assessment	Paired markup into two possible classes	Expert markup
Illumination	Paired markup into two possible classes	Neural network classifier for 13 levels of illumination
Blur	Paired markup into two possible classes	Expert markup
Head Pose	Values of three angles for an individual image	Neural network algorithm for determining rotation angles of the head
Face Occlusion	Visibility type for 23 areas of an individual image	Neural network classifier of face occlusion

Table 2. Dataset characteristics.

Characteristic	Quantity
Total number of pairs	417 240
Total number of images	334 660
Blur, pairs	124 820
Illumination, pairs	286 240
Pairs of one person	65 010
Face Quality Assessment	all pairs
Head Pose	all images
Face Occlusion	all images

6 Test Dataset and Metrics

Since the most common datasets are designed to evaluate the quality of an arbitrary image that does not necessarily contain a human face, we have created a new dataset for this purpose. The test dataset consists of tracks – sets of 5 to 12 frames containing images of faces belonging to one person. The total number of tracks is 7070, and for each track the markup of the best and worst frames was made in terms of quality assessment. The best frames are those that have the best matches the combination of the following properties: front projection, normal illumination, absence of occlusion, noise and blur, etc. Similarly, the worst frames are those that have the worst correspondence to these properties. Each track was marked by three experts, and the frame was considered the best or worst only if the opinions of at least two experts were the same. It was also required to select as few frames as possible. Fig. 4 shows an example of such a track. For this dataset, we define three metrics: Best Shot Accuracy, Worst Shot Accuracy and Pair Accuracy.

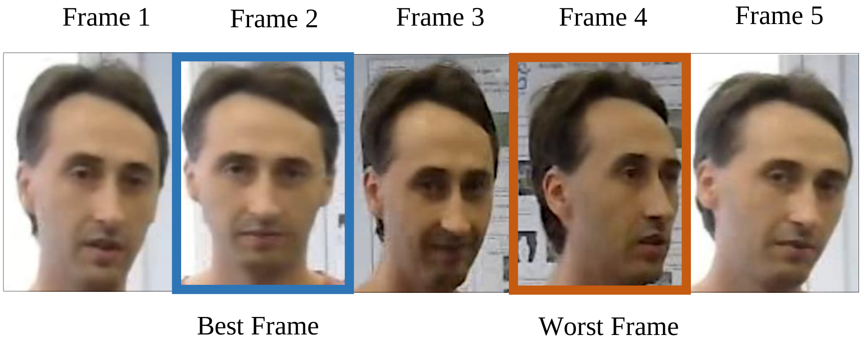

Fig. 4. Track example. Blue indicates the best frame; red indicates the worst frame

Notation for the constructed dataset:

- $S = \{S_t\}_{t=1}^{N}$ – set of N tracks;
- $M_t \in \mathbb{N},\ M_t \in [5, 12]$ – number of frames for each track;
- $S_t = \{F_k^t\}_{k=1}^{M_t}$ – track with the number t, which consists of frames F_k^t;
- $B = \{B_t\}_{t=1}^{N}$ – set of tracks with the best frames, $B_t \subset S_t$;
- $W = \{W_t\}_{t=1}^{N}$ – set of tracks with the worst frames, $W_t \subset S_t$.

Notation for the algorithm output:

- $Q = \{Q_t\}_{t=1}^{N}$ – collection of N sets with quality scores;
- $Q_t = \{QF_k^t\}_{k=1}^{M_t}$ – set of quality scores for a track with the number t.

We define two indicator functions:

$$I_t^B = \begin{cases} 1, F_m^t \in B_t \\ 0, F_m^t \notin B_t \end{cases}, m = \underset{\forall k \in [1, M_t]}{\operatorname{argmax}} QF_k^t, \tag{7}$$

$$I_t^W = \begin{cases} 1, F_m^t \in W_t \\ 0, F_m^t \notin W_t \end{cases}, m = \underset{\forall k \in [1, M_t]}{\operatorname{argmin}} QF_k^t. \tag{8}$$

