FacialSCDnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs

Enrique Bermejo a,d,* , Enrique Fernandez-Blanco b, Andrea Valsecchi c,a, Pablo Mesejo a, Oscar Ibáñez b,c,a, Kazuhiko Imaizumi d

a Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada 18071, Spain
b Faculty of Computer Science, CITIC, University of A Coruña, A Coruña 15071, Spain
c Panacea Cooperative Research S. Coop., Ponferrada 24402, Spain
d Second Forensic Biology Section, National Research Institute of Police Science, Chiba 277-0882, Japan

A R T I C L E   I N F O

Keywords:
Subject-to-camera distance
Perspective distortion
Photography
Human identification
Deep-learning
Transfer learning

A B S T R A C T

Facial biometrics play an essential role in the fields of law enforcement and forensic sciences. When comparing facial traits for human identification in photographs or videos, the analysis must account for several factors that impair the application of common identification techniques, such as illumination, pose, or expression. In particular, facial attributes can drastically change depending on the distance between the subject and the camera at the time of the picture. This effect is known as perspective distortion, which can severely affect the outcome of the comparative analysis. Hence, knowing the subject-to-camera distance of the original scene where the photograph was taken can help determine the degree of distortion, improve the accuracy of computer-aided recognition tools, and increase the reliability of human identification and further analyses.

In this paper, we propose a deep learning approach to estimate the subject-to-camera distance of facial photographs: FacialSCDnet. Furthermore, we introduce a novel evaluation metric designed to guide the learning process, based on changes in facial distortion at different distances. To validate our proposal, we collected a novel dataset of facial photographs taken at several distances using both synthetic and real data. Our approach is fully automatic and can provide a numerical distance estimation for up to six meters, beyond which changes in facial distortion are not significant. The proposed method achieves an accurate estimation, with an average error below 6 cm of subject-to-camera distance for facial photographs in any frontal or lateral head pose, robust to facial hair, glasses, and partial occlusion.

1. Introduction

 Facial identification has become an extremely relevant topic during the last decade. The revolution of deep learning and automatic facial recognition systems have led to a market expansion from the fields of law enforcement and forensic science to areas in the private sector: retail, multimedia applications, or security. Moreover, the development of imaging technology has improved both the quality and the availability of photographic data, which has also contributed to the application of multi-modal identification techniques by using 3D facial models or medical images (Prior et al., 2009; Yoshino et al., 2000). Specifically, in the field of forensic science, the ability to identify people from only facial features can be addressed by the application of different kind of techniques, e.g., facial comparison (Evison & Vorder Bruegge, 2010; Spaul, 2009), facial reconstruction (Wilkinson, 2010), or craniofacial superimposition by means of skeletal remains (Clement & Ranson, 1998; Damas et al., 2020; Rosario Campomanes-Álvarez et al., 2015), among others.

 Individualization or identification techniques are manually performed by experts with or without the assistance of automatic systems. Forensic experts analyze available data, often obtained in uncontrolled scenarios as evidence from caseworks, and then evaluate the anatomical characteristics of an unknown individual compared to a known individual (one-to-one), or many others in large and standardized databases (one-to-many). For example, facial comparisons can involve approaches such as holistic comparisons, morphological analysis, photo-anthropometry, or image superimposition (Zeinstra et al., 2018). For the analysis to be reliable and conclusive, available data,
in this case facial pictures, are required to be in acceptable conditions (quality, resolution, focus, or illumination need to meet a minimum standard), and the scene context (camera viewpoint, head pose, facial expression) needs to be as neutral and representative as possible (Edmond et al., 2009; Spaun, 2009). These requirements are established to ensure that facial traits in the images are faithful to the individual’s anatomical features and to enable the robust application of forensic techniques.

