2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Deep Learning-based Face Recognition and the Robustness to Perspective Distortion

Naser Damer[1], Yaza Wainakh[1], Olaf Henniger[1], Christian Croll[3], Benoit Berthe[4], Andreas Braun[1], Arjan Kuijper[1 2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
[3]KIS PhotoMe Group, Echirolles, France
[4]IDEMIA, Colombes, France
Email: naser.damer@igd.fraunhofer.de

***Abstract*—Face recognition technology is spreading into a wide range of applications. This is mainly driven by social acceptance and the performance boost achieved by the deep learning-based solutions in the recent years. Perspective distortion is an understudied distortion in face recognition that causes converging verticals when imaging 3D objects depending on the distance to the object. The effect of this distortion on face recognition was previously studied for algorithms based on hand-crafted features with a clear negative effect on verification performance. Possible solutions were proposed by compensating the distortion effect on the face image level, which requires knowing the camera settings and capturing a high quality image. This work investigates the effect of perspective distortion on the performance of a deep learning-based face recognition solution. It also provides a device parameter-independent solution to decrease this effect by creating more perspective-robust face representations. This was achieved by training the deep learning model on perspective-diverse data, without increasing the size of the training data. Experiments performed on the deep model in hand and a specifically collected database concluded that the perspective distortion effects face verification performance if not considered in the training process, and that this can be improved by our proposal of creating robust face representations by properly selecting the training data.**

I. Introduction

Face recognition is one of the most socially accepted biometric characteristics as it represents the natural method that people recognize each other [1]. The usability and ever increasing accuracy of this technology is driving its implementation in a wide range of applications such as border control, identity management, forensics, and surveillance.

Varying the distance between the camera and the face causes varying levels of distortions. The main distance-dependent distortion is the perspective distortion. This distortion causes objects to look different in a captured image because of the relative scale of nearby and distant features. Perspective distortion in face images have been shown to alter the human social-judgment of these faces to be, for example, more trustworthy or attractive [2]. This distortion has also been proved to affect the human performance in face recognition tasks [3] on very low capture distances (i.e. 25cm). Therefore, it is also expected to affect the performance of automatic face recognition.

Previous works dealing with perspective distortion in face images focused on estimating the distance between the camera and the face [4]. This was performed to model and/or compensate for this distortion for applications like videochat [5]. Works that considered the degradation of face recognition performance caused by perspective distortion was limited to small distances between the face and the camera (12.7 cm to 88.6 cm) [6]. They also tried to minimize this degradation by compensating the distortion on the image level, a process that can be prone to variations in the image quality and also requires previous knowledge of face depth and distance. They tried to estimate this distance and relay on a modeled depth. However these estimations cannot be always accurate [7][8]. The perspective distortion effect might be also significant to presentation attack detection [9], especially on smart phones.

This work investigates the effect of perspective distortion on the performance of deep learning-based face recognition. It does that on a wide range of distances between the camera and the captured face (50 cm to 300 cm). This have exposed the effect of this distortion on face verification performance, e.g. the false non-match rate (FNMR) at 0.1% false match rate (FMR) was 0% when comparing reference and imposter images captured at 100 cm distance, while the FNMR rose to over 7% when comparing the same reference images to images captured at 300 cm distance in our experiment. Moreover, instead of the parameter-based and quality-sensitive compensation of the distortion on the image level, this work proposes creating face representations that are more robust to this distortion within a deep learning-based approach. This is achieved by including a wider range of distortions while training the deep network (without the need for larger data). Given the example of comparing a reference captured at 100 cm and a probe captured at 300 cm, the FNMR at 0.1% FMR was successfully reduced by 48% using the proposed approach.

