Head model reconstruction and animation method using color image with depth information

Y.K. Kozlova [1], *V.V. Myasnikov* [1,2]
[1] *Samara National Research University, 443086, Samara, Russia, Moskovskoye shosse 34;*
[2] *IPSI RAS – Branch of the FSRC “Crystallography and Photonics” RAS,*
443001, Samara, Russia, Molodogvardeyskaya 151.

Abstract

The article presents a method for reconstructing and animating a digital model of a human head from a single RGBD image, a color RGB image with depth information. An approach is proposed for optimizing the parametric FLAME model using a point cloud of a face corresponding to a single RGBD image. The results of experimental studies have shown that the proposed optimization approach makes it possible to obtain a head model with more prominent features of the original face compared to optimization approaches using RGB images or the same approaches generalized to RGBD images.

Keywords: 3D reconstruction, 3D animation, virtual reality, augmented reality, FLAME, RGB image, depth information, RGBD, point cloud, optimization.

Citation: Kozlova YK, Myasnikov VV. Head model reconstruction and animation method using color image with depth information. Computer Optics 2024; 48(1): 118-122. DOI: 10.18287/2412-6179-CO-1334.

Introduction

3D reconstruction [1] of a digital model of a human head is a rather complex and topical problem in computer vision. The relevance of this problem in recent years has greatly increased, associated with the introduction of augmented and virtual reality technologies into everyday life. In practice, only a reconstructed digital head model is less interesting than, for example, its subsequent animation. Such application of technology allows for solving the problem of virtual human telepresence [2], [3]. Also, this technology can be used to compress information. This application is relevant to avoid the loss of information we transmit over a communication channel with variable bandwidth, as described in section 3. Thus, the task involves constructing an accurate virtual model of the human head using data from a camera and animating it using a sequence of RGB frames. In our research, a stereo camera was employed to build the human head model, which, in addition to capturing the RGB image, provides depth information for creating a 3D point cloud.

Head reconstruction methods can be classified into parametric [4], [5], non-parametric [6], [7] and mixed [8]. The FLAME parametric model was chosen as the basis for head reconstruction in experimental studies [5]. The primary advantage of parametric methods is their ability to achieve high surface extrapolation accuracy, even in regions where data may be sparse or unavailable. However, parametric methods also have some drawbacks. For instance, they do not account for hair reconstruction and may not capture fine facial details or accessories related to the head and face surface. Animation of the head model using parametric models is implemented by introducing the stage of optimizing the parameters responsible for the position of the head and facial expression of the subject tracked in the frame. On the other hand, non-parametric methods do not rely on a predefined model but instead directly reconstruct the head structure from the available data. These methods can overcome the limitations of parametric approaches, as they do not impose rigid constraints on the model. However, they may be more noise-sensitive and require more extensive data for accurate reconstruction. Animation for non-parametric models can be achieved by dynamically updating reconstructed 3D data using a sequence of temporal frames, allowing for realistic facial expressions and movements. Another approach involves using surface deformation methods to alter the facial shape between frames smoothly, enabling smooth animation without relying on predefined parameters. Mixed methods combine elements of both parametric and non-parametric approaches, aiming to leverage the strengths of each. By incorporating a parametric model as a base and refining it with non-parametric data, mixed methods attempt to achieve more accurate and detailed reconstructions.

The structure of the work is as follows. Section 1 describes a method for reconstructing a human head from a single RGBD image/frame (a frame that combines an RGB color image and depth information of the scene being captured) and then animating it from an RGB video. Section 2 presents the procedure for conducting experimental studies and the results obtained. Section 3 presents examples of possible application of the proposed method. The work ends with a conclusion, in which decisions are formulated based on the results obtained during experimental studies.

