Verifying Kinship from RGB-D Face Data

Felipe Crispim(✉), Tiago Vieira, and Bruno Lima

Institute of Computing, Federal University of Alagoas, Maceio, AL, Brazil
fcc@ic.ufal.br

Abstract. We present a kinship verification (KV) approach based on Deep Learning applied to RGB-D facial data. To work around the lack of an adequate 3D face database with kinship annotations, we provide an online platform where participants upload videos containing faces of theirs and of their relatives. These videos are captured with ordinary smartphone cameras. We process them to reconstruct recorded faces in tridimensional space, generating a normalized dataset which we call Kin3D. We also combine depth information from the normalized 3D reconstructions with 2D images, composing a set of RGBD data. Following approaches from related works, images are organized into four categories according to their respective type of kinship. For the classification, we use a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM) for comparison. The CNN was tested both on a widely used 2D Kinship Verification database (KinFaceW-I and II) and on our Kin3D for comparison with related works. Results indicate that adding depth information improves the model's performance, increasing the classification accuracy up to 90%. To the extent of our knowledge, this is the first database containing depth information for Kinship Verification. We provide a baseline performance to stimulate further evaluations from the research community.

Keywords: Kinship Verification · Face biometrics · Structure from motion · 3D reconstruction

1 Introduction

In their daily lives, people receive information and interact in a three-dimensional world. Despite the high complexity and cost of its shapes, nowadays, researchers in computer vision and similar fields are increasingly improving techniques and extracting the benefits of this additional dimension. Even applications to unlock smartphones by 2D facial recognition can be deceived (commonly known as face recognition spoofing mechanisms). Using 3D information improves system's robustness as it is less dependent on environment interference such as illumination. Indeed, Apple has added active infrared illumination for face recognition so users can unlock their mobile devices using its FaceID technology. Regardless, a son was able to unlock his mother's phone due to their facial resemblance.

J. Blanc-Talon et al. (Eds.): ACIVS 2020, LNCS 12002, pp. 215–226, 2020.
https://doi.org/10.1007/978-3-030-40605-9_19

(a) (b) (c) (d) (e)

Fig. 1. Challenges related to facial Kinship Verification. People in (a) and (b) are twins (sister and brother, respectively). Individuals in images (b) and (c) are father and son, respectively. On the other hand, unrelated people can present similar faces as can be seen in the resemblance between actors Chad Smith (d) and Will Ferrell (e).

Kinship Verification (KV) from face images is a challenging task (*cf.* Fig. 1) due to many difficulties. It must handle similarities between non-relatives (which might reduce inter-class similarity distances) while dealing with relatives with different appearances (which could increase within-class distance). There are also variations in gender, age, ethnicity, and half-sibling categories. As a consequence, verifying kinship from faces is a broadly open research topic in computer vision and biometrics.

To the extent of our knowledge, works associated with KV are limited to two-dimensional information, either through photos or videos. There is, however, a degradation of these methodologies due to variations in lighting, expression, pose and registration [16]. In this work, we present a Kinship Verification (KV) approach by considering not only 2D images, but also depth information from faces.

In order to overcome the lack of a suitable face image database containing both depth information and kinship annotations, we provide an online platform where individuals can upload videos from their faces and their kin. From the face videos, we compute the 3D reconstruction and obtain the depth for each individual's face. RGBD images from each pair of faces are then fed into two classifiers: a Support Vector Machine, to provide a baseline, and a Convolutional Neural Network (CNN). The goal is to analyze whether depth information can contribute to the Kinship Verification problem. Moreover, we test our CNN classifier on widely used kinship database benchmarks, namely, KinFaceW-I and KinFaceW-II. When only 2D face images are considered, results are consistent with current state-of-the-art. When depth information is taken into account, results experiment an improvement of 5 (five) percentage points in accuracy.

Our contribution is two-fold. Firstly, we collected the first database with 3D information and kinship annotations. Secondly, we tested both a traditional and a contemporaneous classification techniques to provide performance accuracies to serve as baseline for future approaches. Additionally, we verify that using 3D facial information improves the model's performance, suggesting that this is an interesting path to pursue in future applications.

This paper is organized as follows. Section 2 presents an overview of works associated with Kinship Verification (KV) published so far. We provide information on methods of collecting, pre-processing, normalizing and classifying face images in Sect. 3. Results are reported in Sect. 4 and we draw final remarks and suggest future work in Sect. 5.