Several studies have previously identified the limitations of current approaches, especially when automatic recognition systems are considered. Pose, illumination and expression are commonly referred to as the main challenging factors (Atsuichi et al., 2013; Jain et al., 2012; Tripathi & Jalal, 2021). However, aspects such as occlusion, gender, ethnicity, or age also have a relevant influence on the identification procedure in both human practitioners and automatic systems (Buolamwini, 2018; Li et al., 2011; Phillips et al., 2011). In this work, our interest will reside in the impact of perspective distortion on facial images, i.e., the deformation of facial traits (ears, nose, and head shape) as a result of the subject being close to the camera during the photograph acquisition (Ward et al., 2018). The effect of perspective distortion in social interactions is known to introduce bias in human perception (Bryan et al., 2012; Trebicj et al., 2016), as well as impair the performance of facial recognition (Liu & Chaudhuri, 2003; Liu & Ward, 2006). For that reason, it is also expected to have a negative effect in automatic recognition systems (Damer et al., 2018; Riaz & Beetz, 2012; Valente & Soatto, 2015).

There is a strong interest in improving the accuracy and robustness of current identification techniques to maximize their quality and value as court evidence. As a result, various efforts have been made to address different sources of uncertainty involved in the procedure from an anatomical point of view (Bicalho et al., 2018; Campomanes-Alvarez et al., 2018). However, it is important to highlight that knowledge of the scene context and acquisition parameters of the camera is essential prior to the application of any anatomical analysis. In most situations, configuration parameters of the camera can be extracted when the metadata of the image is available, while head pose can be easily obtained by visual inspection or by using an automatic pose estimation tool (Lathuilière et al., 2017; Merckx et al., 2010; Suman, 2008). In contrast, the estimation of subject-to-camera distance (SCD) poses a major challenge while having a substantial influence on the superimposition or facial comparison between two pictures, as thoroughly analyzed by Stephan (2015). There is an intimate relationship between distance and focal length that can result in a misperception where the size of an object can be analogous for different combinations of SCD and focal length (He et al., 2018).

The estimation of SCD in facial images opens the possibility of quantifying the differences in distortion among two sets of photographs and, more importantly, to reproduce the conditions of the original scene when 3D facial models or skeletal remains are available. This feature is an essential consideration for both manual and automatic identification techniques, improving the reliability of one-to-one comparisons by measuring and controlling a major source of uncertainty.

In summary, the main contribution of this paper is two-fold:

- The introduction of a deep learning approach to estimate SCD of facial photographs for its use in human identification applications. A loss metric is also defined to guide the learning procedure based on the perspective distortion.
- A novel benchmark dataset to train and evaluate the behavior of the machine learning model in realistic scenarios.

2. Background

Pictures, either analog, digital, or computer-generated images, are essentially the result of projecting light from a 3D scene through a point onto a 2D plane. This is known as perspective projection, and its geometry is represented by a camera model. Such model provides the necessary mathematical formulations to compute the coordinates of a point in the photograph from its corresponding point in the 3D scene. Thus, the camera model specifies a set of camera parameters that will define the specific characteristics of the photographic scene at the moment of acquisition. For instance, those parameters can define the camera position and its orientation, the angle of view (focal length), the depth-of-field (aperture), or the sensor size that will determine the physical dimensions of the image.

For the sake of simplicity, a pinhole camera model is assumed to represent an ideal system not affected by optical distortions of lenses, i.e., barrel, pinhole, or tangential (Kingslake, 1992). In practice, camera calibration or photographic software can easily correct the effects of different lens distortions. However, a completely separate category of optical aberration that can affect a camera model is perspective distortion. Its effect is solely determined by the relative position of the camera in the scene and it is often neglected in applications involving facial analysis. In portrait photography, this parameter will determine the subject-to-camera distance. A combination of wide lenses and short distances will generate a large magnification of facial features, introducing a noticeable variation in the perceived depth between nose and ears. Such effect is common in the ‘selfie’ picture format (Ward et al., 2018), where elements closer to the camera (nose, eyes) appear enlarged in contrast to far elements (ear, hair), which appear compressed. Fig. 1 visually illustrate the effects of facial distortion in relation with the SCD.

There is no consensus in the literature to establish a tentative distance where perspective distortion can be considered negligible for facial analysis. Distances around 2 m are often considered free from facial distortion due to the influence of human perception and regular habits of social interaction. In a recent study on facial distortion, Stephan (2015) analyzed how the relationship between SCD and perspective follows a logarithmic decay. Considering average facial dimensions, the decay reaches an acceptable value of 1% distortion at 12 m. According to this study, photographs where SCD is above 6.1 m could be compared within that 1% distortion, which renders an anatomical error of 2 mm in real-sized facial comparisons.