1. Description of the method of reconstruction and animation of the head model

Figure 1 shows a diagram of the method of reconstruction and animation of the head model using single RGBD image. The method consists of four main stages:

1. Preprocessing of input data;
2. Reconstruction of the human head model;

3. Texturing the human head model;
4. Animation of the human head model based on an RGB video sequence.

Preprocessing of input data

The algorithm's input is an RGBD image and the corresponding point cloud. We use an RGB image to search for a face and extract two-dimensional key points [9]. Then, to obtain 3D key points of the face, the calculated two-dimensional key points of the face and the generated point cloud are compared. The preprocessing procedure is completed by filtering the initial point cloud. Filtering consists of cutting off such points not included in the range of values formed from the values of 3D coordinates of key points. A convex hull is calculated for the 3D coordinates of the key points of a person's face. Then all points of the original point cloud are filtered according to the criterion of their occurrence in the resulting convex hull. The result of this stage is a set of 2D face key points, 3D face key points, and a filtered point cloud.

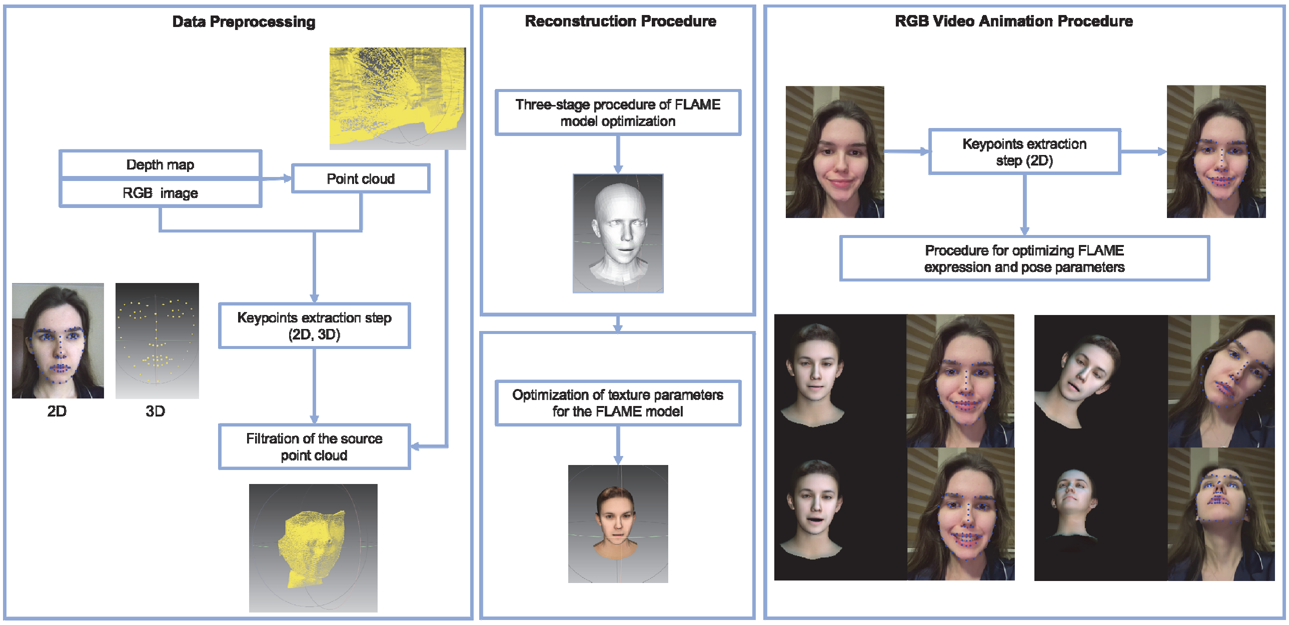
Data Preprocessing
Depth map
RGB image
Point cloud
Keypoints extraction step (2D, 3D)
2D
3D
Filtration of the source point cloud
Reconstruction Procedure
Three-stage procedure of FLAME model optimization
Optimization of texture parameters for the FLAME model
RGB Video Animation Procedure
Keypoints extraction step (2D)
Procedure for optimizing FLAME expression and pose parameters

Fig. 1. Diagram of the method of reconstruction and animation of the head model using single RGBD image

Stage of the human head model reconstruction

At this stage, we are optimizing the FLAME parametric model of the human head. The input is the 3D key points of the face and the filtered point cloud, which we received at the previous stage of the method. The procedure for optimizing the parameters describing the head model is iterative and includes three optimization blocks.