II. State of the art

Images captured at different distances from a lens have two main types of distortion depending on this distance. First is the lens (geometrical) distortion that results from optical aberration within the lens. These distortions (barrel, pincushion, or mustache) leads to the deformation of straight lines in the

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

captured image. The effect of this distortion depends on the object distance to the lens as it has larger effect on the edges of the image area. This kind of distortion is usually not noticed in portraits as they rarely contain straight lines. However, lens distortion is usually minimum when using perspective cameras such as the ones typically used for face recognition [10]. Reversing these minimal effects is a well studied issue in practice[11]. The second type of distortion, which may have the main effect on the face recognition scenario, is the perspective distortion. This distortion is directly linked to the distance between the parts of the object of interest and the lens (camera) placement [10]. As with the human eye, closer objects looks larger. This causes converging verticals when imaging or looking at 3D objects, as closer parts of the objects look relatively bigger than the further parts.

The psychological effects of the perspective distortion in face images have been previously studied [12][13][2]. These studies pointed out that the perspective distortion has an effect on people's judgments on face images. Different distances from the face to the camera resulted in different judgments in terms of interactiveness, peacefulness, or trustworthiness.

As with other biometric characteristics, face recognition performance is affected by variations in the captured data. Studies have previously focused on measuring this effect, modeling it, and building algorithms that are more robust to these variations. Such variations are in face pose [14], expressions [15][16], age [17], occlusion [18], sensors [19], and illumination [20][21]. The human performance in face recognition under different levels of perspective convergence have been studied by Liu et al. [3][22]. After showing the participants a set of face images, they were asked to decide if new images were already shown in the initial set. This test proved that with a high level of perspective distortion (i.e. distance of 30cm), test participants had lower accuracy in the face recognition task.

Previous works tried to compensate for the perspective distortion by matching a 3D face template to the image within videochat applications [5]. As the perspective distortion depends on the distance from the camera and the face depth, Flores et al. [4] proposed a solution to estimate this distance by using effective perspective n-point. Velente and Soatto [6] also worked on estimating the distance between the face and the camera. Their goal was to model and learn to compensate the perspective distortion of face images. However, they did not explicitly try to estimate the necessary face depth. They have also shown the performance drop in automatic face recognition when matching images captured at different distances. They only used conventional face representations like eigenface and sparse representation coding [18]. Velente and Soatto [6] experiments included small distances between the face and the camera (12.7 cm to 88.6cm), which lower range (12.7 cm) is an unrealistic range when considering real applications in identity management, forensics, border control, or surveillance. They focused on compensating the perspective distortion on the image level, which can be sensitive to uncontrolled capture environments and requires information about the camera setup and face depth, rather than learning to extract features that are robust to perspective distortion.

Deep learning is becoming dominant in many applications that utilize machine learning. Biometrics is one of these applications where deep learning has significantly advanced the state-of-the-art accuracies. This was achieved mainly by learning to extract discriminant representations for individuals characteristics. Solutions based on deep learning were previously presented to learn representations of different characteristics such as face [23], iris [24], periocular region [25], and off-line handwritten signature [26]. In face recognition, including variations (pose, expression) in the training data have helped achieve higher recognition accuracies [27]. Such variations require larger training data, which can be hard to acquire, or data augmentations that help train accurate networks using smaller databases [27][28]. However, to our knowledge, no previous work has specifically addressed the issue of enriching the perspective distortion variation in the training data of deep learning-based face recognition solutions.

III. Methodology

Perspective distortion affects the appearance of face images depending on the distance between the face and the camera. This work studies the effect of this distortion on face biometric verification performance based on a deep learning solution. It also suggests building deep learning models that produce representations that are more robust to perspective distortion. To achieve this, a deep learning model is built and trained on data with different levels of perspective distortion. This section presents the technical structure of the studied solution.

Deep learning models: deep convolutional neural networks (CNN) have proved to be very successful in the field of face recognition [29][23]. The base architecture used in the proposed approach was modeled after the well-tested, well-performing, and relatively small OpenFace NN4.SMALL2 model [30]. OpenFace [30] is a general-purpose face recognition library based on the deep neural network (DNN) architecture suggested in [23]. However, the proposed solution is still trained from scratch, as the aim of this work is to investigate the effect of perspective distortion on deep models trained on limited data with different distortion characteristics, rather than achieving state-of-the-art face recognition performance.