The indicator function (7) corresponds to the choice of the best frame – it is equal to 1 if and only if the frame with the highest quality score is contained in the set of the best frames. Similarly, (8) corresponds to the selection of the worst frame. Based on the indicator functions, we define two metrics:

$$\text{Best Shot Accuracy} = \frac{\sum_{t=1}^{N} I_t^B}{N}, \tag{9}$$

$$\text{Worst Shot Accuracy} = \frac{\sum_{t=1}^{N} I_t^W}{N}. \tag{10}$$

Since each track consists of frames of three types (best, worst and normal), we can also create image pairs for each track that consist of two different types of frames. The resulting pairs are marked up into two classes (the quality of the first image is better or worse than the quality of the second one), which is uniquely determined based on the types of images in the pair. Pair Accuracy is defined as the percentage of correctly classified specified pairs, the total number of which is 173 940.

7 Experiments

On the test datasets four algorithms are compared: Baseline, Baseline with Modified Loss, Baseline with Attributes, Baseline with Modified Loss and Attributes. On Fig. 5 the scheme of the baseline algorithm along with two modifications is given.

During training a polynomial change of the learning rate with the degree of polynomial 2 is used, and the initial value of the learning rate is 0.001. The total number of epochs is 25. We also use the Ranger optimizer, which is a combination of two methods proposed in [14] and [15]. Augmentation is the same transformation of both images in a pair (changing saturation, illumination, contrast, additive Gaussian noise, blurring, cropping the image, etc.). Inference latency of the baseline algorithm is 0.002 second on a single core of Intel Core CPU i5-9400.

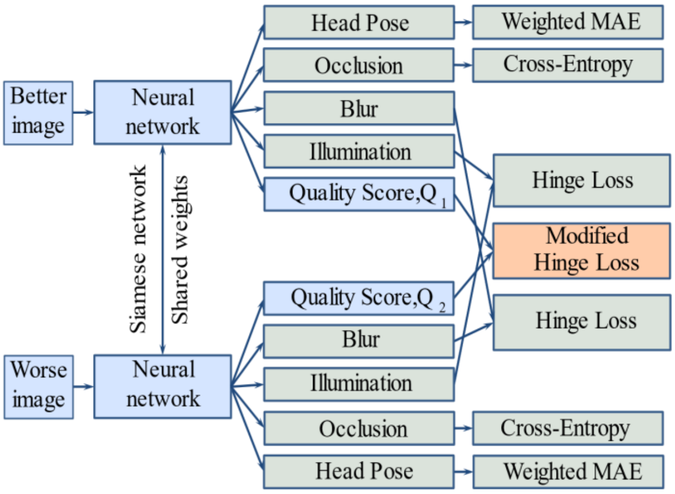

Fig. 5. Scheme of the resulting algorithm. Blue is Baseline, green is Attributes and orange is Modified Loss

The results obtained on the test dataset are shown in Table 3. Experimental assessment shows that the use of both modifications achieves the best result and increases the accuracy of selecting the best frame by 1.3%, the accuracy of selecting the worst frame by 1.9%, pair accuracy by 0.9%. The advantage of the proposed modifications is that they do not increase the inference time of the baseline algorithm.

Table 3. Experimental assessment

Algorithm	Best Shot Accuracy	Worst Shot Accuracy	Pair Accuracy
Baseline	0.710	0.703	0.855
Baseline with Modified Loss	0.714	0.712	0.858
Baseline with Attributes	0.721	0.722	0.862
Baseline with Modified Loss and Attributes	0.723	0.722	0.864

