ORIGINAL PAPER

Kinship verification from facial images and videos: human versus machine

Miguel Bordallo Lopez[1] · Abdenour Hadid[1] · Elhocine Boutellaa[1] · Jorge Goncalves[2] · Vassilis Kostakos[2] ·
Simo Hosio[3]

Abstract
Automatic kinship verification from facial images is a relatively new and challenging research problem in computer vision. It consists in automatically determining whether two persons have a biological kin relation by examining their facial attributes. In this work, we compare the performance of humans and machines in kinship verification tasks. We investigate the state-of-the-art methods in automatic kinship verification from facial images, comparing their performance with the one obtained by asking humans to complete an equivalent task using a crowdsourcing system. Our results show that machines can consistently beat humans in kinship classification tasks in both images and videos. In addition, we study the limitations of currently available kinship databases and analyzing their possible impact in kinship verification experiment and this type of comparison.

Keywords Kinship verification · Face analysis · Biometrics · Crowdsourcing

1 Introduction

It is common practice for humans, to visually identify relatives from faces. Relatives usually wonder which facial attributes do a new born inherit from each parent. The human ability of kinship recognition has been the object of many psychological studies [21,24]. Inspired by these studies, automatic kinship (or family) verification [30,84] has been recently considered as an interesting and open research problem in computer vision and it is receiving an increasing attention by the research community.

Automatic kinship verification from faces aims to determine whether two persons have a biological kin relation by comparing their facial attributes. This is a difficult task that sometimes needs to deal with subtle similarities that often escape the human eye.

Kinship verification has a role in numerous applications. In addition to biological relation verification, kinship estimation is an important feature in the automatic analysis of the huge amount of photographs daily shared on social media, since it helps understanding the family relationships in these photographs. It can also be used for automatically organizing family albums and generating family trees based on present or historical photographs. In addition to image classification, kinship verification proves also useful in cases of missing children and elderly people with reduced cognitive capabilities, as well as in kidnaping cases.

All these applications assume an automatic kinship verification system able to assess kin relationships from limited input data. However, and despite the recent progress, kinship verification from faces remains a challenging task. It inherits the research problems of face verification from images captured in the wild under adverse pose, expression, illumination and occlusion conditions.

In addition, kinship verification should deal with wider intra-class and inter-class variations. Moreover, automatic kinship verification can face new challenges since unbalanced datasets naturally exist in a family, and a pair of input images may be from persons of different sex and/or with a large age difference.

✉ Miguel Bordallo Lopez
miguel.bordallo@oulu.fi

1 Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

2 School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

3 Center for Ubiquitous Computing, University of Oulu, Oulu, Finland

This paper aims at answering one question: How do humans compare against machines in kinship verification tasks? Among the sizable literature in kinship verification, studies on human perception of kinship remain sparse, and they are often conducted with a small subset of the available data. In addition, many times, the experimental setup for comparing human and machine performance in kinship verification tasks is inherently different.

In this context, we investigate the state of the art of automatic kinship verification approaches from both images and videos and compare their performance with the one obtained by asking humans to complete an equivalent task. To assess the capability of the automatic methods, we use an equivalent setup based in crowdsourcing that allows fair comparison of machines and humans. In addition, we analyze for the first time the possible sources of experimental bias when making this type of comparison.

The main contributions of this paper include: (i) An extensive review of the literature in kinship verification, covering psychological studies and computational models; (ii) a crowdsourcing system for measuring human performance in kinship verification tasks; (iii) analysis of the comparative performance between humans and machines, showing that machines can consistently beat humans in kinship classification tasks in both images and videos; (iv) a description of the limitations of currently available databases and their potential sources of bias.

2 Review of literature on kinship verification

2.1 Psychological aspects of kinship

The human ability to recognize members of our kin has posed many evolutionary benefits. In fact, kin recognition and kinship verification are a process that favors assessing the close relations in groups and predicts an observable differential treatment between members and non-members of the family [34]. The recognition of offspring would be especially important in the allocation of parental investment and in assessing the investment of others [50]. The perception of resemblance has shown to have an effect in paternal investment [58], where self-resemblance is important for the fathers [71], or in the probability of spouse/child abuse [16].

In addition to their own families, humans are also able to match faces of siblings to whom they are not related [14,49,54] and assess the relatedness of pairs of close and distant. This ability is referred as *allocentric kin recognition* and is the focus of human and automatic kinship verification studies.

Already in 1984, Porter et al. [60] showed that strangers are able to match photographs of mothers to their infants, while mothers can recognize photographs of their babies just a few hours after the birth. This suggests that there is indeed facial resemblance among kin, a treat referred in some contexts as a social mirror that can affect the behavior of individuals.

Since humans rely on visual information for many important tasks, facial resemblance is expected to be an indicator used by people to recognize kinship relationships. In fact, there appear to be cues to genetic relatedness in facial features. Since humans possess neural areas, such as the fusiform gyrus, specifically trained to respond to faces, kinship verification from facial information seems to have its own recognition mechanism [59]. A human presented with a face of someone kin-related, as opposed to a totally unknown face, activate brain regions involved in self-face recognition (e.g., anterior cingulate gyrus and medial frontal gyrus). In addition, when presented a face of a kin-related person to someone we know, activates "friend" recognition areas (posterior cingulate and cuneus), again suggesting a need to process for identification [59].

In this context, there have been numerous psychological studies that try to assess the human performance in kinship recognition. In these studies, the participants are asked to assess facial pictures of people belonging to the same family, either between children and parents [4], pairs of siblings [21] or even two adult faces [24]. The results of the previous studies show some interesting consensus on the characteristics of the human ability:

It develops with age. Many studies have focused on the mechanisms of kin recognition in humans and other species. However, a few have addressed the development of such abilities. In humans, adults can match photographs of children and parents faces [40], but children do not perform as well [39]. For example, children aged 5–11 can match photographs of infants to parents at levels above chance, but not parents to infants. Also a consistent finding is that people show better performance on discriminating own age faces [8,36].

Both sexes are equally good at it. Matching kin-related individuals seems to be more difficult when both individuals in a kin relationship are of different sexes. Experiments show that the percentage of verification success increases when comparing mother and daughter or father and son, decreasing for father and daughter or mother and son [58]. Also, the detection of resemblance in children's faces activates different parts of the brain with different activation levels in men and women. However, this is probably due to different decision mechanisms, since there is no evidence of different assessment capabilities between men and women [50,54].

No general rule can be extended to all relationships or family members. In 1995, Christenfeld et al. [20] implied that the infants resemble more their fathers. This could be because in social situations, mothers and their families and friends are more likely to say a newborn resembles its father more than its mother, perhaps to reassure the father of paternity [4,52]. However, subsequent efforts to replicate this work have found evidence claiming that humans are able to better match infant with mothers [52], or to both parents equally well [14,15]. Moreover, studies of child resemblance [57], or of resemblance across the first years of a child's life [4] did not come to a common decision of precise similarity measurement among members.

The whole face should be considered for facial resemblance analysis. This question was first raised by Dal Martello and Maloney [21] which designed their study utilizing children facial segments without considering their gender. The upper half of the face, including the eye region, seems to provide more information than the lower part, since the mouth area changes across development and has fewer stable cues to relatedness [6,21,31]. Experiments using only the left and right halves of the face showed no statistical significance when compared with the whole face [22]. However, DeBruine et al. [24] continued Martello's work utilizing only adult faces, concluding that the performance in the kinship assessment improves when using both halves of the face. They claim that both halves provide independent cues that can be optimally combined in the kin recognition tasks.

The human assessment of facial similarities is usually performed in "patches". The feature types and cues that provide information for kinship verification are still poorly known [4,24]. Meissner and Brigham [53] concluded that recognizing faces may indeed be the result of the processing of shapes and distances of different facial parts or "patches." In line with this result, the spatial information such as the ratio of the distance between these patches is not well processed in the recognition tasks [5], since providing some of the facial "patches" separately does not substantially decrease the verification performance [21].

2.1.1 Conclusion

Summarizing, this set of consensus findings show that facial resemblance among the members of a family can be present in different facial parts or patches, and manifest differently across various family members, showing that the human kinship verification process is learned in a way that is member and patch specific. Based on these findings, we could assume that an automatic verification system that wants to mimic the human abilities should be constructed using information of the different specific kinship relations among different members and different facial parts evaluated separately.

2.2 Computer and machine learning approaches to kinship verification

To the best of our knowledge, the attempts to design computational models based on psychological studies for automatic visual kinship verification started in 2010 and is described in the work of Fang et al. [30]. Using anthropometric methods, this work extracts a number of features that are then ranked by feature performance, selecting the top 14 verification factors. The study concluded that the best feature for family (kinship) verification is the left eye grayscale patch, with average accuracy of 72% on the collected dataset containing 286 samples. However, with this approach, there is no assurance that all fiducial patches are detected correctly for holistic-based approaches or the presence of unique facial features such as mole is not as dominant as other parts of the face.

Since then, significant progress has been made in automatic kinship verification, and a number of approaches have been reported in the literature [3,10,11,13,25,28,31,33,38,43, 45,47,51,61,73,78,81,83–85]. Table 1 presents a summary of the most relevant methods and reported results.