The estimation of the perceived depth of a scene is hindered by an optical principle, the relationship between SCD and focal length (He et al., 2018). As the focal length parameter controls the field of view, a similar scene could be obtained at different combinations of SCD and focal length, disregarding the effects of perspective distortion. Accordingly, in order to estimate the SCD, knowledge of the focal length is imperative. This constraint can be overcome by accessing the digital image metadata to obtain information about the camera type, sensor size, focal length, or image resolution when available. The angle of view is controlled by two camera parameters: focal length and sensor size, both of which determine how the scene context is captured.

A standardized size of 36 mm × 24 mm is usually considered for the image sensor. This standard, also known as full frame or 35 mm equivalent, is established as reference system to provide comparison among different digital camera models and manufacturers with respect to the 35 mm film camera format. The focal length of any lens is relative to a specific camera film or sensor size; e.g., a 50 mm lens meant for a 35 mm film camera will exhibit a larger focal length when mounted on a camera with a smaller sensor size. For this reason, focal length is often reported as “35 mm equivalent focal length” format, meaning that the focal length figure reported is relative to a 35 mm camera sensor, as opposed to the specific camera sensor. The choice of the 35 mm format is due to historical reasons. In this paper, we also follow this convention. FacialSCDnet can be used with any camera sensor size by simply providing the 35 mm equivalent focal length value, which can be obtained using Eq. (1). It should be noted that the 36 is not a typo, but a correction value.

\[
\text{focal length (35 mm)} = \frac{\text{focal length} \times 36 \text{ mm}}{\text{camera sensor width}} \tag{1}
\]
3. Related work

In the field of machine learning, the topic of depth estimation from a single photograph has recently gained a lot of attention. This problem is addressed as a dense regression task where each pixel of the photograph is mapped to a depth value by using implicit pictorial cues (Van Dijk & De Croon, 2019). However, the estimation results in a depth map where information is encoded relative to the objects of the scene, not the camera itself. Therefore, information about SCD cannot be extracted unless the real size of the objects is known. Nevertheless, this approach has many applications such as pose estimation, object detection, or segmentation (Cho et al., 2021; Shotton et al., 2013). An interesting technique for depth estimation is based on defocus estimation (Gur & Wolf, 2019; Maximov et al., 2020). Instead of relying on texture, size, or perspective cues, the blurring effect of out-of-focus objects is used to guide the estimation, taking into account the camera’s depth-of-field.

On a related subject, different approaches have been proposed to address the effects of perspective in facial images, aiming at correcting or recovering natural features from distorted facial images (Fried et al., 2016; Shih et al., 2019). Such methods are oriented towards multimedia applications with marginal use in identification applications as they involve an additional manipulation of the photographic evidence. Nevertheless, they provide useful knowledge for its application in the field at hand. For instance, Zhao et al. (2019) introduced a method for perspective undistortion where the first step is based on a DL network for camera distance prediction. In this work, SCD is not estimated but classified into different sampled distances to determine the amount of distortion to correct in subsequent steps.

The first method to address the metric estimation of SCD from a single photograph was proposed by Flores et al. (2013), where a set of facial landmarks is used in a series of synthetic images to estimate SCD independent of the camera focal length. PerspectivesX, proposed by Stephan (2017) to model uncertainty in the craniofacial superimposition technique. It is also based on the location of an anatomical feature, the palpebral fissure length between two easily determinable and accurate landmarks. This method allows for an accurate estimation of SCD for a known focal length. The limitations of this technique include as requirement the manual interaction of an expert to annotate the landmarks, and a head rotation below 30°. The approach consists on the application of a straightforward equation, yet it requires the estimation of the real size of the palpebral fissure from landmarks according to anatomical studies that depend on a specific age group, sex, and population.

4. Proposed method

In order to achieve the automation of the SCD estimation, we propose the application of a Deep Learning (DL) approach capable of regressing the metric distance of individuals directly from photographs. Following a deep architecture we avoid a critical constraint, i.e., the requirement of detecting a particular anatomical feature to guide the estimation procedure. As seen in the previous section, such constraint limits the application of current methods to frontal or quasi-frontal facial images. With this approach we aim to ease the estimation of SCD at any head pose from frontal to lateral profile.