2 Related Work

Face verification and recognition have been extensively tackled by the scientific community and there are many commercial applications available [2,8]. Currently, new problems are increasingly more popular, such as age estimation [23] and expression recognition [12]. Among emerging areas of facial analysis, Kinship Verification (KV) is a recent topic in biometrics, provided that the first work addressing this problem was published by Fang et al. in 2010 [7]. Interest has grown due to possible applications in: (i) Forensic science, such as mitigation of human trafficking, disappearance of children and refugee crises; and (ii) Automatic annotation and reduction of the search space in large databases. Research interest also comes from the fact that KV is a challenging and open problem.

Indeed, faces carry many information about a person and they have been quite used in tasks of recognition. Somanath and Kambhamettu [20] explain how faces can verify blood-relations. As the faces develop, their features become sharper and more alike to their relatives. Parts of faces such as the eyes, nose, ears, cheekbones, and jaw, can supply helpful attributes as relative position, shape, size and color to kinship verification.

As Dibeklioglu explains [6], there are basically two types of kinship analysis: verification and recognition. The former consists of binary identification, which means identifying whether two people are related or not. The latter differentiates the type of kinship between two people (siblings or father-son).

Conventional KV methods use engineered (hand-crafted) features to be extracted from faces for further combination and classification such as Local Binary Patterns (LBP), Gabor features and others [3,22]. We highlight that, to the extent of our knowledge, no work presented thus far has used depth images to tackle the KV problem. As opposed to using hand-crafted features, Deep Learning (DL) techniques for computer vision such as Convolutional Neural Networks (CNNs) are widely used sources for obtaining additional, hierarchical representations [9].

Robinson et al. presented in 2018 the largest kinship database with 1000 family trees [15]. They used a semi-supervised labelling process to improve a pre-annotated clustering. Although the amount of images is large and, therefore, well suited for Deep Learning approaches, they provide 2D face images only, with no depth information whatsoever.

Ozkan et al. have proposed the use of Adversarial Generative Networks to synthesize faces of children by analyzing their parents [14]. One possible application is the increment in database size by generating new samples. The approach is applied to 2D images only.

3 Methodology

In this section we describe the process of collecting, preprocessing and classifying our dataset entitled Kin3D (Sect. 3.1). Then we explain the classification methods we tested, namely CNN and SVM, as well as the metrics we compute to assess model's performances.

3.1 Kin3D Database Collection

Our goal to build our Kin3D dataset is threefold: (1) to prepare, as much as possible, a reduced noise facial dataset with depth information; (2) to include labeled relatives in a database; (3) to classify kinship using known machine learning techniques as deep neural network. As a result, we have reconstructed pairs of relatives and use both RGB and depth images (D) to conduct our validation experiment.

In order to construct the Kin3D dataset, we faced two challenges. Firstly, face scanners are neither cheap nor much portable to be transported into people's homes as an attempt to scan relatives. Besides, *smartphones* are showing increasingly higher spatial and exposure resolutions, which results in 3D reconstructions with greater quality. Secondly, people might not be willing to attend an event to have their faces scanned if there is no type of compensation.

In order to facilitate collecting videos from participants, the task of recording their faces was executed with their own smartphone cameras, in uncontrolled environments. Besides, university students who willingly engaged in this activity had extra scores taken into account.

To enable this research, an online form was created in order to collect annotated data from users. Through the form, one can send a video of his face and another one for each of his relatives. Each video must comprise the lower, median and upper regions of the face, with three different slopes as shown in first stage (upper-left) of the pipeline in Fig. 2. Thus, details that would be occluded on a frontal static image are also captured. The only suggestions were smooth camera movement and a reasonably bright recording environment.

3.2 Structure-from-Motion Pipeline and 3D Reconstruction

From each video, a reduced number of frames is extracted so that they cover all movement. Structure-from-Motion technique takes these frames as input to reconstruct the face as a 3D mesh.