Before describing the optimization blocks, we introduce some notations. Let the FLAME parametric model be denoted as a function $M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$ that returns a mesh, where $\vec{\beta}$ is the human head shape parameters, $\vec{\theta}$ is the human head posture parameters, $\vec{\psi}$ is the human face expression parameters. FLAME uses Linear Blend Skinning with 5023 vertices and four joints. The four joints represent the eyes, jaw and neck. The parameters of the posture of the human head describe the positions and rotations of these joints. Based on the model parameters $\vec{\beta},\vec{\theta},\vec{\psi}$, one can directly obtain the coordinates of 3D key points and the point cloud containing the vertices of the resulting 3D model. We will denote the function for obtaining 3D key points as $kpts_{3D}\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$, the function for obtaining 2D key points as $kpts_{2D}\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$, and the function for obtaining a point cloud as $PCD\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$. We perform an orthogonal projection onto the $\vec{\eta}$ plane for all points from FLAME so that the points obtained from FLAME correspond to the points obtained in the previous step of the method. Such a transformation will be denoted as a function $OP\left(x,\vec{\eta}\right)$. This transformation corresponds with some accuracy to the transformation the camera applies to the physical world. Wh some accuracy, for the reason that in this problem, we neglect the perspective projection. The position of the plane $\vec{\eta}$ for the orthogonal projection will also be subject to optimization.

By optimization, we mean changing the parameters $\vec{\beta},\vec{\theta},\vec{\psi}$ or $\vec{\eta}$ in order to minimize the criterion that describes the degree of deviation of the predicted data obtained from the reconstructed model from the true data extracted from the human face on the RGBD frame. Further, the criterion will be referred to, among other things, as a loss function. The first two optimization blocks use the L_2 measure or the Euclidean distance as a component of the loss function. The value L_2 of the measure in the general case can be calculated by the formula (1):

$$L_2\left(x_{GT},x\right)=\sqrt{\frac{1}{N}\sum_{i=0}^{N-1}\left(x_{GT_i}-x_i\right)^2}, \qquad (1)$$

where N – the number of key points or the number of points in the target point cloud, xGTi – target coordinates of the i-th key point or target coordinates of the i-th point from the point cloud x_i – predicted coordinates of *i*-th key point or *i*-th closest vertex of $PCD\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$. We also introduce a loss function, which we will use to estimate the proximity measure of points in the target point cloud to the mesh of the model being optimized $M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)$. To the greatest extent, this loss function allows us to reflect the small features of the human face in the resulting model. The value of this loss function will be calculated as the averaged sum of the squared distances from points in the target point cloud to the nearest triangular faces in the mesh using formula (2):

$$L_{3D}\left(M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),PCD_{GT}\right)= \\ =\frac{1}{K}\sum_{i=0}^{K}dist_{point\_to\_mesh}\left(M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),PCD_{GT_i}\right)^2, \quad (2)$$

where K – the number of points in the target point cloud, PCD_{GT_i} – target coordinates of the i-th point from the point cloud PCD_{GT} (ground truth or target sample), $dist_{point\_to\_mesh}(x, y)$ – function that returns the Euclidean distance from point y to the nearest face of mesh x. The first block is responsible for optimizing the parameters $\vec{\eta}$ and $\vec{\theta}$ relative to the input RGBD frame based on the 3D coordinates of the key points of the face (not including the key points that describe the location of the face oval). This stage allows us to perform mutually correct positioning of the model by the orthogonal projection and the person's face both on the RGB frame and in space. We calculate the value of the loss function using formula (3):

$$Loss_1 = L_2\left(OP\left(kpts_{3D}\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),\vec{\eta}\right),kpts_{GT_{3D}}\right)\to\min_{\vec{\eta},\vec{\theta}}. \quad (3)$$