Recent technological developments have increased the popularity of neural networks with the surge of DL and its application to different domains (Lecun et al., 2015). In particular, the use of Convolutional Neural Networks (CNNs) has widespread for applications involving image processing due to their capabilities for feature learning. However, DL architectures are known to require a significant amount of training data to negate the effects of over fitting and provide generalizable results. To overcome this problem a technique known as transfer learning is usually applied, where complex or deep architectures can be reused and adapted to a different problem in less time and requiring a smaller amount of data. For that purpose, architectures known to perform well in a wide range of problems have been usually considered, e.g., ResNet (He et al., 2016) or VGG-16 (Simonyan & Zisserman, 2015).

In particular, we take advantage of the pre-trained parameters of the VGG-16 network (Simonyan & Zisserman, 2015), widely recognized...
for its ability of migration learning tasks. As an example, architectures based on VGG have been successfully proposed for facial recognition (Parkhi et al., 2015), which motivated our choice for this contribution. Therefore, the convolutional layers of the VGG-16 model used in our proposal are pre-trained with the ImageNet dataset (Russakovsky et al., 2015), and fine-tuned with a dataset specifically designed for the addressed problem, introduced in Section 5.1.

Due to physiognomic variations of facial shape and size across different individuals, estimating SCD at large distances poses a significant challenge. The relationship between SCD and perspective distortion is that of a logarithmic decay. This means that small SCD values correspond to a large distortion, while the distortion decreases promptly as the SCD increases, becoming almost unnoticeable beyond 6 m. As a consequence, a small mistake in estimating the SCD will have a large effect on the predicted amount of distortion depending on whether the true SCD value was large or small. For this reason, we employed an error function (or metric) that penalized misestimation of low SCD values. Thus, the loss function is defined as the averaged absolute error of the relative facial distortion as follows:

\[
\text{Distortion} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n},
\]

where \(y_i\) correspond to the true label, and \(x_i\) to the predicted value of the measured facial distortion, computed as the distortion factor (DF):

\[
D_f = \frac{1}{1 + \frac{SCD}{d}}.
\]

In this equation, \(d = 12.6572\) cm corresponds to a value derived from geometric calculations performed by Stephan (2015) to experimentally obtain the distortion factor over a representative human head of average size, relative to the SCD in a photograph. According to adult human measurements, \(d\) is similar to the average facial depth (Brinkley, et al., 2016). Thus, this constant value is used as a conversion factor to establish a relationship between facial features and the perspective distortion affecting the photographs.

5. Experimental framework

5.1. Dataset generation protocol

The specific constraints of the problem of SCD estimation require the focal length of the photographs to be known beforehand. *In the wild* datasets, commonly used in DL applications as is the case of facial recognition, landmark detection, or pose estimation, are not appropriate for this purpose. Such datasets based on *found data* available on the internet lack the necessary information to address this problem, as images are usually cropped and image metadata is not available. Burgos-Artizzu et al. (2014) introduced a dataset specifically designed for SCD estimation, named Caltech Multi-Distance Portraits (CMDP), where images were cropped and processed to fit a common format. Unfortunately, focal length varies along with the sampled distances in the CMDP dataset, which renders it impractical for this problem. Therefore, in the absence of a publicly available dataset for this purpose, we detail in this section the procedure followed to gather a specific dataset of facial images at a distance.

In order to collect an optimal amount of data required to train the CNN, the proposed dataset features two collections of data that expand the number of available images. First, a synthetic set of images have been compiled by using the Stirling ESRC 3D Face database,\(^2\) generating simulated photographs using 3D face models with different expressions. In particular, a total number of 315 facial models from 54 different individuals are used to generate approximately 150 K synthetic photographs. The second collection corresponds to digital photographs of 28 individuals following an acquisition protocol detailed in Fig. 2. Additionally, a subset of 3D facial models for 15 volunteers among the participants have been acquired using a 3D physiognomic range finder (Fiore, NEC Corp., Tokyo, Japan) (Ogawa et al., 2015). This subset is used to complement the collection with additional synthetic images and to validate the distance markers used during the acquisition. Photographs have been acquired using two digital cameras, a Nikon D5200 and a Fujifilm X-T30, equipped with a 18–55 mm zoom lens, and a Huawei P20 smartphone which camera is equivalent to a focal length of 27 mm. A total number of 20 K digital images comprise this second set. The structure of the collected dataset is detailed below:

- **Focal length.** Four different focal lengths (in full-frame or 35 mm equivalent) have been considered: 27 mm, 35 mm, 50 mm, 85 mm. These focal lengths were chosen as standard or common lens in the market, from wide angle to medium telephoto lens, in an effort to represent commonly available photographs. In particular, most smartphones nowadays contain a camera lens equivalent to 26–28 mm, while 85 mm is regarded as preferred choice for portraits with normal facial features.

- **Distance.** As shown in Section 2, perspective distortion affects facial images as a logarithmic decaying function of the distance. For that reason, a total of twelve different distances from 50 cm to 6 m were sampled, measured from the focal plane of the camera to the eye plane of the individuals. Specifically, distances were spaced at increasing intervals of 10 cm, 20 cm, 30 cm, 50 cm, and 1 m.

- **Pose.** Seven different head poses were photographed from left profile to right profile, roughly at each 20° of rotation. This movement was performed twice, as participants were instructed to remain in a neutral facial expression the first pass, and smile or speak during the second one. In total 14 photographs were obtained per participant at a determined distance, per each focal length. To add variability in the sample, some individuals were photographed standing while others were seated in a rotatory chair. As for the augmented dataset, synthetic photographs were simulated for each facial model at a random rotation in the horizontal axis (−90°, 90°), and inclinations in the vertical and distal planes in the range (−15°, 15°). As facial models have a fixed expression, 10 simulations were generated for each distance and focal length.

To summarize, the photographic dataset gathered to validate the proposed method include a representative sample of various head poses, facial expressions and SCD distances, in addition to different physiognomies, occlusions related to facial hair or accessories such as glasses for people of diverse ethnic groups, gender and age. The dataset will be publicly available at the skeleton-id.com website. (see Section 7, data availability).

5.2. Experimental methodology

To deal with the ambiguity generated by different combination of SCD and focal lengths, we define four deep learning models, each one associated with one of the focal lengths considered in the dataset. For simplicity, we refer to this system as FacialSCDnet.

The structure of each CNN is based on VGG-16, initialized with the pre-trained Imaganet weights and stripped of the top layer. In order to adapt the architecture to the specific problem of SCD estimation, the five convolutional blocks are preserved, and two fully connected layers are attached and trained from scratch. The last layer of the model consists of a dense linear activation that performs the regression task. Thus, the output of the network determines a predicted metric distance for a particular facial photograph. Fig. 3 provides a visual representation of the CNN architecture.

---

\(^2\) Stirling ESRC 3D Face database is available at [http://pics.psych.stir.ac.uk/ESRC/index.htm](http://pics.psych.stir.ac.uk/ESRC/index.htm)
Fig. 2. Synthesis of the photograph acquisition protocol depicting the conditions of the scene: focal lengths used, range of distances, and head pose rotation angles. Different distances were considered for separate volunteers to cover a wide variety of distances in the dataset.

Fig. 3. Diagram representing the network architecture based on VGG-16. Fully connected layers were re-trained to adapt the network to the regression problem of SCD estimation.

Table 1
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Options</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>[adam,adagrad,rmsprop]</td>
<td>adam</td>
</tr>
<tr>
<td>Learning rate</td>
<td>[10⁻⁶, 10⁻³]</td>
<td>10⁻⁴</td>
</tr>
<tr>
<td>Batch size</td>
<td>[32,128]</td>
<td>64</td>
</tr>
<tr>
<td>Patience</td>
<td>[2,4]</td>
<td>3</td>
</tr>
<tr>
<td>Early stopping</td>
<td>[4,8]</td>
<td>6</td>
</tr>
</tbody>
</table>

A hyperparameter tuning phase was considered to determine an optimal configuration for the proposed architecture. In particular, the Tree-structured Parzen Estimator Approach (TPE) (Bergstra et al., 2011) algorithm was employed to guide and speed-up the search procedure, and the final parameters used to train the models are shown in Table 1.