The chosen CNN architecture uses the inception module [31][32][33]. It is initially inspired by the work of Lin et al. [34], and it introduces 1x1 convolutions, ignoring the layer's spatial dimensions and rather finding cross-channel correlations. The deep structure used in this work is compactly illustrated as a block diagram in Figure 1. It is worth noting, that batch normalization (BN) and Rectified Linear Unit (ReLU) activation is applied before and after every convolutional or fully-connected layer. Besides BN, there is another type of normalization applied, the local response normalization (LRN). LRN was introduced by Krizhevsky et al. [35] to further reduce error rates. The idea behind LRN is to apply, e.g. L^2 normalization, to units of adjacent feature maps at the same spatial position. Finally, the average pooling

layer makes the transition from the convolutional stage to the fully connected stage, effectively reducing the neurons of the subsequent layer to a single vector. The used architecture results in a final embedding size of 128.

Triplet loss function: under conventional classification task scenario, the most common type of loss function used when training a CNN is the cross-entropy loss. It employs the Maximum Likelihood Estimation (MLE) principle, or its computationally advantageous variant, the Negative Log-Likelihood (NLL) [36]. The reason it is suitable for classification tasks, is that the output layer of typical CNNs is a softmax function (a generalization of the logistic function). The output is thus designed to be interpreted as a probability distribution, describing the likelihood of the exclusive presence of each of the individual classes in an image. On the same note, the class labels can likewise be interpreted as probability distributions, if represented as a sparse one-hot vector, containing a single entry corresponding to the class index. Therefore, the goal of a cross-entropy loss function is to progressively reduce the Kullback-Leibler Divergence (KLD) between the predicted and the actual probability distributions.

While classification loss is still a credible solution to verification problems, it requires a large amount of training examples per class. Therefore, other types of loss functions are rather used for verification tasks. The most straightforward method is to simply input pairs of genuine or imposter example pairs to the system and reduce the supervised learning problem to a binary choice, by providing the appropriate genuine/imposter labels. In this work, a more recently proposed solution is used, namely the triplet loss [23][37]. Instead of providing genuine/imposter pairs, the idea is to provide triplets of examples. Each triplet consists of an anchor example (sample of a particular subject), a positive (genuine) example (sample of the same subject), and a negative (imposter) example (sample of a different subject). In order to establish compatibility between the examples representations/embeddings in the otherwise unstructured feature space, they are L^2-normalized and thus reside on the surface of the unit hypersphere. Ultimately, the goal is to ensure that the similarity between the anchor and the positive examples representation (e.g., measured by the squared L^2 distance) is greater than that to the negative example by at least a predefined margin (0.2 in this work).

Regularization and termination: strategies explicitly designed to reduce the generalization error, possibly at the expense of increasing the training error, are referred to as regularization techniques [36]. This work utilized the Dropout approach for regularization [38], as one of the most versatile and efficient regularization techniques, with a keep ratio of 80%. The commonly used L^2 parameter norm penalty was not used in this work. This technique has the intuitive interpretation of heavily penalizing weights of large magnitude, and thus encouraging the DNN to use all the inputs subtly, rather than some of the inputs extensively [39].

Early stopping was also used for regularization. This is utilized along the Adam optimizer [40] (used to train all models) by monitoring different termination ratios for mini-batches (450 image examples). An active triplet is one that violates the triplet loss condition. The triplet ratio, which describes the portion of genuine pairs that can build an active triplet, was set to 10^{-4}. The active triplets' ratio is also used, which is the ratio of training triplets that are active (this was set to 10^{-2}). The third termination criteria used is the triplet loss itself, terminated when reaching the threshold 10^{-4}. Different parameters have been tested and the final used parameters represent the best overall performance.