The second block is responsible for optimizing the parameters $\vec{\beta}$ and $\vec{\psi}$ based on the 3D coordinates of the key points of the face and the filtered point cloud. It is important to clarify that optimization is also performed for the parameters $\vec{\eta}$ and $\vec{\theta}$. This is necessary for the best accuracy, as there may be a situation in which the shape of the head will be significantly different from the initialized shape of the FLAME model (for example, the head is more elongated relative to the original model). In such a situation, at the first stage of optimization, the parameters $\vec{\eta}$ only approximately allow reaching the truly key points, which is characterized by a significant value of the loss function. The final loss function for the second optimization block is calculated by the formula (4):

$$Loss_2 = L_2\left(OP\left(kpts_{3D}\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),\vec{\eta}\right),kpts_{GT_{3D}}\right)+ \\ +L_{3D}\left(M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),PCD_{GT}\right)\to\min_{\vec{\beta},\vec{\theta},\vec{\psi},\vec{\eta}}. \quad (4)$$

Finally, we need the third block to refine the shape of the head (face) by optimizing the $\vec{\beta}$ parameters based on the filtered point cloud. The loss function for a given block is calculated by the formula (5), which we define as the average sum of the squared distances from points in the target point cloud to the nearest triangular faces in the mesh:

$$Loss_3 = L_{3D}\left(M\left(\vec{\beta},\vec{\theta},\vec{\psi}\right),PCD_{GT}\right)\to\min_{\vec{\beta}}. \quad (5)$$

Stage of the human head model texturing

The original FLAME model does not have a model describing texture information. Therefore, in this work, the Basel Face Model (BFM) [10] is used for optimization, which, after optimization, returns a FLAME-compatible UV map. The lighting and texture parameters are optimized at the current stage, denoted as $\vec{\lambda}$ and $\vec{\rho}$, respectively. We pass the optimized parameters to the input of the rendering algorithm. The result of rendering is an image of an orthogonal projection of the model with a superimposed texture and optimized parameters $\vec{\eta}$ and $\vec{\theta}$ under given lighting conditions. The role of the loss function, which we calculate by the formula (6), is the average value of the absolute pixel-by-pixel difference between the original RGB image and the rendered model with the texture applied, optimized by the parameters $\vec{\eta}$ and $\vec{\theta}$ so that the rendered head image occupies a position similar to the position heads in the frame. This loss function, in essence, shows how similar the rendered human head model is to the original RGB image. The result of the stage is a UV scan, which we impose on the mesh.

$$Loss_4 = \frac{1}{I\cdot J}\sum_{i=0}^{I-1}\sum_{j=0}^{J-1}\left|img_{GT_{i,j}} - img_{i,j}\right|\to\min_{\vec{\lambda},\vec{\rho}}. \quad (6)$$

Stage of the head model animation

The stage of the animation procedure receives a reconstructed head model with a UV scan and an RGB video sequence as input. For each video sequence frame, two-dimensional face key points are extracted using the previously mentioned key point extraction algorithm, which then participates in the optimization procedure for the model parameters $\vec{\theta}$ and $\vec{\psi}$. The role of the loss function, the value of which we calculate by formula (7), is the L_2-measure by the two-dimensional coordinates of the key points of the RGB frame of the video sequence and the two-dimensional coordinates of the corresponding points in the orthogonal projection of the mesh of the head model:

$$Loss_5 = L_2\left(kpts_{GT_{2D}},kpts_{2D}\left(\vec{\beta},\vec{\theta},\vec{\psi}\right)\right)\to\min_{\vec{\theta},\vec{\psi}}. \quad (7)$$

Thus, the key feature of the proposed approach to head reconstruction, in contrast to the basic FLAME approach, is using a 3D face point cloud as additional refinement information for optimizing the parametric model. To introduce a 3D point cloud into the optimization procedure, the approach to optimizing the FLAME parametric model was modified, namely, the loss function was modified at the second stage of optimization, and a third stage was added, in which the facial features of the model are refined based on the point cloud of the face.