It can be seen that typical current best performing methods follow a similar structure in their methodology, combining several face descriptors, applying metric learning approaches to compute distances between pairs of features and utilizing this distance to learn a threshold that is able to perform a binary classification tasks. Here, we provide a review of the most significative findings of recent research:

Handcrafted features. Handcrafted features designed for facial representation have shown very good performance in different face analysis tasks. This is also the case in kinship verification, where some of the first approaches in the literature were based on low-level handcrafted feature extraction and SVM or KNN classifiers. For instance, Zhou et al. [83] used a spatial pyramid learning descriptor, later refined into a Gabor gradient orientation pyramid [84], an approach also used by Xia et al. [73,74], while Kohli et al. [43] used self-similarity of Weber faces.

Local descriptors based on texture analysis such as variants of HOG [23], LBP [1] or LPQ [2], have also been exploited. The most recent well-performing methods include different descriptors such as Weber local descriptor (WLD) [18], three-patch-based LBPs (TPLBP) [70], over-

Table 1 Summary of the existing methods for kinship verifications

Class	Method	Year	Feature	Classifier	Database	F–S	F–D	M–S	M–D	Mean	Human
Local features	Computational model [30]	2010	Face appearance and geometry	KNN	Cornell KinFace	–	–	–	–	70.7	67.2
	Leveraging [69]	2014	Face appearance and geometry	SVM	Family101	–	–	–	–	92.0	–
	MPDFL [78]	2014	LBP+LE+ SIFT	SVM	KinFaceW-I	73.5	67.5	66.1	73.1	70.1	–
					KinFaceW-II	77.3	74.7	77.8	78.0	77.0	–
					Cornell KinFace	74.8	69.1	77.5	66.1	71.9	–
					UB KinFace	–	–	–	–	67.3	–
	GGOP [84]	2012	Gabor	SVM	1000 images Internet	65.5	65.5	73.5	74.5	69.7	69.7
	Multi-perspective [12]	2015	TPLBP+LPQ +WLD	SVM	KinFaceW-I	85.8	85.3	86.7	87.5	86.3	63.8
					KinFaceW-II	82.2	84.0	81.2	84.8	83.1	66.8
	mRMR [80]	2016	HDLBP	Dissimilarity	TSKinFace	–	–	–	–	89.7	–
Deep learning	CNN [81]	*2015	CNN	CNN	KinFaceW-I	71.8	76.1	84.1	78.0	77.5	–
					KinFaceW-II	81.9	89.4	92.4	89.9	88.4	–
	Deep+shallow [13]	2016	Deep+Spatiotemporal	SVM	UvA-NEMO Smile	88.3	93.1	90.5	91.2	91.0	–
	Deep PI[62]	2017	VGG	SVM	Family in the Wild	–	–	–	–	71.2	–
Metric learning	Ensemble metric learning [66]	2012	SIFT+ PHOG	Kernel sim	VADANA	–	–	–	–	80.2	–
	DMML [77]	2014	LBP+SPLE+ SIFT	SVM	KinFaceW-I	74.5	69.5	69.5	75.5	72.3	–
					KinFaceW-II	78.5	76.5	78.5	79.5	78.3	–
					Cornell KinFace	76.0	70.5	77.5	71.0	73.8	–
	MNRML[48]	2014	LBP+LE+SIFT+TPLBP	SVM	KinFaceW-I	72.5	66.5	66.2	72.0	69.9	71.0
					KinFaceW-II	76.9	74.3	77.4	77.6	76.5	74.0
	LLMMML[37]	2017	LBP, SIFT	SVM	KinFaceW-II	–	–	–	–	80.0	–
	DDMML[46]	2017	LBP, HOG	SVM	KinFaceW-II	–	–	–	–	84.3	–
Other methods	MLCSL [19]	2015	LBP	Cos-Sim	KinFaceW-I	84.5	81.0	81.0	82.6	83.3	–
					KinFaceW-I	88.0	82.4	84.0	82.6	84.2	–
			HOG	Cos-Sim	KinFaceW-I	86.6	80.9	77.1	85.1	82.4	–
					KinFaceW-II	86.8	82.8	84.4	83.2	84.3	–
	RSBM [61]	2015	SIFT	SVM	TSKinFace	83.0	80.5	82.8	81.1	82.0	–

Verification accuracies are reported in %

complete LBP (OCLBP) [9] or Fisher vector faces (FV) [64].

A typical methodology, based on the combination of hand-crafted features, can be seen in the baseline systems used in Kinship verification competitions [45,47]. Based on reference HOG and LBP implementations, faces are described and encoded by dividing each frame into 8×8 non overlapping blocks, extracting nine-dimensional HOG features and multi-scale LBP features from each block. Finally, all the blocks' features are concatenated to form a 2880-dimensional face feature vector.

Learning deep features. Most of the kinship verification work is mainly based on shallow handcrafted features and hence is not associated with the recent significant progress in the machine learning field that suggests the use of deep features. Motivated by the increasing success of deep learning approaches in image representation and classification in general [65] and face recognition in particular [67], Zhang et al. [81] recently proposed a convolutional neural network architecture for face-based kinship verification. The proposed architecture is composed of two convolution max pooling layers followed by a convolution layer and a fully connected layer. A two-way soft max classifier is used as the final layer to train the network. The network takes a pair of RGB face images of different persons as an input, checking the possible kin relations. However, their reported results do not outperform the shallow methods presented in the FG15 kinship competition on the same datasets [45]. The reason behind this may be the scarcity of training data, since deep learning approaches require the availability of enough training samples which is not the case for available face kinship databases. Recently, Boutellaa et al. [13] show that the combination of shallow and deep features can be complementary and indeed improve the results of shallow features even further, while Robinson et al. showed that the same approach of using deep features to describe the face characteristics could be used even in very challenging and complete datasets with reasonable accuracy [62] However, since deep learning approaches require large amounts of training data, they might not be applicable in every situation.

Color information. The conversion of color images into grayscale can simplify the classification process, but at the same time, it eliminates useful characteristics with discriminative power. Kin-related pairs tend to share facial features related to structural or textural information, such as the shapes of eyes, mouth and nose, but also others related to chrominance information such as hair, eyes or skin color. In addition, grayscale conversion can reduce edge information, making it harder to distinguish even textural capabilities. Recent work hints that the study of joint color–texture infor-

mation that utilizes information from three different channels of digital images can better describe the characteristics of kin-related pairs [72].

Feature selection and fusion. The combination of several types of features, exploiting their possible complementarity, seems to show performance advantages in the verification of kin relations. For example, in the last kinship competition [45], all the proposed methods used three or more descriptors while the best performing method employed four different local features (LBP, HOG, OCLBP and Fisher vectors). Since the combination of several features at different scales generates very large feature vectors, dimensionality reduction techniques have been extensively used. Among the most used methods are principal component analysis (PCA), independent component analysis (ICA), variations such as the whitened principal component analysis (WPCA) or learning the weight of each feature component using sparse ℓ_1 regularized logistic regression.

Typical methods for feature selection try to fuse several features, while only keeping the most discriminative features. Wang and Kambhamettu [69] combined texture and facial geometry in a single classification system. They combined a Gaussian mixture model of LBP features and the projection of facial landmarks in the Grassman manifold. The work of Bottino et al. [11] applied feature fusion on four different textural features: LPQ, WLD, TPLBP and FPLBP. For each feature, the difference between vectors of a pair images is computed and normalized and the four features are concatenated forming the pair descriptor. The minimum redundancy maximum relevance (nRMR) algorithm was applied to perform feature selection, and SVM was utilized for classification.

Vote-based feature selection schemes are able to reduce the model parameters [45]. An improvement in the robustness has been demonstrated with the use of different features that employ pyramid features extracted from different scales and orientations [69,84], while the most recent methods focus on the extraction of high-density local features [80].

Metric learning. Complementarily to dimensionality reduction and feature selection, various metric learning approaches have been investigated to tackle the Kinship verification problem, showing major progress in the field. Metric learning aims at automatically learning a similarity measure from data rather than using handcrafted distances. This is in line with the intuition that faces from member of the same family should look similar, but not necessarily the same. As a first attempt, Somanath and Kambhamettu [66] applied ensemble metric learning. The training data are initially clustered using different similarity kernels. Then a final kernel is

learned based on the initial clustering. For each kin relation, the learned kernel ensures that related pairs have a greater similarity than unrelated pairs. Lu et al. [48] learned a distance metric where the face pairs with a kin relation are pulled close and those without a kin relation are pushed away. Recently, Zhou et al. [85] applied ensemble similarity learning for solving the kinship verification problem. They learned an ensemble of sparse bilinear similarity bases from kinship data by minimizing the violation of the kinship constraints between pairs of images and maximizing the diversity of the similarity bases. Yan et al. [77] and Hu et al. [38] learned multiple distance metrics based on various features, by simultaneously maximizing the kinship constraint (pairs with a kinship relation must have a smaller distance than pairs without a kinship relation) and the correlation of different features. Other utilized methods include the triangular similarity metric learning (TSML) [82] or distance metrics learn using side information-based linear discriminant analysis (SILD) [41], while the most recent methods rely on multi-metric learning [37] or deep metric learning [46,68].

Classification methods. If the extracted face features are discriminative enough, the classification can be simply performed with a linear classifier. Cosine similarity distance and thresholding have been used with mixed performance, while K-nearest neighbor classifier that measures the second-order distance of the features to find the nearest class seems to offer better performance [30]. Model-based classification utilizing a biclass support vector machine (SVM) has shown superior performance when the amount of data allows the partition into meaningful splits of training and testing data [28,45].

If the amount of data is sufficient, a classification stage can be learned together with facial descriptions using deep learning methodology and convolutional neural networks [81]. When this is not the case, other non-model-based classification methods such as canonical correlation analysis [19] or online sparse similarity learning [44] have been tried with different results.