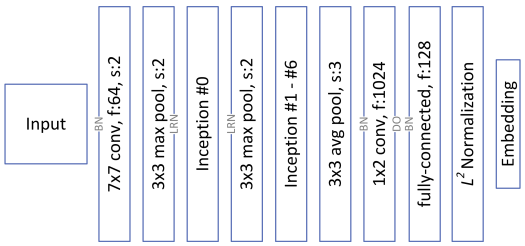

Fig. 1: The network trained and used to produce the face template.

IV. Database

The database used in this work was collected on a specifically built setup (designed by KIS S.A.S / Photo-Me Group) to measure the effect of the perspective distortion on face recognition. The capture of the images followed the specifications of the International Civil Aviation Organisation (ICAO) [8]. The data was collected from 435 users in two sessions. The first session started by capturing the images from the nearest distance (50 cm), then capturing the images at higher distances until 300 cm. The second session captured the images starting from the furthest distance down to the 50 cm. Images at 10 different distances between the camera and the face were captured (50, 60, 70, 80, 90, 100, 150, 200, 250, and 300 cm), a sample set of images is shown in Figure 2. This resulted in 8700 images of 435 subjects, each with 10 reference images (session 1) and 10 probe images (session 2).

The camera used was the Canon EOS 6D with 36x24 mm CMOS sensor with 20.2 MP. The camera was set to manual focus by SDK, 100 ISO, aperture of 22, flash synchronization 1/160, and the white balance set to flash. The used lens was Canon EF 50mm f/1.8 STM with distortion of 0.76% and high sharpness that allows for captures at all the considered distances. The capture was performed under front and background flash, and an ambient light that allowed for ICAO compliant images from all distance.

V. Experimental setup

Two models were trained with different levels of perspective diversity. The first, referred to as model *A*, was trained on an image set with less diverse of perception distortion. Model *A* was trained on 6 images per training subject. These images were the ones captured at a distance of 50, 60, 70, 80, 90, and 100 cm from the face. The second model *B*, was trained with more perspective diverse images. This was done

(a) 50 cm (b) 60 cm (c) 70 cm (d) 80 cm (e) 90 cm (f) 100 cm (g) 150 cm (h) 200 cm (i) 250 cm (j) 300 cm

Fig. 2: face images collected at different distances visualize the effect of the capture distance on the perspective distortion. This is more visual at the face edges in in the nose area.

to enable the network to neglect image variations caused by perspective distortion and focus on stable features that increase intra-class variability and inter-class similarity of the identities themselves. Model *B* was trained using the same number of images per training subject (6 images) to enable a fair comparison with model *A*. The images used to train model *B* were the ones captured at the distances 50, 100, 150, 200, 250, and 300 cm. From these images, face regions were detected using a multi-scale sliding window detector based on histogram of oriented gradients (HOG) [41]. This resulted in a square face region that is scaled to 96x96 pixels.

The data was split into 4 parts based on the identities. The parts were used in a four fold evaluation, where each of the four evaluations used 3 parts for training and the fourth for evaluation. This was performed to enable an evaluation on the whole data. Running four evaluations per model also helped presenting a fair comparison between both deep learning models, as the deep learning process involves random processes and does not repetitively land in the same optimization points. The results reported in this paper are averaged error rates over the 4 evaluations. While the training of each model contained only 6 images per identity, the evaluation of the models was performed on images of all distances (10 images per identity).

Evaluations of both models *A* and *B* were run separately. To give a clearer view on the effect of the perspective distortion on face recognition, each image (distance) of each of the reference evaluation identities was compared to each image (distance) of the probe evaluation identities. The achieved performance was reported as FNMR values at different fixed FMR values to show the tradeoff errors values at different operational points.

VI. RESULTS

Figure 3 presents the FNMR values achieved by models *A* and *B* at the FMR values of 0.1%, 1% and 10%. These error rates are shown for each combination of distances between references and probes.