2. Experimental studies

A dataset was collected from several people to conduct experimental studies using the ZED2 stereo camera. Each person has an RGB image, a depth map, and a point cloud.

Due to the specifics of the data when solving the problem of reconstructing the head model, we can evaluate the quality of the method only visually. We compared the proposed method with the original head model reconstruction method proposed in [5] and the DECA reconstruction method [4] based on FLAME optimization, supplemented by the stage of refining facial irregularities using a trained ResNet50 architecture encoder [11] with an added fully connected layer to translate an RGB image into latent space.

In the original implementation of DECA for reconstruction, the authors use information about the 3D coordinates of the key points of the face, which they obtained using the method proposed in [12]. This method takes an RGB image as input, so we expect the output 3D coordinates to only approximate the spatial coordinates for key face points. For a more meaningful comparison, we pass the target 3D coordinates of the key points to DECA.

Figure 2 shows the results of a comparison of methods for reconstructing a human head model.

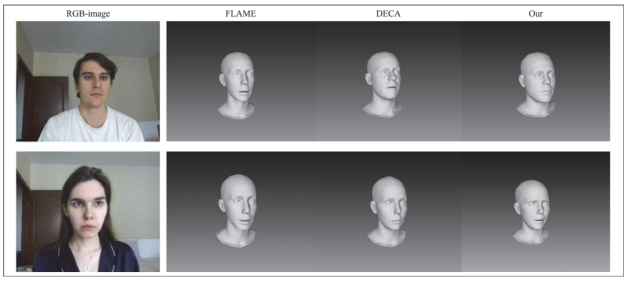

Fig. 2. Comparison of methods for reconstructing the human head model

From the obtained results, we can conclude that the classical approach to optimizing the FLAME model based on the two-dimensional coordinates of key points allows only the plane to obtain some similarity between the model and a person. As in DECA, the introduction of an additional encoder-based refinement step has allowed facial features to be refined. However, the result is still only an approximation. The approach to optimization proposed by us made it possible to obtain the result of the reconstruction closest to reality. The average error values for data from the entire data set at each optimization stage of the head model reconstruction procedure are presented in Table 1. It is important to note that the increase in the error value at the final optimization stage is due to a different nature of the data entering the optimization criterion.

Tab. 1. Average values of loss functions at the stage of reconstruction of the human head over the entire dataset

Optimization stage	Loss function value
First optimization block	0.049
Second optimization block	0.019
Third optimization block	0.0017

After the reconstruction stage, we performed texturing of the resulting models. We present the texturing result for one of the models in Figure 3. The average value of the loss function over the entire dataset at the current stage was 0.038.

The final stage of the method is the animation of the reconstructed head model based on the RGB video sequence. It is important to note that the input video sequence can contain articulation for any person. That is, the reconstructed model has no rigid dependence on the person corresponding to it. First, the extraction of key points for each frame is performed, after which the parameters of the reconstructed head model are optimized, as described in the corresponding subsection devoted to the description of the proposed method. Figure 4 shows an example of the current stage for arbitrary video sequence frames. During optimization at the current stage, the maximum error value for a frame was 0.04. This value was chosen experimentally and is considered the most appropriate for the correct display of articulation. Choosing this value allows us to find a compromise between speed and quality of work. We set it as a threshold value in the optimization cycle. That is, the optimization of the facial expression for the final mesh per frame is performed until the value of the error function becomes less than 0.04.