The model training process followed a two-stage sequence. First, the models are trained over the larger synthetic dataset to learn the relationship between SCD and facial features. Then, models are fine-tuned using the real dataset. Such approach allows the CNNs to perform the estimation task by focusing only on facial features while other body parts are ignored.

The two datasets, synthetic and real, were shuffled and split into train, validation, and test subsets following a 55:15:30 proportion. To prevent data leakage, the test set was constructed by selecting all the images (with different head poses) at a particular distance from random individuals. In addition to that, we ensure that all the real photographs (at every distance and focal) from at least three individuals were unseen by the CNNs until the test stage. A series of performance measures were considered to both guide and validate the regression model. As mentioned in Section 2, the loss function is based on the relative facial distortion. In addition, the mean absolute error (MAE) and mean relative error (MRE) were computed between the predicted and the actual SCD of each photograph.

All the experiments were performed using a platform configured with forty-cores CPU Intel(R) Xeon(R) CPU E5-2630 (2.20 GHz), and NVIDIA TITAN XP 12 GB GPU running on Ubuntu 16.04.2 LTS. The software tools included CUDA 8.0, CUDNN 7.5, Python 3.6. The experiments were implemented in the framework Tensorflow 1.14 using the pre-trained model and weights from Keras 2.2.4. Image augmentation was used during training to increase the robustness of the network. An online augmentation was performed to provide different random backgrounds to the images, adding rotation, blur, noise, and saturation, color, and illumination changes into the training images (see Fig. 4). In addition, a partial occlusion of the images was also considered by removing parts of the original image (cutout) or a percentage of the total image (pixel dropout). The library Imgaug was considered for that purpose in order to avoid overfitting the data. Table 2 summarizes the considered augmentation techniques, the probabilities for their application during each training batch, and the ranges for the different effects applied to the input images.

6. Experimental results

The results presented in this section correspond to the evaluation of the four different neural networks over the test partition of the data. Table 3 summarizes the results of the three considered metrics (MAE, MRE, and relative distortion). The 99th percentile of the resulting error for each metric is included in this table to gain insight of the prediction
Table 2
Distortion margins for the augmentation techniques, and probabilities for its random application on training images.

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation (% per axis)</td>
<td>1.0</td>
</tr>
<tr>
<td>Rotation (°)</td>
<td>1.0</td>
</tr>
<tr>
<td>Blur</td>
<td>0.25</td>
</tr>
<tr>
<td>Gray scale</td>
<td>0.25</td>
</tr>
<tr>
<td>Hue and saturation</td>
<td>0.25</td>
</tr>
<tr>
<td>Gaussian noise</td>
<td>0.25</td>
</tr>
<tr>
<td>Simplex noise</td>
<td>0.1</td>
</tr>
<tr>
<td>Cutout (% image size)</td>
<td>0.1</td>
</tr>
<tr>
<td>Pixel dropout (% of pixels)</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Fig. 4. Example of augmentation techniques applied to different photographs during training.

Fig. 5. Histograms depicting error distribution for the different focal lengths considered in terms of (a) prediction error (cm), and (b) mean relative error (%).

In an attempt to provide interpretability for the behavior of the models regarding the regression task, Fig. 7 depicts a number of attention layer representations for the neural networks. In particular, different visualization techniques, i.e., gradient saliency, guided backpropagation, deep Taylor decomposition, and layer-wise relevance propagation, are shown here to provide insight of the image features guiding the prediction results (Alber et al., 2019). Such figure also serves as confirmation that the models are focusing in relevant parts of the facial physiognomy to perform the predictions. From these examples, it is possible to identify how background information is mostly ignored, as well as facial hair without heavily misleading the prediction. Meanwhile, facial features seem to draw most of the attention in contrast to other parts of the body regardless its posture (seating or standing).

A complementary experimentation has been performed, and summarized in Table 4, with the aim to quantitatively study the influence...
Table 3
Prediction results for each FacialSCDnet model according to MAE, MRE, and relative distortion metrics. Mean, standard deviation, median, and 99th percentile error results are summarized for each metric. Measurement of the coefficient of determination is also shown.