The performance of model *A* clearly deteriorates when comparing images captured at difference distances. This is clearer when the difference in the capture distance is higher, and thus the difference in the perspective distortion. The effect of perspective distortion on the performance is larger when involving images captured at smaller distances (e.g. 50 cm) due to the large variations in the perspective distortions at these distances. For example, comparing references captured at 50 cm distance with probes captured at 60 cm distance results in a three-fold increase in the FNMR at 0.1% FMR. This is not the case when considering references captured at 300 cm, the performance is relatively stable for probes captured at the same distance and down to 150 cm. This is a result of the larger perspective distortion variations at smaller distances in comparison to larger capture distances.

Model *B* is trained on the same amount of data as model *A* (number of subjects and number of images per subject). However, model *B* was trained on data that represent a more diverse range of perspective distortion in an effort to produce face representations that are more robust to this distortion. This robustness can be measured by the verification performance improvement introduced to face comparisons between images captured at different distances, in comparison to model *A*. Figure 3 clearly presents this performance improvement in model *B* induced by the more diverse training data. In this case, the verification performance between faces captured at different distances achieved significantly lower FNMR in comparison to model *A*. For example, considering the verification performance of a reference captured at 50 cm, compared to probes at all distances, model *B* achieved an average FNMR of 5.18% at 1% FMR, down from 7.68% FMR for model *A*. For a reference captured at 300 cm distance, this average FMR value was 4.32% for model *A* and 1.92% for model *B*. It must be noted that this improvement in the verification performance in the cases where the capture distances are different did not significantly effect the already high performance of verifying face images captured at the same distance. These results clearly uncover the vulnerability of deep learning based face recognition solutions to perspective distortion when it is not explicitly considered. They also show the positive improvements of the overall verification performance when training over more perspective-diverse data, without the need to increase the size of the training data. It must be mentioned, that the same limited size data was tested on several state-of-the-art solutions from industrial vendors and showed no errors (FNMR = 0%) for any combination of reference and probe capture distance [7], as reported in annex F of the ICAO draft technical report in [8]. However, this work is concerned with the effect of the nature of the training data on the verification performance and is trained on data of a limited size.

VII. CONCLUSION

Motivated by the increasing interest in the evermore reliable face recognition technologies, this work investigated the effect of perspective distortion on the utilized deep learning-based face verification performance. Using a specially col-