Facial dynamics from videos. The role of facial dynamics in kinship verification is mostly unexplored, as most existing work focus on analyzing still facial images instead of video sequences. While most of the published work copes with kinship problem from images, it has not be until recently that kinship verification from videos has been conducted, starting with the work of Dibeklioglu et al. [28]. In this seminal work, the authors combined facial expression dynamics with temporal facial appearance as features and used SVM for classification. The combination of facial dynamics and static features is able to exploit both the textural and temporal information present in face videos, improving the description of kin-related face pairs. Lately,

the apparition of other works incide on the use of more advanced features [13,26] or on metric learning methods [75,76] to improve the results obtained by simple facial dynamics.

2.2.1 Conclusions

Summarizing, the published papers and organized competitions dealing with automatic kinship verification have shown some promising results over the last few years. From the literature it can be extracted that exploiting combination of several complementary features such as texture, color, facial dynamics and deep information offers the best performance. Feature selection, dimensionality reduction and metric learning descriptors utilized before classifiers based on model training can improve the classification scores even further. The recent apparition of deep learning methods and their discriminatory power shows already to be promising and could push the computer accuracy further.

2.3 Comparing human and computer assessment

Until now, only a handful of studies have tried to compare the performance of humans against automatic methods for kinship verification. However, the experimental evaluation present in the literature is mostly done in very controlled conditions, with a small number of human participants all belonging to the same groups and only across a reduced subset of the available data utilized for the automatic assessment.

In the first automatic kinship study, Fang et al. [30] compared an automatic kinship verification method against the human performance. However, they utilized only 16 participants and a small random subset comprising only 20 image pairs. Zhou et al. [84] used a small subset of 100 random pairs presented to 20 participants, all with ages between 20 and 30, a setup repeated by Lu et al. [48].

3 Kinship verification by humans

To assess the performance of humans in kinship verification tasks, we have gathered information by scoring a set of kinship verification pairs using the Amazon Mechanical Turk service (MTurk) crowdsourcing service [7]. MTurk allows to crowdsource different human intelligence tasks (referred to as HITs) to a group of human workers. Each HIT corresponds to the assessment of one pair of images or videos by one person. In our experiments, MTurk workers remain anonymous, since personally identifying information ore demographics were not collected. In total, 304 different individual workers were included in the experiment, and 10 different annotations from different users were made for each pair (HIT).

Fig. 1 Crowdsourced human estimation of kinship using Amazon Mechanical Turk. Overview of one HIT task

The experimental task was designed trying to keep the HITs assigned to workers as simple as possible, in order to reduce unintentional mistakes. The HITs consists on displaying a pair of face images (or videos) with a prompt question asking if the individuals in both images have a kin relation, with the following question: "*Are these two people related (i.e., part of the same family?)*". Along with the question, the workers are presented with two buttons showing "*Yes*" and "*No*" answers (see Fig. 1).

Using this setup, we have collected kinship verification scores for three different datasets. Two datasets, Kinship Face in the Wild I & II [48], including positive and negative pairs, were used in their original form. The entire datasets consist on 3066 image pairs, with a resolution of 64×64 pixels, and positive and negative kinship pairs are equally distributed.

The third dataset, the UvA-Nemo Smile database [27], comprises 550 video pairs, half containing deliberate (posed) smiles and half of them containing genuine (spontaneous) smiles. The videos were presented to the workers using a rescaled resolution of 480×270 pixels. For this database, the assessment scores were collected using two different setups: showing the original videos or showing just the first frame of the video, for a total number of 2200 pairs. Thus, in total we used 5266 unique HITs to Amazon's Mechanical Turk.

A description of the datasets and other available databases can be seen in Appendix A, included as supplemental material.

Typically, crowdsourcing experiments present a subset of the collected data that might be unreliable. To reduce this effect, we implemented several quality assurance mechanisms that have shown to be successful in literature. First, we only allowed workers that had at completed at least 1000 HITs in the platform and had at least 99% or more

of these HITs approved by the requesters. Second, we collected the time taken by the user to answer the question and compared it with the average times. HITs that were completed in an abnormally short amount of time were rejected. Third, we used a common crowdsourcing quality assurance mechanism called gold standard, which entails the creation and inclusion of tasks that have known answers to the requested crowdsourcing job [29]. The inclusion of these pre-labeled questions allowed us to capture the reliability of its workers. Therefore, all answers from workers that performed badly on the pre-labeled tasks can be removed, potentially improving the accuracy of the crowdsourced results [32]. In this case, we added pairs of cartoon and movie characters that are obviously not related (e.g., Bert from Sesame Street and Jabba the Hutt from Star Wars). All contributions from workers that answered these gold standard HITs incorrectly were rejected. Finally, all HITs were answered by exactly 10 different workers in order to provide a more reliable crowd answer to each individual HIT. Ultimately, a total of 55,643 HITs were completed out of which 2983 were rejected and 52,660 were approved. The payment of each approved HIT was 1 cent, for a total cost of $526.66.

As kinship verification is essentially a problem with binary classification, we decided to assign a value of 1 when an MTurk worker identified the pair as having a kin relation and 0 otherwise. To compute the scores utilized for classification, we have simply averaged the answers of the ten workers, obtaining scores from 0 to 1, that represent the "confidence" value of the humans for each individual pair. Since these scores are analogous to the confidence values obtained by automatic machine-based classification methods, this scoring was preferred against other usual ones such as majority voting.

4 Kinship verification by machines

Based on our prior work [13], we propose a hybrid methodology for kinship verification from facial images and videos that exploits the complementarity of deep and shallow features. As illustrated in Fig. 2, our proposed approach consists on five main steps. It starts with detecting, cropping and aligning the face images based on eye coordinates and other facial landmarks. Then, two types of descriptors are extracted: shallow spatiotemporal texture features and deep features.

As spatiotemporal features, we extract local binary patterns (LBP) [1], local phase quantization (LPQ) [2] and binarized statistical image features (BSIF) [42]. These features are all extracted from Three Orthogonal Planes (TOP) of the videos. To take benefit of the multi-resolution representation [17], the three features are extracted at multiple scales, varying their parameters. For the LBP descriptor, the
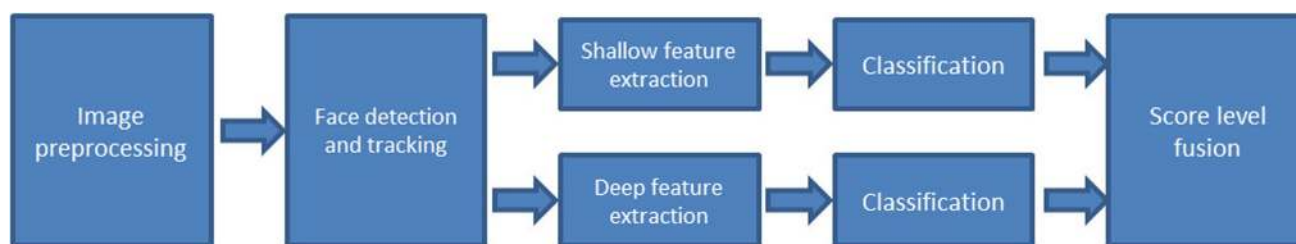
Fig. 2 Overview of the proposed hybrid methodology for automatic kinship verification

selected parameters are $P = \{8, 16, 24\}$ and $R = \{1, 2, 3\}$. For LPQ and BSIF descriptors, the filter sizes were selected as $W = \{3, 5, 7, 9, 11, 13, 15, 17\}$.

Deep features are extracted by convolutional neural networks (CNNs) [55]. In VGG-face, the input of the network is an RGB face image of size 224×224 pixels. To extract deep face features for kinship verification, we input the video frames one by one to the CNN and collect the feature vector issued by the fully connected layer fc7. (All the layers of the CNN except the class predictor fc8 layer and the softmax layer are used.) Finally, all the frames' features of a given face video are averaged, resulting in a video descriptor that can be used for classification.

Two feature pairs corresponding to both components of a kin relationship are then combined. The resulting vector is used as an input to several support vector machines (SVM) for classification. The scores of the classifiers are then fused using a weighted sum. As research in both psychology and computer vision revealed, since different kin relations render different similarity features, all different kin relations are treated differently during the model training.

5 Experimental results and analysis

Following our methodology for human and machine assessment of kinship verification, we have conducted extensive experiments in three datasets: KinFaceW-I, KinfaceW-II and UvA-NEMO Smile, described in Appendix A. The experiments are performed using the methods described in Sects. 4 and 3, following the evaluation protocols recommended in the literature [28,48]. In this context, we have separated the datasets in different kin relations (4 in KinFaceW, 7 in Smile), used cross-validation (fivefold in KinFaceW, leave-one-out in Smile) and utilized still images and video frames (aligned low-resolution facial images in KinFaceW, first video frame in Smile) for the evaluation. We report mean accuracy results measured on the receiver operating characteristic (ROC) curves. Table 2 summarizes the comparative performance of humans and machines from facial images. Figure 3 depicts the ROC curves of human assessment and different automatic verification methods.