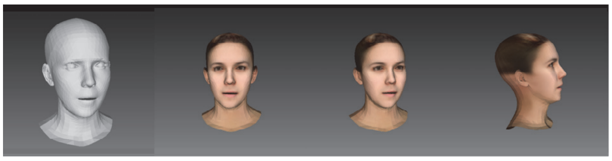

Fig. 3. Texturing result of the reconstructed human head model

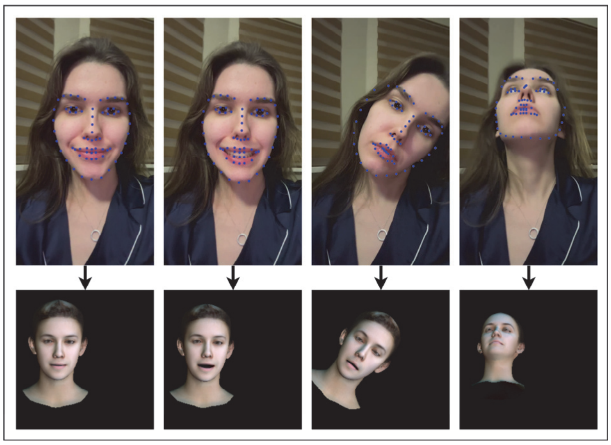

Fig. 4. An example of how the animation stage works on arbitrary frames of the input video sequence

3. Possible application of the method

The method proposed in the paper can be used for teleconferencing in a communication channel with a small bandwidth. In this case, we perform some information compression. In the first stage, to establish a connection with the interlocutor, we transmit information about the parameters $\vec{\beta}$, $\vec{\lambda}$ and $\vec{\rho}$, pre-calculated on our device. We also assume that the FLAME model is installed and configured on the interlocutor's side to the initial parameters. The information that we initially pass to initialize the model ($\vec{\beta}$, $\vec{\lambda}$ and $\vec{\rho}$), takes up 1508 bytes:

$$\left(N+M+K\right)4,\ \vec{\beta}\in\mathbb{R}^N,\vec{\rho}\in\mathbb{R}^M\ \vec{\lambda}\in\mathbb{R}^K,$$

where $N=300$, $M=50$, $K=27$, and 4 represents the size in bytes of a single real number.

After presetting, for each frame, we will have to transfer a set of parameters $\vec{\theta},\vec{\psi}$, $\vec{\eta}$. This will be 435 bytes:

$$\left(L+P+T\right)4,\ \vec{\theta}\in\mathbb{R}^L,\vec{\psi}\in\mathbb{R}^P\ \vec{\eta}\in\mathbb{R}^T,$$

where $L=6$, $\mathrm{P}=100$, $\mathrm{T}=3$, and 4 again represents the size in bytes of a single real number.

If, however, an RGB frame with a minimum comfortable resolution of 256×256 acts as information for transmission over a communication channel, then we need 196608 bytes to transmit it (in a situation where the frame is not compressed). Thus, when using the proposed approach, we can reduce the amount of the minimum required channel bandwidth by approximately 450 times.

The disadvantage of this approach to using animation is that to obtain a three-dimensional model of a human head, an RGBD frame is needed, which can only be obtained using specific equipment - a stereo camera. Also, such technology can be dangerous because it is not known who is hiding behind the speaker's face. This danger arises because attackers could use animation, a model previously built for another person.

Conclusion

The contribution of this work lies in the proposed method for reconstructing a three-dimensional model of the human head. The proposed method differs from the existing ones by a new procedure for optimizing the parameters of the FLAME model. This approach allows us to get models with more human faces and many small details. It is planned to continue this work in the future. To obtain a full-fledged human telepresence technology based on RGBD images, we need to reconstruct a high-quality human head completely. Here it is possible to improve the proposed method by moving from a parametric model to a mixed one. Namely, to supplement the method with the use of separate models that add details to the reconstructed model. Such details can be accessories like glasses, jewellery, and skin imperfections, such as scars, acne, and wrinkles. Also, perform the reconstruction of human hair, for which it is also necessary to create a model designed to solve this problem. Then, all the details obtained from different models must be combined. The result of such a combination will be a high-quality and believable model of a person's head or avatar. A separate area for research is obtaining a realistic texture for a reconstructed head model, which, for example, can be represented as a UV map. It is also important to note that when implementing the reconstruction from a single RGBD frame, it is rather challenging to capture all parts of the head. As a rule, a frontal image of a person's head is used to obtain the best results in the reconstruction of the head model. With such an arrangement of a person in the frame, capturing the most significant amount of information is realized.