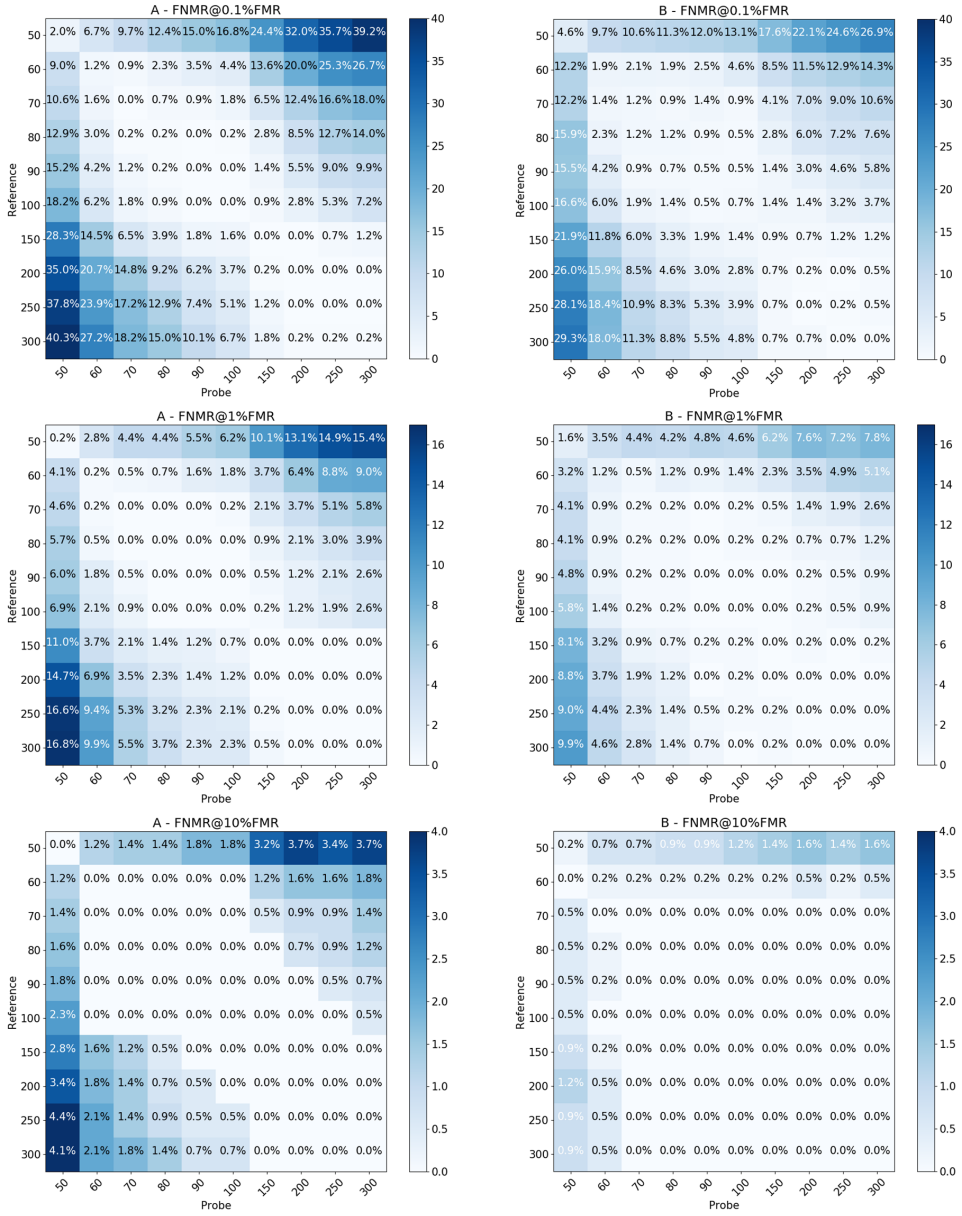

Fig. 3: verification performance (FNMR@FMR) for model A and model B. The significant reduction in FNMR induced by the perspective-diverse training data of model B is visible.

lected database, the achieved results showed that perspective distortion diversity effects the performance of this solution, especially at short capture distances. A previous work studied this effect, however on hand-crafted features, and proposed a solution that reverses the distortion effect on the image level, which requires the knowledge of camera parameters and a high quality capture, and the depth of face. However, this work proposes creating more perspective-robust face representations by simply providing a more perspective-diverse training to build the utilized deep model. This proved to significantly decrease the verification error rates when comparing face images with very different levels of distortion, e.g. the FNMR at 1% FMR was decreased by the perspective-robust representations by around 50%, when comparing references captured at 50 cm with probes captured at 300 cm.

ACKNOWLEDGMENT

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within CRISP.

REFERENCES

[1] A. Krupp, C. Rathgeb, and C. Busch, "Social acceptance of biometric technologies in Germany: A survey," in *2013 BIOSIG - Proceedings of the 12th International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 4-6, 2013*, ser. LNI, vol. 212. GI, 2013, pp. 193–200.

[2] P. Perona, "A new perspective on portraiture," *Journal of Vision*, vol. 7, no. 9, pp. 992–992, mar 2010.

[3] C. H. Liu and A. Chaudhuri, "Face recognition with perspective transformation," *Vision Research*, vol. 43, pp. 2393–2402, oct 2003.

[4] A. Flores, E. M. Christiansen, D. J. Kriegman, and S. J. Belongie, "Camera distance from face images," in *Advances in Visual Computing - 9th International Symposium, ISVC 2013, Rethymnon, Crete, Greece, July 29-31, 2013. Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 8034. Springer, 2013, pp. 513–522.

[5] B. Super, B. Augustine, J. Crenshaw, E. Groat, and M. Theims, "Perspective improvement for image and video applications," Jul. 23 2013, uS Patent 8,494,224.

[6] J. Valente and S. Soatto, "Perspective distortion modeling, learning and compensation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 9–16.