Table 2 Comparison of human and machine performance (accuracy) in three datasets: KinFaceW-I and KinFaceW-II and UvA-NEMO Smile

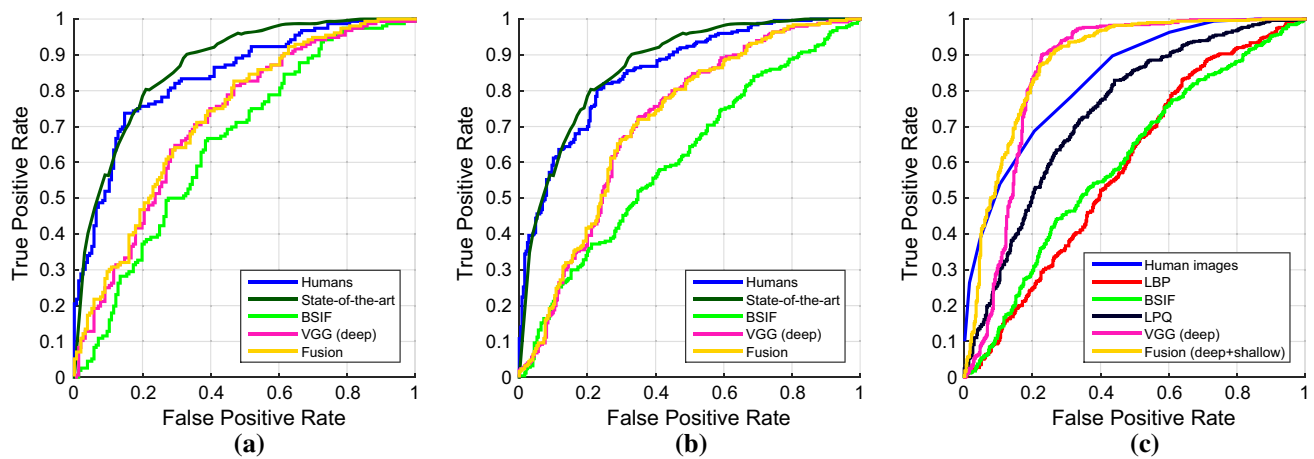| Database	KinFaceW-I	KinFaceW-II	Smile
LBP	62.5	60.9	60.1
LPQ	65.7	67.1	71.7
BSIF	62.3	62.7	63.5
VGG	67.9	64.3	84.7
Deep+Shallow [ours]	68.4	66.5	**87.8**
State of the art	83.7 [45]	86.6 [45]	87.8 [ours]
Humans	78.6	83.5	80.2

Bold represents best result

The experiments show that different automatic methods obtain varied results. On the KinFaceW-I and KinFaceW-II datasets, composed of low-resolution images in uncontrolled environments, the performance of simple textural (LBP, LPQ, BSIF) and deep features (VGG) offers similar results, still far from the performance offered by state-of-the-art methods that use different metric learning and feature fusion techniques [45]. This might be due to the low resolution of the images, which is not a good match for VGG features, and to the inherent bias in the experimental datasets caused by collecting the cropped faces from the same images [10] (see more details in Section 6).

On the other hand, in the UvA-NEMO Smile database, composed of high-resolution video frames taken in controlled conditions, deep features show *state-of-the-art* performance [13]. The results obtained with VGG features can be improved even further by combining them with shallow features using score-level fusion. This shows that the characteristics of the images in the dataset have a noticeable influence in the performance of the automatic methods that are still not able to generalize well across databases.

The experiments assessing the human performance show results that range from 75 to 85%. These results show that humans are still able to outperform many of the most recent automatic methods. However, it can be seen that humans show a slightly worse performance when compared against automatic methods tailored for specific databases and particular conditions.

Fig. 3 Comparing humans versus machine best performing methods on KinFace and Smile databases. **a** KinFaceW-I. **b** KinFaceW-II. **c** SmileDB

Table 3 Classification accuracy (percent) using fivefold validation on the KinFaceW-I dataset

Method	F-S	F-D	M-S	M-D	Mean
LBP	62.5	51.5	61.5	61.5	59.2
LPQ	65.7	69.9	64.4	65.3	66.3
BSIF	62.3	68.0	63.0	60.3	63.4
VGG	67.9	64.6	69.8	64.6	66.7
Deep+Shallow [ours]	68.8	68.8	70.5	65.5	68.4
State of the art [45]	83.0	80.6	82.3	85.0	82.7
Humans	78.2	75.8	74.6	85.8	78.6

Table 4 Classification accuracy (percent) using fivefold validation on the KinFaceW-II dataset

Method	F-S	F-D	M-S	M-D	Mean
LBP	65.4	56.6	60.6	61.0	60.9
LPQ	68.1	70.2	64.3	65.8	67.1
BSIF	63.4	68.7	55.8	63.2	62.7
VGG	65.6	62.6	64.8	64.4	64.3
Deep+Shallow [ours]	66.5	68.8	65.4	65.4	66.5
State of the art [45]	89.4	83.6	86.2	85.0	86.0
Humans	86.0	76.8	84.4	86.6	83.5

5.1 Kinship verification for different relationships

When considering different kin relationships, the performance of humans and machines shows some interesting differences. Tables 3, 4 and 5 summarize the comparative results of automatic methods versus human assessment in separated kin relationships. Figure 4 shows the ROC curves for each different kin relationship. The kin relations are coded as follows: father–son (F-S), father–daughter (F-D), and so on.

The results show that humans are noticeable better when assessing people of the same gender, especially mothers and

daughters. This can be seen across all three databases and is in line with previous psychological studies [58]. However, automatic methods seem to offer similar performance for all types of relationships, although a small tendency to better results can be observed for the father–son relationship.

When considering the age difference, it can be seen that humans show a tendency to assess better brothers and sisters, especially when they are of the same gender, than parents and children. This is expected, since different distinctive features might appear with the age advances. For computers, this difference is not as significant, and no clear trends can be perceived.

5.2 Kinship verification from smile videos

To examine the role of facial dynamics in the assessment of kinship for both humans and machines, we have carried out an experiment that quantifies the difference of verifying kinship relations from videos against still images. For this, we compare the results obtained employing the first frame from each video of the database against the full 10-second videos. Figure 5 shows the ROC curve comparing the performance of videos against still images for the pool of all relationships.

The superiority of the performance obtained with videos compared with still images is observed for both humans and machines (using both shallow and deep features). This clearly demonstrates the importance of face dynamics in verifying kinship between persons. Again, deep features extracted from still face images demonstrate high discriminative ability, outperforming both the spatial texture features extracted from images and the spatiotemporal features extracted from videos.

However, observing carefully, it can be seen that the difference between images and videos seems to be even more significant for computers. For example, using still images, humans seem to have a significantly better performance when

**Table 5** Classification accuracy (percent) using leave-one-out validation on the UvA-NEMO Smile dataset

Method	F-S	F-D	M-S	M-D	B-B	S-S	B-S	Mean
LBP	58.0	65.5	60.3	63.5	57.1	56.4	59.9	60.1
LPQ	60.6	68.1	73.2	71.1	67.9	86.1	75.0	71.7
BSIF	56.4	66.4	64.6	54.9	73.2	65.7	63.6	63.5
VGG	84.0	92.2	80.5	84.6	83.9	89.8	78.0	84.7
Deep+Shallow [ours]	86.6	94.1	85.3	87.0	86.5	92.2	83.3	87.8
Humans	73.7	66.7	71.7	81.5	96.2	88.7	82.8	80.2
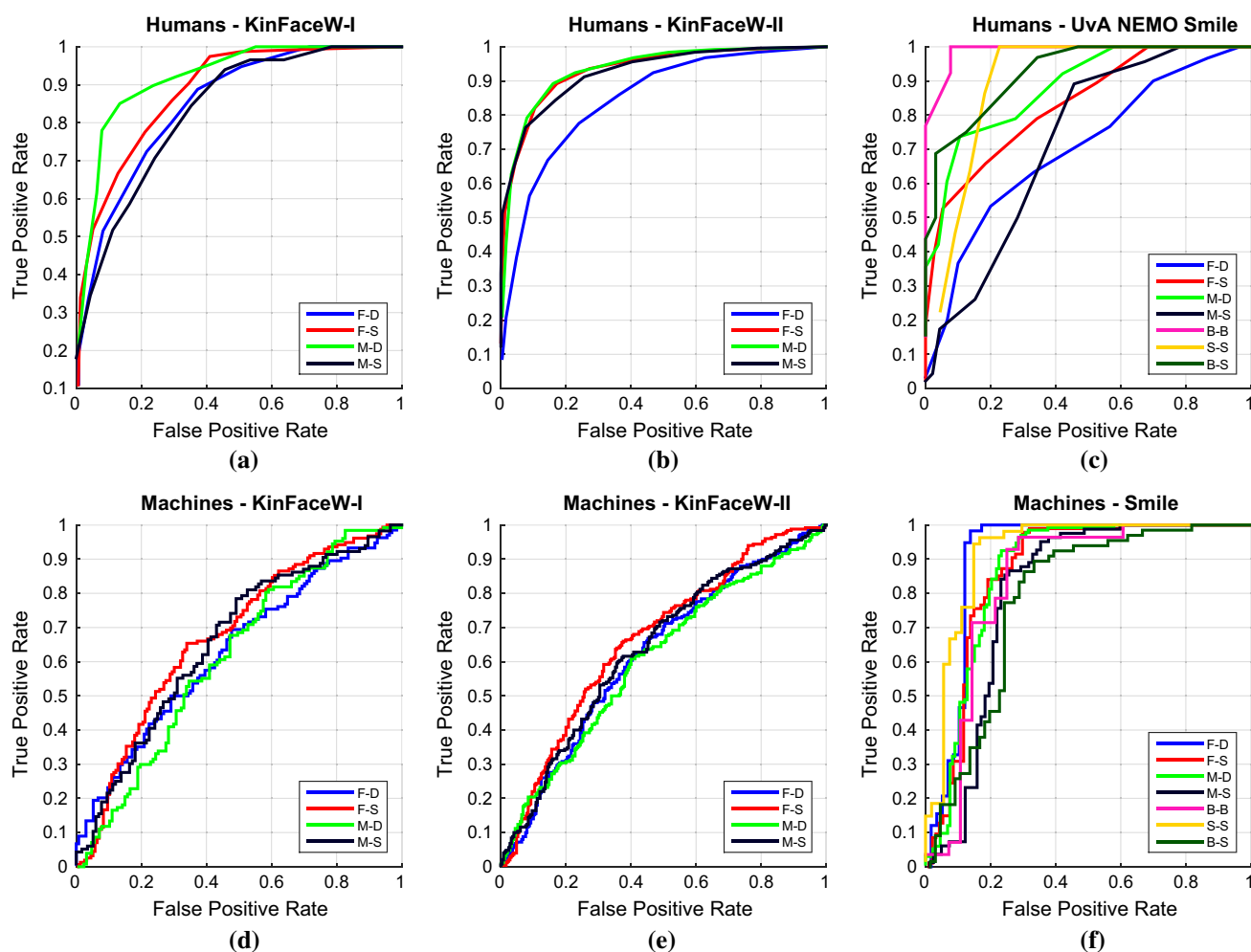
Fig. 4 Human and machine performance across different kin relationships, measured on KinFaceW-I &II and Smile databases. **a** Humans KinFaceW-I. **b** Humans KinFaceW-II. **c** Humans SmileDB. **d** Machine KinFaceW-I. **e** Machine KinFaceW-II. **f** Machine SmileDb