Prior to the 3D reconstruction, we perform a 2D normalization using frontal face features such as the eyes. Frames that present face as frontal as possible are used to compose our RGB dataset, being aligned and cropped into a image of 64×64 pixels as follows:

- Face region of interest and its landmarks detection in the RGB image;
- From eyes' landmarks the image is translated, rotated and scaled so that the eyes lie on a horizontal line;

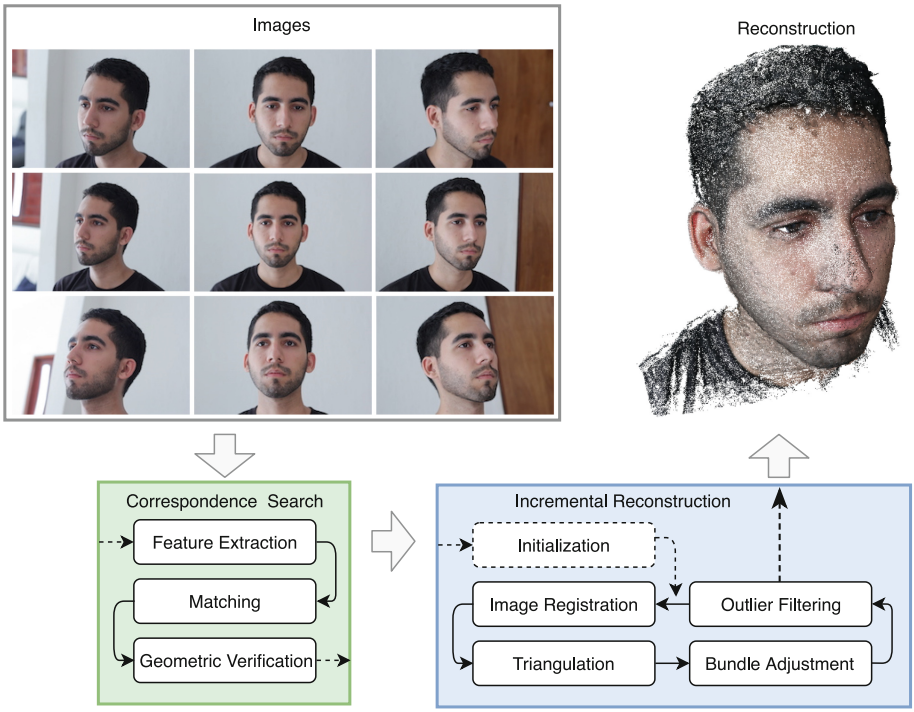


Fig. 2. Structure-from-Motion (SFM) pipeline. A set of images (upper-left) serves as input for feature detection and matching (lower, green box). After geometric verification, the 3D reconstruction happens (blue box) resulting in a set of (x, y, z) and (R, G, B) values, as shown in the upper-right corner of the figure. (Color figure online)

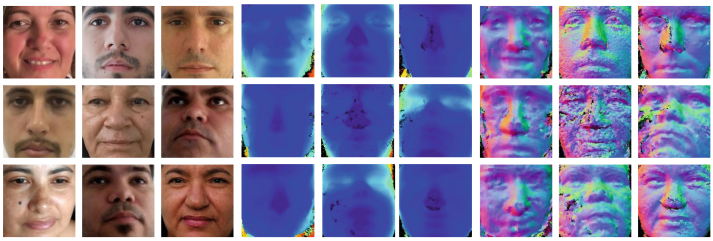
– The matrix used as affine transformations in RGB images is also used to normalize the normal and depth images, as shown in Fig. 3.

The main process of 3D reconstruction is illustrated in Fig. 2. The input is a set of overlapping facial images of the same person. That process of reconstructing is divided into three parts [17,18]:

1. Feature detection and extraction;
2. Feature matching and geometric verification;
3. Structure and motion reconstruction.

In the reconstruction process, pixels from each subsequent frame are matched while the vertices of the 3D mesh are generated. After the mesh is complete, a depth and a normal map are extracted from frontal view, as shown in Fig. 3.

At the end of this process, most of the resulting point cloud show a reasonable amount of noise due to the uncontrolled environment: illumination, camera quality and distance between the recorded person and the camera. Using the software MeshLab [5], we perform a cleaning process by removing the farthest noise just



(a) (b) (c)

Fig. 3. Examples of picture channels obtained from post processing of SFM. (a) Normalized 2D face images. (b) Depth maps computed from the 3D reconstruction. (c) Map of normal vectors represented as one RGB image. (Color figure online)

selecting and excluding it, and for the merged noise from face we remove them by color selection. Figure 4 shows an example of that cleaning process.

(a) (b)

Fig. 4. Point cloud with and without noise. (Color figure online)

Graphic Processing Units (GPUs) are designed for heavy workload and throughput. Their parallelism helps to speed up many operations in which they would last hours or days to be finished on some Central Process Units (CPUs). We have used a GPU to generate our point clouds and depth images from about 70 facial images extracted from a video.