<table>
<thead>
<tr>
<th>Evaluation Metrics</th>
<th>f27</th>
<th>f35</th>
<th>f55</th>
<th>f85</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE (cm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (sd)</td>
<td>5.9 (9.8)</td>
<td>4.5 (8.5)</td>
<td>5.3 (11.0)</td>
<td>3.4 (5.0)</td>
<td>4.8 (9.1)</td>
</tr>
<tr>
<td>Median</td>
<td>2.7</td>
<td>2.2</td>
<td>2.6</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td>99th perc.</td>
<td>51.0</td>
<td>37.4</td>
<td>76.6</td>
<td>24.2</td>
<td>51.7</td>
</tr>
<tr>
<td>MRE (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (sd)</td>
<td>2.9 (3.1)</td>
<td>2.3 (2.8)</td>
<td>2.6 (2.8)</td>
<td>2.2 (2.1)</td>
<td>2.5 (2.8)</td>
</tr>
<tr>
<td>Median</td>
<td>1.7</td>
<td>1.3</td>
<td>1.7</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>99th perc.</td>
<td>14.1</td>
<td>13.9</td>
<td>15.0</td>
<td>9.0</td>
<td>13.4</td>
</tr>
<tr>
<td>Distortion (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (sd)</td>
<td>0.2 (0.2)</td>
<td>0.2 (0.2)</td>
<td>0.2 (0.2)</td>
<td>0.2 (0.2)</td>
<td>0.2 (0.2)</td>
</tr>
<tr>
<td>Median</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>99th perc.</td>
<td>1.1</td>
<td>1.1</td>
<td>0.9</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Adjusted $R^2$ score</td>
<td>0.994</td>
<td>0.996</td>
<td>0.994</td>
<td>0.998</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Fig. 6. Graphics illustrate the resulting prediction error (cm) for each one of the considered models: (a) depicts the direct comparison of actual to predicted SCD distances; and (b) violin plot that outlines the comparison of the statistical error distribution for distances below and above a threshold of 2 m.

of the real data amount on the model accuracy. One of the main motivations for carrying out this experimentation is to know the approximate number of facial photographs needed to fine-tune a new SCD estimation model. This number becomes relevant when optimizing the acquisition of data to train specific models for focal distances other than the four used in this work. Results show a sustained accuracy improvement for each test, which suggests the considered amount of data is adequate to obtain a competitive performance. The main conclusion derived is that networks based on longer focal lenses may require less data to achieve a similar level of accuracy.

6.1. Comparison against state-of-the-art

A subset of synthetic photographs are used here to establish a comparison with PerspectiveX (Stephan, 2017). The use of 3D models allows us to accurately locate the landmarks required for the measurement of the palpebral distance. In particular, 1080 frontal or quasi-frontal images (with a maximum of 15° of lateral rotation) from the synthetic dataset were considered for this comparison. Results are displayed in Table 5. It is worth noting that PerspectiveX factors two error sources in the predictions: the estimation of the palpebral fissure length, and the estimation of the SCD from a known measurement. However, despite its simplicity, the results obtained by PerspectiveX are within the 1% distortion considering the average values. The 99th percentile shows how the distortion error can reach a 3% error in some situations, much higher than results obtained by our proposal (1.1%). In terms of metric errors, the advantage of FacialSCDnet is noticeable, where the average MAE ($4.8 \pm 9.1$ cm) contrasts with the results of PerspectiveX ($38.9 \pm 30.2$ cm).

When compared to PerspectiveX, FacialSCDnet proved to be a reliable and robust method, able to provide highly accurate estimations,
especially at closer SCD distances. Furthermore, CNNs are not constrained by the necessity of identifying anatomical features to guide the prediction task, increasing the flexibility of our proposal to deal with lateral photographs. Such benefit is also relevant when comparing with other approaches such as Google MediaPipe Iris (2020). Authors report different results when comparing photographs with facial accessories such as glasses (4.8 ± 3.1% MRE), or without (4.3 ± 2.4% MRE). In contrast, FacialSCDnet manages to obtain more accurate results (2.5 ± 2.8% MRE) in an heterogeneous dataset, which outlines its robust performance when considering different head poses, expressions, and facial occlusion from facial hair or glasses.