compared against automatic methods based on shallow textural features. However, when we compare the performance obtained using videos, machine methods increase their performance to levels comparable to humans, even when using only shallow spatiotemporal features. If the automatic methods take deep features into account, the machines are able to outperform humans even further. In addition to the importance of spatiotemporal information, these results suggest that increasing the available information for training also has an impact in the performance of the automatic methods.

5.3 Kinship verification from spontaneous and posed smiles

Facial expression seems to have a hereditary component that is able to tie kin-related people. For example, Peleg et al. [56] demonstrated the similarities of spontaneous facial expressions such as smiles between born-blind people and their sighted relatives. To test the influence of spontaneous expression against deliberate or posed ones, we have conducted experiments on the smile database, separating the dataset in
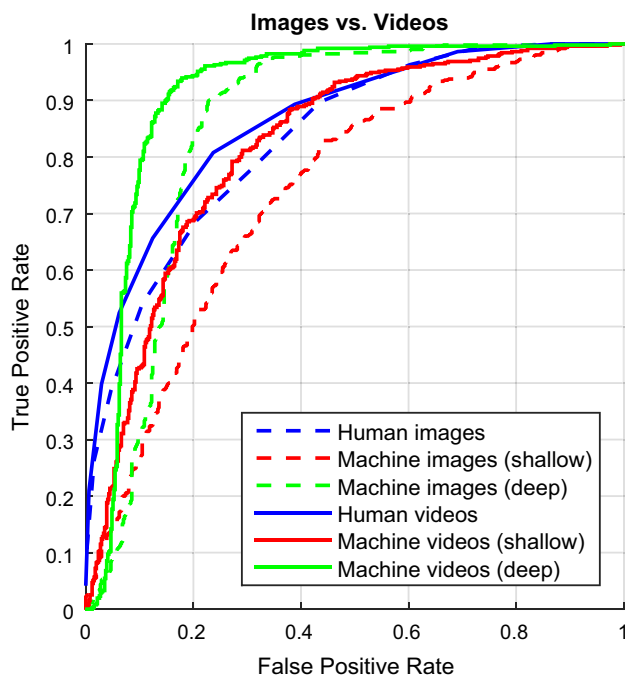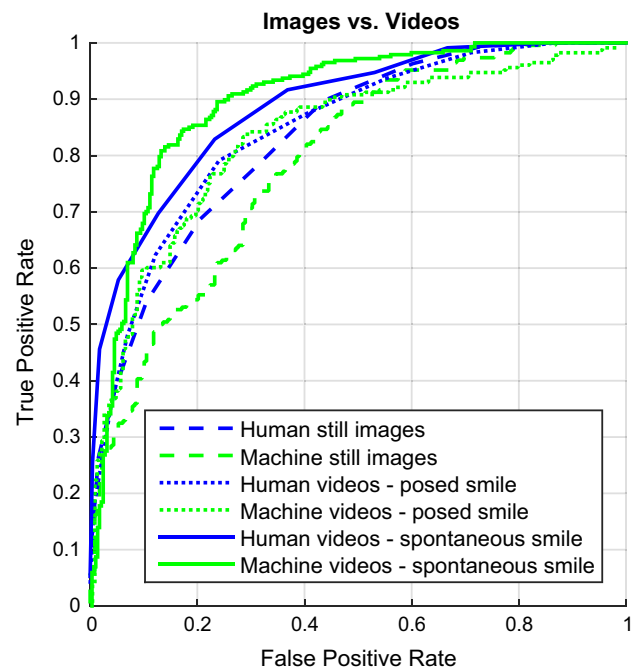
Fig. 5 Performance (accuracy) comparison of humans and machines measured from still images and videos in the UvA-NEMO Smile database

Fig. 6 Performance (accuracy) comparison of humans and machines measured from videos containing posed and spontaneous smiles in the UvA-NEMO Smile database

two different groups. In the first one, the videos show subjects that were instructed to pose with a smile. In the second one, the subjects were exposed to an external stimulus that was able to produce an spontaneous smile. Figure 6 shows ROC curves comparing the performance in kinship verification of humans and machines for both types of pairs, depicting posed and spontaneous smiles.

The results show that both posed and spontaneous smiles provide humans and machines with information that increases their classification performance compared with still images. However, as intuitively expected, spontaneous smiles provide more information than posed ones. This is the case for both humans and machines and can be explained by the learned characteristics of the posed expressions, which make them less specific to particular subjects and their family members.

6 Limitations, open issues and future directions

The comparison of the performance of humans and computers in kinship verification requires the careful design of a set of experiments that guarantee the most possible fairness and the accounting for all types of bias. These experiments usually rely on collections of images compiling a database where positive and negative examples of kin relationships are depicted. Available datasets usually provide images, annotations and verification protocols for separate kin relationships,

such as father–son, father–daughter, mother–son or mother–daughter. With this setup, kinship verification can be viewed as a typical binary classification problem.

However, the particularities of these sets of images can have an effect on the verification accuracy and the exploitation of possible knowledge on the data not related to kinship can lead to biased results. Both computers and humans are able utilize this information together with the kinship-specific features to create a more accurate (but biased) confidence value. In this context, the possible database bias that can be subject to exploitation can be divided in two types: Use of privileged information and use of prior knowledge.

6.1 Use of privileged information

When humans are faced with the task of determining if a pair has a kin relationship, prior knowledge on the nature of these kind of relationships can be used. For example, many kinship verification databases contain images that are easier to classify for humans since they consist of pictures taken from famous people. For a trained human, knowing the identity of both persons greatly simplifies the classification of positive examples. The problem is reduced to the verification of two well-known faces, and no attention has to be paid to kin-related features.

In addition, many negative examples of kinship pairs can be deducted by guessing the relative age, gender or ethnicity of both components of the pair. Pairs that depict proposed
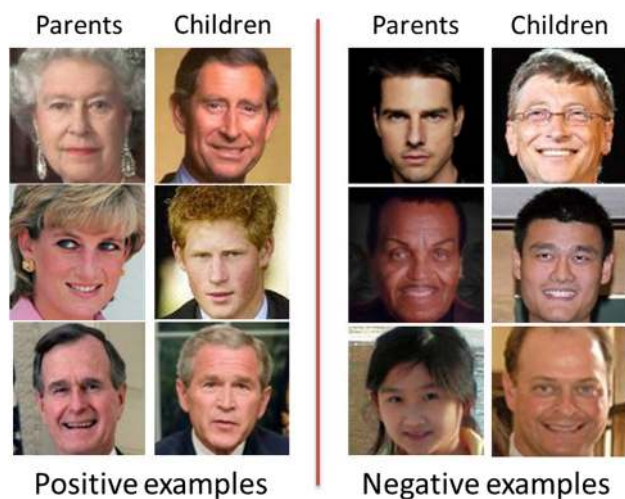
Fig. 7 Examples of image pairs from the *Cornell* dataset. Positive kin pairs can be deducted by humans that know the identity of the famous subjects, without paying attention to kin-related features. Negative examples can be easily discarded for differences in age or ethnicity or known identity

Table 6 Classification accuracy (percent) using age differences on the UvA–Smile dataset and comparison against visual features

Method	Mean
Age difference	70.8
Visual (deep + shallow)	90.9
Fusion (visual + age)	**92.2**

Bold represents best result

Fig. 8 Examples of video frames from the *Smile* dataset. All videos are obtained under controlled conditions, showing similar very similar characteristics

parents and children of very different ethnicities or children noticeably older than the parents can be easily discarded by a human. In case of annotated databases that include the age or ethnicity of the subjects, machines can also exploit the information by including it as a feature for the classification task.