3.3 Classification Models

In order to evaluate our classification models consistently, we performed the following steps:

1. We reproduced the results from [24] to assess the applicability of CNN to classify the widely used 2D kinship databases KinFaceW-I and KinFaceW-II. This first evaluation was useful since the number of samples is very small which may cause the model to overfit. We mitigated this by using data augmentation.
2. Since our database is novel, we test a well known classification method – Support Vector Machine – onto features obtained from 2D face images only. To this end, we follow the procedure proposed recently [21].
3. Finally, we apply the known CNN topology for performing KV. To this end, we apply the same model developed in the Stage 1 to classify our database (RGB and RGBD) to evaluate whether depth improves the accuracy or not.

The basic structure of Convolutional Neural Network (CNN) used in this project consists of several concatenated (in parallel) neural nets. The simplest network contains two convolutional layers which are connected to two max-pooling layers, then followed by two fully-connected layers and ending with a soft-max layer, as shown in Table 1. Input images are of size $39 \times 39 \times 6$ (39 wide, 39 high, 6 color channels).

Table 1. Topology of the Convolutional Neural Network before concatenation.

Conv1	Pool1	Con2	Pool2	FC
conv-32	max-2	conv-32	max-2	FC1-128
				FC2-128

The convolutional layers basically are parameterized by: the number of maps, the size of the maps and filter sizes. Our first convolutional layer receives 6 feature maps of size 39×39 and after, with $5 \times 5 \times 6$ filters, it generates 32 maps. These filters slide, or convolve, around the input image with a stride of 1. Since the image size is small, it's used a padding to ensure that the output has the same length as the original input. To initialize weights of the convolutional layers, we use a normal distribution with zero mean and a standard deviation of 0.05. A Rectifier Linear Unit (ReLU) function is used as activations.

In general, pooling layers are based in operations such as max and average. We have chosen max pooling layers because they show better accuracy. These layers have filters of size 2×2 applied to a stride of 2, downsampling the input image along both width and height. Therefore, every operation takes a max among 4 values.

Just like the humans who look at specific parts of the face to recognize who the person is, in our task we decompose the relative's RGB faces into ten individual parts, following the approach in [24]. Next we input each one to a neural network since it has been known that CNN can learn better than in a holistic way. After the convolutional, max-pooling and fully connected layers, we concatenate the ten networks. The same is also done with the depth images.

In this approach with RGB+Depth, there are twenty nets. The concatenation of the rgb and depth networks are also concatenated. Linked to this layer there is an output layer that is added to complete our network. This layer has one neuron per class in the classification task, totaling 2 neurons to relatives and non-relatives. A softmax activation function is applied.

We compile our model using the optimizer Adam [10], which is an improved version of the stochastic gradient descent algorithm that incorporates time and learning rate adaptive. Furthermore, binary crossentropy as the loss function was applied since our targets are in categorical format.

Given all the possible kinship classes, a one-versus-one (OvO) strategy was adopted. Thus, our CNN was trained to validate the relation between one father (or mother) and his son (or daughter). In addition to this strategy being faster, it is also more appropriate for smaller sets. We generate arbitrary sinthetic negative examples by combining people from different families.

To train and validate our deep convolutional neural network we used Keras and TensorFlow [1,4] on NVIDIA's CUDA programming framework [13].

4 Results

4.1 Database Collection

As a result of people's participation, we collected smartphone videos from 120 individuals. When organized by categories, the number of pairs are described in Table 2. Both positive and negative pairs have the same size. The negative pairs are generated from the swapping between positive pairs. We recognize that Kinship Verification is a very challenging problem and that a much larger dataset is recommended for a more representative approach. Regardless, we evaluate the CNN with care, keeping track of training and validation accuracies to prevent overfitting. We acknowledge that, in order to best represent a broad problem such as kinship, collecting a much larger dataset is advisable. That being said, we are still receiving videos through our form to be rebuilt and analyzed. Moreover, it is important to highlight that we use data augmentation, using rotations and vertical flip, multiplying the number of samples threefold.

Table 2. Number of kin pairs for each kinship category.

Category	Number of pairs	Number of samples after data augmentation
Father-Son (FS)	14	42
Father-Daughter (FD)	9	27
Mother-Son (MS)	27	81
Mother-Daughter (MD)	11	33

4.2 Testing a First Topology

KinFaceW [11] is divided into KinFaceW-I and KinFaceW-II and they are ones of the most known datasets used for Kinship evaluation. Using the concatenated CNN that we adapted from [24], we apply it through the KinFaceW. The main changes in our model was that we used only two convolutional layers and added one fully-connected layer. In relation to weight initialization, Zhang et al. used a Gaussian distribution with zero mean ($\mu = 0$) and a standard deviation of $\sigma = 0.01$. But we noticed that the standard deviation with that value did not allowed our concatenated CNN to learn. Thus, we initialized the weights with standard deviation $\sigma = 0.5$. Accuracies were approximate to those obtained in [24], as can be seen in Table 3.