7. Conclusion

In this paper, we introduced a fully automatic approach for subject-to-camera distance estimation in photographs to deal with the effects of perspective distortion, named FacialSCDnet. A Convolutional Neural Network was considered for distance regression without requiring human interaction nor explicit anatomical information to guide the procedure. This opens the possibility of rapidly estimating differences in facial distortion between photographs when performing any form of facial identification, a factor that is crucial for techniques involving photographic comparison.

To train and assess the performance of our proposal, a novel benchmark dataset was collected using a combination of synthetic and real photographs taken at different distances from 0.5 to 6 m, where facial distortion is more pronounced. The factor of not requiring any anatomical information such as facial landmarks or iris detection imply our method is robust to partial occlusion and profile poses where eyes are not visible, in contrast to the limitations of current methods in the literature.

The results revealed that an accurate estimation of the SCD can be achieved automatically. Regarding the comparison with state-of-the-art methods, FacialSCDnet obtained remarkable results, outperforming such methods when using frontal images. In addition to introduce the flexibility to predict using lateral or profile poses, the proposed method is robust to occlusion (facial hair, glasses), facial expression, and noise.

We acknowledge the dataset used in this contribution is limited and, as reported for other works based on deep learning, can be biased by the reduced number of individuals from the acquired photographs for the real dataset and the fact that it was comprised by only adults. Nevertheless, no differences were found during testing on individuals with distinct features never seen from the networks, nor the use of synthetic or real images. To further increase the robustness of our proposal to gender, ethnicity, and age, we intend to augment the number of photographs available in the dataset. In addition, we aim to extend the functionality of FacialSCDnet by integrating the estimation of head pose along with the SCD prediction. Moreover, as future work we will perform an extensive study on the impact of SCD estimation for computer-assisted human identification methods such as facial comparison (Martos et al., 2018), or craniofacial superimposition (Campomanes-Álvarez et al., 2018; Damas et al., 2020; Rosario Campomanes-Álvarez et al., 2015).

CRedit authorship contribution statement

Enrique Bermejo: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. Enrique Fernandez-Blanco: Conceptualization, Software, Resources, Validation, Writing – review & editing. Andrea Valsecchi: Conceptualization, Methodology, Investigation, Validation, Writing – review & editing. Pablo Mesejo: Conceptualization, Methodology, Investigation, Writing – review & editing. Oscar Ibáñez: Conceptualization, Investigation, Writing – review & editing, Supervision. Kazuhiro Imaizumi: Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Data availability

The photographic dataset described in Section 5.1 will be publicly available at the skeleton-id.com website. Furthermore, the website will host a web service for SCD estimation in photographs based on FacialSCDnet. Any additional data regarding the trained models can be shared with the community upon formal request to the authors.

Acknowledgments

Dr. Bermejo’s work has been supported by the Japan Society for the Promotion of Science (JSPS) as International Research Fellow (Standard Fellowship).

Dr. Valsecchi’s fund is supported by the Spanish Ministry of Science and Innovation grant [Ref: PTQ-17-09306].

Dr. Mesejo’s fund is supported by the European Commission H2020-MSCA-IF-2016 through the Skeleton-ID Marie Curie Individual Fellowship (Ref: 746592).

Dr. Ibáñez work is funded by Spanish Ministry of Science, Innovation and Universities-CDTI: Neotec program 2019 [Ref: EXP-00122609/SNEO-20191236] and also under grant RYC2020-029454-I.

They wish to acknowledge the support received from the Centro de Investigación de Galicia “CITIC”, funded by Xunta de Galicia, Spain and the European Union (European Regional Development Fund-Galicia 2014–2020 Program), by grant EDAI319/01/09.

Additionally, this work was supported by the Grant-in-Aid for JSPS Fellows [Ref: 19F19119], by the Spanish Ministry of Science, Innovation and Universities, and European Regional Development Funds (ERDF), under grant EXASOCO [Ref: PGC2018-101216-B-I00], by Xunta de Galicia, Spain under grant number ED431C 2018/49, and by the Regional Government of Andalusia under grant EXASIFI [Ref: P18-FR-42626]. Funding for open access publication was provided by the University of Granada: CBUA, Spain.

References