An example of this source of bias can be seen in Figure 7 that shows positive and negative kinship pair examples obtained from the Cornell databases. The well-known identity of the subjects simplifies the human classification of positive examples. The discrepancies in the expected age and ethnicity of parents and children simplify the classification of negative examples.

When utilizing annotated databases for training, the same classification strategy utilized by the humans, inferring biometric traits to aid the classification, could be as well exploited by carefully designed automated systems. Automatic verification methods that try to assess a set of biometric traits such as the age and ethnicity of the subjects could improve a kinship verification system where the same data is used during training, making use of all available priors.

To illustrate the possible use of privileged information in automatic verification and its effect, we have performed classification tasks in the UvA–Smile database utilizing only the provided ages of the subjects in the videos. Calculating the age difference between the members of the pairs, we utilize this result to train a set of SVM classifiers that try to estimate if both members of the pair are kin-related based solely on their ages. Table 6 shows the classification result of this strategy compared with the automatic classification based on visual features.

The results show that, while not being a definitive classification strategy, the age of the components of a potential kin-related couple is indeed of importance in the estimation of kinship. In addition, in many cases, the fusion of the classification scores obtained by age difference alone can improve the results of visual classification even further.

6.2 Use of prior knowledge

When exposed to significant amount of training data belonging to a particular dataset, both humans and computers can take advantage of the limitations on the image capturing conditions of the dataset. In many datasets related to face analysis, the capturing conditions of the images are similar, utilizing the same camera, pose and illumination, a fact that simplifies greatly the problems for computers, while not posing a great advantage for humans. For example, some databases capture images or videos in a restricted environment, where the background stays the same and the subjects are assured to remain with a frontal pose and unoccluded, while the image quality and resolution are constant and very high. In this context, the automatic classification problem is simplified by the invariant conditions. Figure 8 shows two images of the Smile database, showing the constant capturing conditions and a color palette to help in the luminance and color normalization and other preprocessing tasks.

To simulate real-world conditions, other existing datasets are obtained under uncontrolled environments with no restrictions in terms of resolution, pose, lighting, background, expression, age, ethnicity or partial occlusion. These datasets expose a more difficult challenge for the computers, since they require the utilization of robust features that are able to cope with the variation in the conditions, while

Fig. 9 Examples of image pairs from the KinFaceW and UBI-Kin datasets. Images are taken in very different conditions regarding pose, background and occlusion and depict very different ages, ethnicities and expressions

Fig. 10 Examples of six image pairs (3 on the top row and 3 on the bottom row) obtained from the *KinFaceW-II* dataset. Kinship pairs are cropped from the same image and show very similar characteristics

Table 7 Classification accuracy (percent) of different methods on the different subsets of KinFaceW-II dataset, including biased simple scoring

Method	F-S	F-D	M-S	M-D	Mean
LBP [48][1]	75.4	66.6	70.6	66.0	69.6
HOG [48][1]	74.2	66.6	70.6	67.0	69.6
NRML$_{LBP}$ [48][2]	79.2	71.6	72.2	68.4	72.8
BIU$_{HOG}$ [45][2]	87.5	80.8	79.8	75.6	80.9
Polito [45][3]	84.0	82.2	84.8	81.2	83.1
LIRIS [45][3]	**89.4**	**83.6**	**86.2**	85.0	**86.0**
Simple scoring[1,2,3]	78.2	73.2	84.2	**88.2**	80.1

Bold represents best result
[1] *Unsupervised*, [2] *Image-Unrestricted*, [3] *Image-Restricted*

Table 8 Classification accuracy (percent) of simple scoring methods on the different subsets of *KinFaceW-II* dataset

Method	F-S	F-D	M-S	M-D	Mean
SSIM	65.8	63.8	67.4	67.4	66.1
RGB distance	72.2	71.8	75.0	74.6	73.4
Luminance distance	68.4	68.2	68.8	70.4	68.9
Chrominance distance	78.2	73.2	84.2	88.2	80.1

humans are able to overcome this variability easier. Figure 9 shows examples from KinFaceW-I and UBI-Kin datasets, depicting images captured under very different conditions.

On the other hand, the selection of the face images that compose the kinship relationship database can also cause bias. For example, the KinFaceW and TSKinFace datasets contain face images collected from the internet depicting four classes of family relationships, including images obtained under uncontrolled and unrestricted environments. However, both images in a positive kin pair are cropped from the same original photographs, a fact that is usually not mentioned in research articles, and which implications are rarely discussed. As an example of this source of bias, Figure 10 shows images taken from the KinFaceW-II dataset where both images of the kinship pair are cropped from the same image.

6.2.1 Experiments on database bias

We have conducted a set of experiments that take advantage of the prior knowledge of the image capturing conditions to obtain competitive but biased results in kinship verification tasks. As expected, knowing the image characteristics such as that both images of a positive pair were cropped from the same image can significantly bias and simplify the classification problem. A classification strategy that tries to determine whether both images in a pair are cropped from the same photograph will show improvements when compared to approaches focusing only on facial features. To illustrate this anomaly, an extremely simple classification method that

requires no training and offers comparable results to the ones obtained with sophisticated methods under the same experimental protocol was presented [10].

In this method, measuring the chrominance distance between the images of a kinship pair produces directly the classification score. Image pairs with smaller distance are more likely to be part of the same photography and hence more likely to be a positive kinship pair. Table 7 shows our obtained results using the simple scoring approach on *Kin-FaceW & II* dataset under the three evaluation protocols. A comparison against other reported methods under the same evaluation protocols is also reported. It can be seen that a simple strategy that focuses solely on the source of bias produces results comparable with sophisticated methods. In addition, this simple method has the capability of being complementary to any other method that focuses only in pure kinship-related features.

Besides the difference in average chrominance, any other method that is able to take into account the characteristics of the captured images is suitable to present discrimination power and can be utilized to compute the simple scoring. Table 8 summarizes the scores obtained on the KinFaceW-II dataset utilizing different methods not related to kinship such as Structural Similarity Measurements (SSIM), distance in the RGB color space, luminance distance in the CIELAB color space and the distance of the chrominance in the Lab color space.

To illustrate that image similarity classification only works on databases with pairs cropped from the same image, we

Table 9 Mean classification accuracy (percent) of the simple scoring method on different databases

Database	Simple Scoring	State of the art	Cropped
UBKinFace	52.2	67.3 [78]	No
Smile	57.7	91.0 [13]	No
KinFaceW-I	71.4	86.3 [12]	Partially
KinFaceW-II	80.1	88.4 [81]	All
TSKinFace	80.2	82.0 [61]	All
CornellKin	**81.4**	73.8 [77]	All

Bold represents best result

Table 10 Cross-database performance using VGG features. Classification accuracy training in one database and testing in another

Training/testing	KFW-I	KFW-II	Smile	TSKin
KFW-I	**66.7**	57.3	59.2	62.0
KFW-II	61.6	**63.8**	59.0	59.8
Smile	53.5	52.3	**88.6**	52.3
TSKin	66.0	61.0	57.6	**66.3**

have used the same method across most of the available databases. Table 9 depicts the results of the experiment. As expected, the classification results based on chrominance difference show diminished discrimination power when no assumption can be made on the image pairs. That is the case of datasets such as UvA-NEMO Smile or UBKinFace.

A possible implication of the limitations in the image capturing processes is that even if the design of the kinship verification methodology does not explicitly target the image capturing conditions as discrimination features, many learning-based methods applied in these datasets could inadvertently be learning features not related to kinship but to the very nature of the images and their conditions, such as background, resolution, luminance or average color. In this context, to minimize the learning bias, we believe that future methods that report results in kinship analysis should be verified and evaluated on several different publicly available datasets, even considering the utilization of cross-database verification strategies.

6.3 Discussion

As seen in the experiments reported above, because of the nature of many of the kinship datasets, there is a high potential for biased results and confusing interpretations when comparing different kinship verification methods. It is highly recommendable that the publications reporting kinship verification results disclose all possible sources of bias in their datasets and the possible implications of them on the reported performance.

In addition, most of the existing visual kinship datasets used for kinship verification purposes contain a relatively small number of images, usually well below 1000 total training image pairs. This number is reduced further if we consider that most of the experiments are performed separately in many different family relationships. The lack of sufficient data results in models that are prone to overfitting in the training data. The generalization capabilities to unseen data, captured in different conditions than the one utilized for training, could potentially lead to unstable predictions.

Table 10 shows the classification accuracies obtained for cross-database performance. The results are obtained using deep features (VGG), training in one database and testing in the other. The diagonal (in bold) shows intra-database results obtained using fivefold validation.

As expected, given the similarity of the data between KinFaceW and TSKin datasets, the results show that the models obtained generalize relatively well, although they offer a slightly lower performance. However, the model trained with the Smile database, which is captured in controlled conditions, is not able to offer high classification accuracies in other databases and viceversa. This suggests that data variability during the training of kinship verification models is indeed of importance.

7 Conclusions

This article focuses on the comparison of human and machine performance on the task of kinship verification. The study of different types of dataset bias and their effects on the experimental accuracy complements the evaluation and offers a guideline for the conduct of future studies.

From the human perspective, psychological studies show that recognizing family members of different subjects is an ability based on facial similarity. Humans are able to guess with certain probability above chance if a pair of persons are part of the same family. Automatic kinship verification methods have helped machines to attain this ability by checking the similarity of features obtained from facial images and videos.