[7] B. Berthe, C. Croll, and O. Henniger, "Face verification robustness and camera-subject distance," *Keesing journal of Documents and Identity*, vol. 55, February 2018.

[8] International Civil Aviation Organisation (ICAO), "ICAO Draft Technical Report: Portrait quality (reference facial images for MRTD)," technical report, Version 0.9, 2017.

[9] N. Damer and K. Dimitrov, "Practical view on face presentation attack detection," in *Proceedings of the British Machine Vision Conference, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[10] R. Swaminathan, M. D. Grossberg, and S. K. Nayar, "A perspective on distortions," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*. IEEE Computer Society, 2003, pp. 594–601.

[11] R. G. Jason P. de Villiers, F. Wilhelm Leuschner, "Centi-pixel accurate real-time inverse distortion correction," vol. 7266, 2008, pp. 7266 – 7266 – 8.

[12] R. Bryan, P. Perona, and R. Adolphs, "Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces," *PLOS ONE*, vol. 7, no. 9, pp. 1–9, 09 2012.

[13] E. A. Cooper, E. A. Piazza, and M. S. Banks, "The perceptual basis of common photographic practice." *Journal of vision*, vol. 12, no. 5, p. 8, May 2012.

[14] S. Du and R. Ward, "Face recognition under pose variations," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 596 – 613, 2006, winners of the student paper competition at the 2005 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) held in Philadelphia PA, provide the state-of-art in their fields of research.

[15] M. Faundez-Zanuy and J. Fabregas, *On the Relevance of Facial Expressions for Biometric Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 32–43.

[16] N. Damer, T. Samartzidis, and A. Nouak, "Personalized face reference from video: Key-face selection and feature-level fusion," in *Face and Facial Expression Recognition from Real World Videos - International Workshop, FFER@ICPR 2014, Stockholm, Sweden, August 24, 2014*, ser. LNCS, vol. 8912. Springer, 2014, pp. 85–98.

[17] N. T. Vetrekar, R. Raghavendra, A. A. Gaonkar, G. M. Naik, and R. S. Gad, "Extended multi-spectral face recognition across two different age groups: an empirical study," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2016, Guwahati, Assam, India, December 18-22, 2016*. ACM, 2016, pp. 78:1–78:8.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[19] Ortega-Garcia et al., "The multiscenario multienvironment BioSecure multimodal database (BMDB)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1097–1111, 2010.

[20] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721–732, 1997.

[21] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," in *2003 IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003), 17 October 2003, Nice, France, Proceedings*. IEEE Computer Society, 2003, pp. 157–164.

[22] C. H. Liu and J. Ward, "Face Recognition in Pictures is Affected by Perspective Transformation but Not by the Centre of Projection," *Perception*, vol. 35, no. 12, pp. 1637–1650, dec 2006.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823.

[24] A. K. Gangwar and A. Joshi, "Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition," in *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*. IEEE, 2016, pp. 2301–2305.

[25] H. Proena and J. C. Neves, "Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks," *IEEE Transactions on Information Forensics and Security*, vol. PP, no. 99, pp. 1–1, 2017.

[26] B. Ribeiro, I. Gonçalves, S. Santos, and A. Kovacec, "Deep learning networks for off-line handwritten signature recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 16th Iberoamerican Congress, CIARP 2011, Pucón, Chile, November 15-18, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7042. Springer, 2011, pp. 523–532.

[27] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, ser. LNCS, vol. 9909. Springer, 2016, pp. 579–596.

[28] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, and J. Verbeek, "Frankenstein: Learning deep face representations using small data," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 293–303, 2018.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1701–1708.

[30] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1–9.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 2017, pp. 4278–4284.

[34] M. Lin, Q. Chen, and S. Yan, "Network in network," *CoRR*, vol. abs/1312.4400, 2013.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 1106–1114.

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[37] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017.

[38] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] A. Karpathy, "Course cs231n: Convolutional neural networks for visual recognition," University Lecture.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. IEEE Computer Society, 2005, pp. 886–893.