Acknowledgments

The reported study was funded by the RF Ministry of Science and Higher Education within the state project of FSRC “Crystallography and Photonics” RAS (project 0026-2021-0014).

References

[1] Goshin YeV, Fursov VA. 3D scene reconstruction from stereo images with unknown extrinsic parameters. Computer Optics 2015; 39(5): 770-775. DOI: 10.18287/0134-2452-2015-39-5-770-776.

[2] Chen L, Cao C, De la Torre F, Saragih J, Xu C, Sheikh Y. High-fidelity face tracking for AR/VR via deep lighting adaptation. arXiv Preprint. 2021. Source: <https://arxiv.org/abs/2103.15876>. DOI: 10.48550/arXiv.2103.15876.

[3] Hu L, Saito S, Wei L, Nagano K, Seo J, Fursund J, Sadeghi I, Sun C, Chen, YC, Li H. Avatar digitization from a single image for real-time rendering. ACM Trans Graph 2017; 36(6): 195. DOI: 10.1145/3130800.3130887.

[4] Feng Y, Feng H, Black MJ, Bolkart T. Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans Graph 2021; 40(4): 88. DOI: 10.1145/3450626.3459936.

[5] Li T, Bolkart T, Black MJ, Li H, Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans Graph 2017; 36(6): 194. DOI: 10.1145/3130800.3130813.

[6] Dou P, Shah SK, Kakadiaris IA. End-to-end 3D face reconstruction with deep neural networks. 30th IEEE Conf on Computer Vision and Pattern Recognition 2017: 1503-1512. DOI: 10.1109/CVPR.2017.164.

[7] Jackson AS, Bulat A, Argyriou V, Tzimiropoulos G. Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression. 2017 IEEE Int Conf on Computer Vision (ICCV) 2017: 1031-1039. DOI: 10.1109/ICCV.2017.117.

[8] Grassal PW, Prinzler M, Leistner T, Rother C, Nießner M, Thies J. Neural head avatars from monocular RGB videos. arXiv Preprint. 2022. Source: <https://arxiv.org/abs/2112.01554>. DOI: 10.48550/arXiv.2112.01554.

[9] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. 2014 IEEE Conf on Computer Vision and Pattern Recognition 2014: 1867-1874. DOI: 10.1109/CVPR.2014.241.

[10] Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T. A 3D face model for pose and illumination invariant face recognition. 2009 Sixth IEEE Int Conf on Advanced Video and Signal Based Surveillance 2009: 296-301. DOI: 10.1109/AVSS.2009.58.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[12] Bulat A, Tzimiropoulos G, Kingdom U. How far are we from solving the 2D & 3D Face Alignment problem? 2017 IEEE Int Conf on Computer Vision (ICCV) 2017: 1021-1030. DOI: 10.1109/ICCV.2017.116.

Authors' information

Yuliya Khanifovna Kozlova (b. 1999) graduated with honors from Samara National Research University (Samara University) majoring in Information Security of Computer-Aided Systems in 2021. Main research interests: image processing, neural networks, computer vision, pattern recognition. E-mail: jganeeva99@gmail.com

Vladislav Valerievich Myasnikov, received his DrSc degree in Physics & Maths (2008). Currently he works as a professor at the Geoinformatics and Information Security department in Samara National Research University and, at the same time, as a leading researcher at the IPSI RAS, a Branch of the Russian Academy of Sciences 'Crystallography and Photonics' RAS. The range of scientific interests: computer vision, pattern recognition and artificial intelligence, machine learning and geoinformatics. He has about 200 publications, including more than 100 articles and three monographs. http://www.ssau.ru/staff/62061001-Myasnikov-Vladislav-Valerevich. E-mail: vmyas@geosamara.ru

Received May 05, 2023. The final version – August 13, 2023.