In our work, the human ability to assess kinship has been evaluated using a crowdsourced approach based on the Amazon Mechanical Turk service. We have extensively studied the state of the art in automatic kinship verification methods. We compared the human assessment with a method that combines both shallow textural features and deep features obtained using deep learning. Experiments for both humans and machines were conducted in three whole datasets (NEMO Smile, KinFaceW-I and II). Human and machine results have been compared in a meaningful way, over the same data showing some interesting insights.

From the computer perspective, the ability of machines in kinship verification seems to have surpassed the human

ability. Humans show improved ability when comparing subjects of same sex and similar age, while machines seem to be able to assess all relations equally. Spatiotemporal information obtained from video sequences of kin-related subjects is shown to be of vital importance in kinship verification, while the use of spontaneous expressions as opposed to posed ones facilitates the kinship assessment even further.

However, the machine capabilities are closely related to the training material utilized in the benchmark databases. While humans are able to perform similarly in all types of conditions, the machines' performance is tied to the particular characteristics of the used dataset and automatic kinship verification methods do not generalize always well. Current kinship datasets have undeniably contributed to the kinship verification research to some extent. However, we discussed that the nature of the images in these datasets have a high potential for biased results. This calls the research community for joint efforts to design new and more reliable databases that pay careful attention to the possible sources of bias, trying to minimize them.

The exploitation of deep learning methods when constructing kinship verification models have been proven useful, providing results that already surpass human ability. The use of more sophisticated network architectures (e.g., ResNet [35]) and more complex databases that help modeling faces (e.g., CASIA-WebFace [79]) could lead to improved results in kinship verification from images.

The recent apparition of new larger datasets [62], including new methodologies based on it [68] and a workshop that provided extensive evaluation [63], can already provide for the opportunity of performing extensive cross-database experiments that assess the portability and generalization of the models.

However, future directions for kinship verification should also take into account more diverse types of data. Until now, the lack of large databases with different modalities (e.g., images and videos) and annotated with multiple traits (e.g., ethnicity, age, and kinship) has been hindering the progress of the field and could constitute the bulk of the future efforts on the research field. These types of datasets would not only improve the robustness of automatic methods, but will in turn increase the accuracy of the assessment of the human abilities, especially if the possible relationships between subfields of facial analysis and soft biometrics can be taken into account.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
2. Ahonen, T., Rahtu, E., Ojansivu, V., Heikkila, J.: Recognition of blurred faces using local phase quantization. In: International Conference on Pattern Recognition, pp. 1–4 (2008)
3. Alirezazadeh, P., Fathi, A., Abdali-Mohammadi, F.: A genetic algorithm-based feature selection for kinship verification. Signal Processing Letters, IEEE **22**(12), 2459–2463 (2015). https://doi.org/10.1109/LSP.2015.2490805
4. Alvergne, A., Faurie, C., Raymond, M.: Differential facial resemblance of young children to their parents: who do children look like more? Evolution and Human behavior **28**(2), 135–144 (2007)
5. Alvergne, A., Oda, R., Faurie, C., Matsumoto-Oda, A., Durand, V., Raymond, M.: Cross-cultural perceptions of facial resemblance between kin. Journal of Vision **9**(6), 23–23 (2009)
6. Alvergne, A., Perreau, F., Mazur, A., Mueller, U., Raymond, M.: Identification of visual paternity cues in humans. Biology letters **10**(4), 20140,063 (2014)
7. Amazon: Amazon mechanical turk. (2016). URL https://www.mturk.com/
8. Anastasi, J.S., Rhodes, M.G.: An own-age bias in face recognition for children and older adults. Psychonomic Bulletin & Review **12**(6), 1043–1047 (2005). https://doi.org/10.3758/BF03206441
9. Barkan, O., Weill, J., Wolf, L., Aronowitz, H.: Fast high dimensional vector multiplication face recognition. In: IEEE International Conference on Computer Vision, pp. 1960–1967 (2013)
10. Bordallo López, M., Boutellaa, E., Hadid, A.: Comments on the "kinship face in the wild" data sets. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2016). https://doi.org/10.1109/TPAMI.2016.2522416
11. Bottino, A.G., De Simone, M., Laurentini, A., Vieira, T.: A new problem in face image analysis: finding kinship clues for siblings pairs (2012)
12. Bottino, A.G., Ul Islam, I., Vieira, T.: A multi-perspective holistic approach to kinship verification in the wild (2015)
13. Boutellaa, E., López, B., M. Ait-Aoudia, S., Feng, X., Hadid, A.: Kinship verification from videos using texture spatio-temporal features and deep learning features. In: International Conference on Biometrics (ICB'16) (2016)
14. Brédart, S., French, R.M.: Do babies resemble their fathers more than their mothers? a failure to replicate christenfeld and hill (1995). Evolution and Human Behavior **20**(2), 129–135 (1999)
15. Bressan, P., Grassi, M.: Parental resemblance in 1-year-olds and the gaussian curve. Evolution and Human Behavior **25**(3), 133–141 (2004)
16. Burch, R.L., Gallup, G.G.: Perceptions of paternal resemblance predict family violence. Evolution and Human Behavior **21**(6), 429–435 (2000)
17. Chan, C.H., Tahir, M., Kittler, J., Pietikainen, M.: Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(5), 1164–1177 (2013)
18. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: Wld: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1705–1720 (2010)
19. Chen, X., An, L., Yang, S., Wu, W.: Kinship verification in multilinear coherent spaces. Multimedia Tools and Applications pp. 1–18 (2015)

20. Christenfeld, N., Hill, E.A., et al.: Whose baby are you. Nature **378**(6558), 669–669 (1995)

21. Dal Martello, M.F., Maloney, L.T.: Where are kin recognition signals in the human face? Journal of Vision **6**(12), 2–2 (2006)

22. Dal Martello, M.F., Maloney, L.T.: Lateralization of kin recognition signals in the human face. Journal of vision **10**(8), 9 (2010)

23. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 vol. 1 (2005)

24. DeBruine, L.M., Smith, F.G., Jones, B.C., Roberts, S.C., Petrie, M., Spector, T.D.: Kin recognition signals in adult faces. Vision research **49**(1), 38–43 (2009)

25. Dehghan, A., Ortiz, E., Villegas, R., Shah, M.: Who do i look like? determining parent-offspring resemblance via gated autoencoders. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 1757–1764 (2014). https://doi.org/10.1109/CVPR.2014.227

26. Dibeklioglu, H.: Visual transformation aided contrastive learning for video-based kinship verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2459–2468 (2017)

27. Dibeklioglu, H., Salah, A., Gevers, T.: Are you really smiling at me? spontaneous versus posed enjoyment smiles. In: Computer Vision ECCV 2012, vol. 7574, pp. 525–538. Springer (2012). https://doi.org/10.1007/978-3-642-33712-3_38

28. Dibeklioglu, H., Salah, A., Gevers, T.: Like father, like son: Facial expression dynamics for kinship verification. In: Computer Vision (ICCV), 2013 IEEE International Conference on, pp. 1497–1504 (2013)

29. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2399–2402. ACM (2010)

30. Fang, R., Tang, K.D., Snavely, N., Chen, T.: Towards computational models of kinship verification. In: Image Processing (ICIP), 2010 17th IEEE International Conference on, pp. 1577–1580. IEEE (2010)

31. Ghahramani, M., Yau, W.Y., Teoh, E.K.: Family verification based on similarity of individual family member's facial segments. Machine Vision and Applications **25**(4), 919–930 (2014)

32. Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., Kostakos, V.: Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. Computer Networks **90**, 34–48 (2015)

33. Guo, G., Wang, X.: Kinship measurement on salient facial features. Instrumentation and Measurement, IEEE Transactions on **61**(8), 2322–2325 (2012). https://doi.org/10.1109/TIM.2012.2187468

34. Hamilton, W.: The genetical evolution of social behaviour. (1964)

35. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, pp. 630–645. Springer (2016)

36. Hills, P.J., Lewis, M.B.: The own-age face recognition bias in children and adults. The Quarterly Journal of Experimental Psychology **64**(1), 17–23 (2011)

37. Hu, J., Lu, J., Tan, Y.P., Yuan, J., Zhou, J.: Local large-margin multi-metric learning for face and kinship verification. IEEE Transactions on Circuits and Systems for Video Technology (2017)

38. Hu, J., Lu, J., Yuan, J., Tan, Y.P.: Large margin multi-metric learning for face and kinship verification in the wild. In: Computer Vision–ACCV 2014, pp. 252–267. Springer (2015)

39. Kaminski, G., Gentaz, E., Mazens, K.: Development of children's ability to detect kinship through facial resemblance. Animal cognition **15**(3), 421–427 (2012)

40. Kaminski, G., Méary, D., Mermillod, M., Gentaz, E.: Perceptual factors affecting the ability to assess facial resemblance between

41. parents and newborns in humans. Perception **39**(6), 807–818 (2010)

42. Kan, M., Shan, S., Xu, D., Chen, X.: Side-information based linear discriminant analysis for face recognition. In: British Machine Vision Conference, pp. 1–12 (2011)

43. Kannala, J., Rahtu, E.: BSIF: Binarized statistical image features. In: International Conference on Pattern Recognition (ICPR), pp. 1363–1366 (2012)

44. Kohli, N., Singh, R., Vatsa, M.: Self-similarity representation of weber faces for kinship classification. In: Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on, pp. 245–250 (2012)

45. Kou, L., Zhou, X., Xu, M., Shang, Y.: Learning a genetic measure for kinship verification using facial images. Mathematical Problems in Engineering **2015** (2015)