Table 3. Verification accuracy (%) on KinFace dataset.

Methods	KinFaceW-I				Methods	KinFaceW-II			
	FS	FD	MS	MD		FS	FD	MS	MD
CNN-Points [24]	71.8	76.1	78.0	84.1	CNN-Points [24]	89.4	81.9	89.9	92.4
Our method	69.2	77.8	72.6	83.8	Our method	86.2	83.3	84.1	88.8

4.3 Our Dataset

Firstly, a Support Vector Machine (SVM) was developed to supply a baseline. SVMs are well suited for classification of complex tasks but with not-so-big datasets. The SVM is applied to 128-dimensional embedded face vectors as described by [19].

Two types of CNN, representative feature learning methods, were trained. The first one learned to verify kinship from only RGB images. The second one learned to verify kinship from RGB plus depth images to assess the contribution of depth information. We evaluate our dataset to the four categories in kinship research community: father-son (FS), father-daughter (FD), mother-son (MS) and mother-daughter (MD). Results are shown in Table 4 and Fig. 5.

Table 4. Verification accuracy (%) on Kin3D dataset.

Methods	Kin3D			
	FS	FD	MS	MD
SVM	69.9	56.6	66.6	73.8
CNN (RGB)	71.5	75.0	86.4	90.5
CNN (RGB-D)	76.4	76.9	88.4	94.3

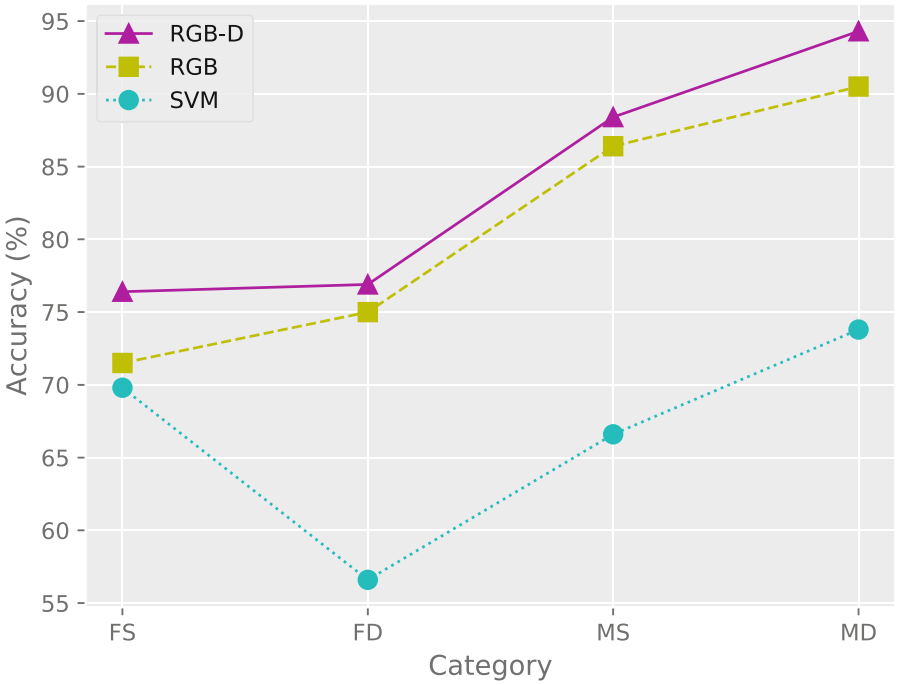


Fig. 5. Classification accuracies.

5 Conclusion

In this paper, we presented a novel use of 3D data for kinship verification (KV). Given the technological advancements in smartphone cameras, computer hardware and image processing software, the addition of 3D information to computer vision and machine learning tasks seems to be the best way to make such tasks more feasible. Our dataset, Kin3D, in addition to information provided about kin relationships also provides information such as age and ethnicity that can be used for studies related to age, synthesis of children faces based on parents, among others.

Overall, the experiments have shown that depth information contributes to the model's performance. We acknowledge that, in order to better represent a broad problem such as kinship, collecting a much larger dataset is advisable. Nevertheless, we hope this database can provide an initial contribution to the research community focused on Kinship Verification (KV) and interested in investigating further the use of 3D information to tackle this task.