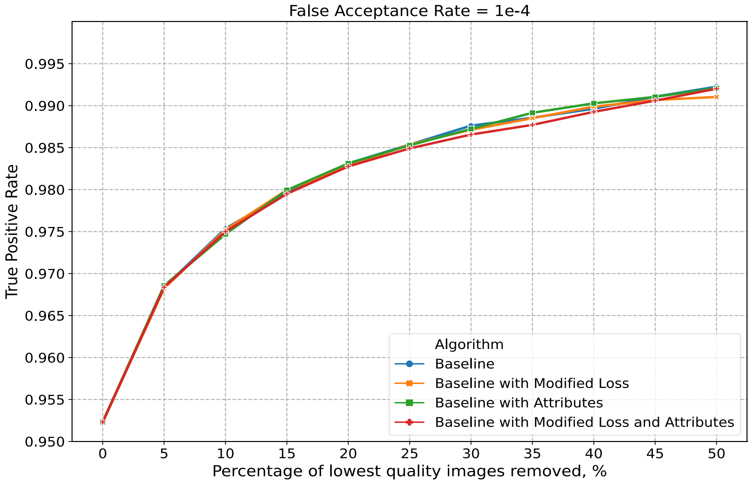

Fig. 6. Dependence of the face recognition system on the quality of input images

To study the dependence of the face recognition system on the quality of input images we use a second test dataset consisting of 61 500 pairs of images of a single person and a face recognition system from the company Video Analysis Technologies [16]. The dataset used has a limited variation in image quality and was not specially designed for this purpose, so it is insufficient to demonstrate the difference between the modifications. But we present the results as additional confirmation of the applicability of the developed algorithms. Fig. 6 shows the dependence of True Positive Rate on the percentage of lowest quality images removed, with a fixed False Acceptance Rate of 0.0001.

8 Conclusion

In this paper the modified neural network algorithm for face quality assessment is proposed. An approach from article [4] is used for baseline algorithm. As modifications the face recognition system is applied to change the loss function and image and face quality attributes are used when training the model. Changes increase the accuracy of selecting the best and worst frame by 1.3% and 1.9%, respectively, without affecting performance.

References

1. Grother, P., Hom, A., Ngan, M., Hanaoka, K.: Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Asssessment. Information Access Division Information Technology Laboratory, NIST (2020).
2. Nikitin, M., Konushin, A., Konushin, V.: Face quality assessment for face verification in video. In: Proceedings of the 24th International Conference on Computer Graphics and Vision GraphiCon'2014, pp. 111–114 (2014).
3. Bagrov, N., Konushin, A., Konushin, V.: Face recognition with low false positive error rate. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 11–15 (2019).
4. Liu, X., Van De Weijer, J., Bagdanov, A.: RankIQA: Learning from Rankings for No-Reference Image Quality Assessment. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1040-1049 (2017).
5. Best-Rowden, L., Jain, A.: Learning Face Image Quality From Human Assessments. In: IEEE Transactions on Information Forensics and Security, vol. 13, no. 12, pp. 3064-3077 (2018).
6. Hernandez-Ortega, J., Galbally, J., Fiérrez, J., Haraksim, R., Beslay, L.: FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In: 2019 International Conference on Biometrics (ICB), pp.1-8 (2019).
7. Terhorst, P., Kolf, J., Damer, N., Kirchbuchner, F., Kuijper, A.: SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness (2020).
8. Nikitin, M., Konushin, V., Konushin, A.: Neural network model for video-based face recognition with frames quality assessment. In: Computer Optics 2017, vol. 41, pp. 732-742 (2017). (In Russian)
9. Kuharenko, A., Konushin, A.: Simultaneous facial attribute classification with convolutional neural networks. In: 11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2003), IPSI RAS Samara, vol. 2, pp. 623–626 (2013).
10. Yang, D., Peltoketo, V., Kämäräinen, J.: CNN-Based Cross-Dataset No-Reference Image Quality Assessment. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), pp. 3913-3921 (2019).
11. Lijun, Z., Xiaohu, S., Fei, Y., Pingling, D., Xiang-dong, Z., Yu, S.: Multi-branch Face Quality Assessment for Face Recognition. In: 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, pp. 1659-1664 (2019).
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778 (2016).

13. Maze, Brianna et al.: IARPA Janus Benchmark - C: Face Dataset and Protocol. In: 2018 International Conference on Biometrics (ICB), pp. 158-165 (2018).
14. Zhang, M., Lucas, J., Hinton, G., Ba, J.: Lookahead Optimizer: k steps forward, 1 step back. NeurIPS, (2019).
15. Liu, Liyuan et al.: On the Variance of the Adaptive Learning Rate and Beyond, (2020).
16. Video Analysis Technologies Homepage, https://tevian.ru.