46. Lu, J., Hu, J., Liong, V.E., Zhou, X., Bottino, A., Islam, I.U., Vieira, T.F., Qin, X., Tan, X., Keller, Y., et al.: The fg 2015 kinship verification in the wild evaluation. Face and Gesture (FG'15) (2015)

47. Lu, J., Hu, J., Tan, Y.P.: Discriminative deep metric learning for face and kinship verification. IEEE Transactions on Image Processing **26**(9), 4269–4282 (2017)

48. Lu, J., Hu, J., Zhou, X., Zhou, J., Castrillón-Santana, M., Lorenzo-Navarro, J., Kou, L., Shang, Y., Bottino, A., Vieira, T.F.: Kinship verification in the wild: The first kinship verification competition. In: Biometrics (IJCB), 2014 IEEE International Joint Conference on, pp. 1–6. IEEE (2014)

49. Lu, J., Zhou, X., Tan, Y.P., Shang, Y., Zhou, J.: Neighborhood repulsed metric learning for kinship verification. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(2), 331–345 (2014)

50. Maloney, L.T., Dal Martello, M.F.: Kin recognition and the perceived facial similarity of children. Journal of Vision **6**(10), 4–4 (2006)

51. Mateo, J.M.: Perspectives: Hamilton's legacy: Mechanisms of kin recognition in humans. Ethology **121**(5), 419–427 (2015)

52. Mathews, S., Kambhamettu, C., Barner, K.: Am i your sibling?; inferring kinship cues from facial image pairs. In: Information Sciences and Systems (CISS), 2015 49th Annual Conference on, pp. 1–5 (2015). https://doi.org/10.1109/CISS.2015.7086888

53. McLain, D.K., Setters, D., Moulton, M.P., Pratt, A.E.: Ascription of resemblance of newborns by parents and nonrelatives. Evolution and Human behavior **21**(1), 11–23 (2000)

54. Meissner, C.A., Brigham, J.C.: Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy, and Law **7**(1), 3 (2001)

55. Nesse, R.M., Silverman, A., Bortz, A.: Sex differences in ability to recognize family resemblance. Ethology and Sociobiology **11**(1), 11–21 (1990). https://doi.org/10.1016/0162-3095(90)90003-O

56. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)

57. Peleg, G., Katzir, G., Peleg, O., Kamara, M., Brodsky, L., Hel-Or, H., Keren, D., Nevo, E.: Hereditary family signature of facial expression. Proceedings of the National Academy of Sciences **103**(43), 15921–15926 (2006)

58. Platek, S.M., Burch, R.L., Panyavin, I.S., Wasserman, B.H., Gallup, G.G.: Reactions to children's face: Resemblance affects males more than females. Evolution and Human Behavior **23**(3), 159–166 (2002)

59. Platek, S.M., Keenan, J.P., Mohamed, F.B.: Sex differences in the neural correlates of child facial resemblance: an event-related fmri study. NeuroImage **25**(4), 1336–1344 (2005)

60. Platek, S.M., Kemp, S.M.: Is family special to the brain? an event-related fmri study of familiar, familial, and self-face recognition. Neuropsychologia **47**(3), 849–858 (2009)

61. Porter, R.H., Cernoch, J.M., Balogh, R.D.: Recognition of neonates by facial-visual characteristics. Pediatrics **74**(4), 501–504 (1984)

61. Qin, X., Tan, X., Chen, S.: Tri-subject kinship verification: understanding the core of a family. Multimedia, IEEE Transactions on **17**, 1855–1867 (2015)

62. Robinson, J.P., Shao, M., Wu, Y., Fu, Y.: Families in the wild (fiw): Large-scale kinship image database and benchmarks. In: Proceedings of the 2016 ACM on Multimedia Conference, MM '16, pp. 242–246. ACM, New York, NY, USA (2016)

63. Robinson, J.P., Shao, M., Zhao, H., Wu, Y., Gillis, T., Fu, Y.: Recognizing families in the wild (rfiw): Data challenge workshop in conjunction with acm mm 2017. In: Proceedings of the 2017 Workshop on Recognizing Families In the Wild, RFIW '17, pp. 5–12. ACM, New York, NY, USA (2017)

64. Simonyan, K., Parkhi, O., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: Proceedings of the British Machine Vision Conference. BMVA Press (2013)

65. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014). URL http://arxiv.org/abs/1409.1556

66. Somanath, G., Kambhamettu, C.: Can faces verify blood-relations? In: Biometrics: Theory, Applications and Systems (BTAS), International Conference on, pp. 105–112. IEEE (2012)

67. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition, pp. 1891–1898. IEEE, Washington, DC, USA (2014). https://doi.org/10.1109/CVPR.2014.244

68. Wang, S., Robinson, J.P., Fu, Y.: Kinship verification on families in the wild with marginalized denoising metric learning. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 216–221 (2017)

69. Wang, X., Kambhamettu, C.: Leveraging appearance and geometry for kinship verification. In: Image Processing (ICIP), 2014 IEEE International Conference on, pp. 5017–5021. IEEE (2014)

70. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Real-Life Images workshop at the European Conference on Computer Vision (2008)

71. Wu, H., Yang, S., Sun, S., Liu, C., Luo, Y.J.: The male advantage in child facial resemblance detection: Behavioral and erp evidence. Social neuroscience **8**(6), 555–567 (2013)

72. Wu, X., Boutellaa, E., Lpez, M.B., Feng, X., Hadid, A.: On the usefulness of color for kinship verification from face images. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2016)

73. Xia, S., Shao, M., Fu, Y.: Toward kinship verification using visual attributes. In: Pattern Recognition (ICPR), 2012 21st International Conference on, pp. 549–552. IEEE (2012)

74. Xia, S., Shao, M., Luo, J., Fu, Y.: Understanding kin relationships in a photo. Multimedia, IEEE Transactions on **14**(4), 1046–1056 (2012)

75. Yan, H., Hu, J.: Video-based kinship verification using distance metric learning. Pattern Recognition **75**, 15–24 (2018)

76. Yan, H., Lu, J.: Video-based facial kinship verification. In: Facial Kinship Verification, pp. 63–80. Springer (2017)

77. Yan, H., Lu, J., Deng, W., Zhou, X.: Discriminative multimetric learning for kinship verification. Information Forensics and Security, IEEE Transactions on **9**(7), 1169–1178 (2014)

78. Yan, H.C., Lu, J., Zhou, X.: Prototype-based discriminative feature learning for kinship verification (2014)

79. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)

80. Zhang, J., Xia, S., Pan, H., Qin, A.: A genetics-motivated unsupervised model for tri-subject kinship verification. In: Image Processing (ICIP), 2016 IEEE International Conference on, pp. 2916–2920. IEEE (2016)

81. Zhang, K., Huang, Y., Song, C., Wu, H., Wang, L.: Kinship verification with deep convolutional neural networks. In: British Machine Vision Conference (BMVC), pp. 148.1–148.12 (2015)

82. Zheng, L., Idrissi, K., Garcia, C., Duffner, S., Baskurt, A. (eds.): Triangular Similarity Metric Learning for Face Verification. IEEE (2015)

83. Zhou, X., Hu, J., Lu, J., Shang, Y., Guan, Y.: Kinship verification from facial images under uncontrolled conditions. International Conference on Multimedia, MM '11, pp. 953–956. ACM, New York, NY, USA (2011)

84. Zhou, X., Lu, J., Hu, J., Shang, Y.: Gabor-based gradient orientation pyramid for kinship verification under uncontrolled environments. In: Proc. 20th ACM international conference on Multimedia, pp. 725–728 (2012)

85. Zhou, X., Shang, Y., Yan, H., Guo, G.: Ensemble similarity learning for kinship verification from facial images in the wild. Information Fusion pp. – (2015)

Miguel Bordallo Lopez received his Master's and Doctoral degrees from the University of Oulu in 2010 and 2014, respectively, where he currently works at the Center for Machine Vision and Signal Analysis as a Research Scientist. His research interests include real-time face analysis and computer vision in embedded platforms.

Abdenour Hadid is an Associate Professor at the Center for Machine Vision and Signal Analysis at the University of Oulu. His research interests include computer vision, machine learning, and pattern recognition with a particular focus on biometrics.

Elhocine Boutellaa received his Master's and Doctoral degrees from Ecole nationale superieure d'Informatique, Algeria, in 2011 and 2017, respectively. He has been working as a research associate at the CDTA, Algeria, from 2011 to 2017. His research interests include face analysis and biometrics.

Vassilis Kostakos is a professor of computer science at the School of Computing and Information Systems, University of Melbourne, Australia. His research interests include ubiquitous computing, HCI, and social systems. Kostakos received a PhD in computer science from the University of Bath, UK.
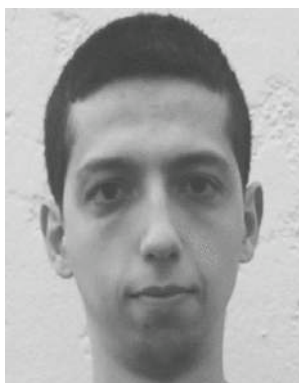
Jorge Goncalves is a lecturer at the University of Melbourne, Australia. His research interests include ubiquitous computing, HCI, crowdsourcing, public displays, and social computing. Goncalves received a PhD in computer science from the University of Oulu.

Simo Hosio is a researcher in the Center for Ubiquitous Computing at the University of Oulu, Finland. His research interests include social computing, crowdsourcing, and public displays. Hosio received a PhD in computer science from the University of Oulu, Finland.