Further work consists in constantly increment the database size and investigate whether each face's Point Cloud (PC) could provide further information to the models. Some questions could be tackled, such as; (i) how can neural nets performances be compared to traditional curvatures analyses; (ii) can we use

generative adversarial networks to synthesize childrens faces by analyzing their parents. Studies are being conducted and results will be eventually reported.

Acknowledgments. The authors would like to acknowledge the Msc's grant provided by FAPEAL (state's research support foundation) and Institute of Computing's students who participated in the research activities.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI 2016, pp. 265–283. USENIX Association, Berkeley (2016). http://dl.acm.org/citation.cfm?id=3026877.3026899
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997). https://doi.org/10.1109/34.598228
3. Bottino, A., Islam, I.U., Vieira, T.F.: A multi-perspective holistic approach to kinship verification in the wild. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, pp. 1–6. IEEE, May 2015. https://doi.org/10.1109/FG.2015.7284834, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7284834
4. Chollet, F., et al.: Keras (2015). https://github.com/fchollet/keras
5. Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G.: MeshLab: an open-source mesh processing tool. In: Scarano, V., Chiara, R.D., Erra, U. (eds.) Eurographics Italian Chapter Conference. The Eurographics Association (2008). https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136
6. Dibeklioglu, H.: Visual transformation aided contrastive learning for video-based kinship verification. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
7. Fang, R., Tang, K.D., Snavely, N., Chen, T.: Towards computational models of kinship verification. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 1577–1580. IEEE, September 2010. https://doi.org/10.1109/ICIP.2010.5652590, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5652590chenlab.ece.cornell.edu/projects/KinshipVerification/
8. Freire, A., Lee, K.: Face recognition in 4- to 7-year-olds: processing of configural, featural, and paraphernalia information. J. Exp. Child Psychol. **80**(4), 347–371 (2001). https://doi.org/10.1006/jecp.2001.2639. http://www.sciencedirect.com/science/article/pii/S0022096501926396
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). http://www.deeplearningbook.org
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. https://www.arxiv-vanity.com/papers/1412.6980/
11. Lu, J., Zhou, X., Tan, Y.P., Shang, Y., Zhou, J.: Neighborhood repulsed metric learning for kinship verification. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 331–345 (2014). https://doi.org/10.1109/TPAMI.2013.134
12. Nigam, S., Singh, R., Misra, A.K.: Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. Multimed. Tools Appl. **77** (2018). https://doi.org/10.1007/s11042-018-6040-3

13. NVIDIA Corporation: CUDA Programming Guide 9.0. NVIDIA Corporation (2018)
14. Ozkan, S., Orkan, A.: KinshipGAN: synthesizing of kinship faces from family photos by regularizing a deep face network. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 2142–2146 (2018). https://doi.org/10.1109/ICIP.2018.8451305
15. Robinson, J.P., Shao, M., Wu, Y., Liu, H., Gillis, T., Fu, Y.: Visual kinship recognition of families in the wild. IEEE Trans. Pattern Anal. Mach. Intell., 1 (2018). https://doi.org/10.1109/TPAMI.2018.2826549
16. Savran, A., Gur, R., Verma, R.: Automatic detection of emotion valence on faces using consumer depth cameras. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 75–82, December 2013. https://doi.org/10.1109/ICCVW.2013.17
17. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
18. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31
19. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
20. Somanath, G., Kambhamettu, C.: Can faces verify blood-relations? In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 105–112, September 2012. https://doi.org/10.1109/BTAS.2012.6374564
21. Thilaga, P.J., Khan, B.A., Jones, A.A., Kumar, N.K.: Modern face recognition with deep learning. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1947–1951, April 2018
22. Vieira, T.F., Bottino, A., Laurentini, A., De Simone, M.: Detecting siblings in image pairs. Vis. Comput. **30**(12), 1333–1345 (2014). https://doi.org/10.1007/s00371-013-0884-3
23. Xing, J., Li, K., Hu, W., Yuan, C., Ling, H.: Diagnosing deep learning models for high accuracy age estimation from a single image. Pattern Recognit. **66**, 106–116 (2017)
24. Zhang, K., Huang, Y., Song, C., Wu, H., Wang, L.: Kinship verification with deep convolutional neural networks. In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) Proceedings of the British Machine Vision Conference (BMVC), pp. 148.1–148.12. BMVA Press, September 2015. https://doi.org/10.5244/C.29.148