Augmenting Automated Kinship Verification with Targeted Human Input

Completed Research Paper

Danula Hettiachchi
The University of Melbourne
danula.hettiachchi@unimelb.edu.au

Niels van Berkel
Aalborg University
nielsvanberkel@cs.aau.dk

Simo Hosio
University of Oulu
simo.hosio@oulu.fi

Miguel Bordallo López
VTT Research Centre of Finland
miguel.bordallo@vtt.fi

Vassilis Kostakos
The University of Melbourne
vassilis.kostakos@unimelb.edu.au

Jorge Goncalves
The University of Melbourne
jorge.goncalves@unimelb.edu.au

Abstract

Kinship verification is the problem whereby a third party determines whether two people are related. Despite previous research in Psychology and Machine Vision, the factors affecting a person's verification ability are poorly understood. Through an online crowdsourcing study, we investigate the impact of gender, race and medium type (image vs video) on kinship verification - taking into account the demographics of both raters and ratees. A total of 325 workers completed over 50,000 kinship verification tasks consisting of pairs of faces shown in images and videos from three widely used datasets. Our results identify an own-race bias and a higher verification accuracy for same-gender image pairs than opposite-gender image pairs. Our results demonstrate that humans can still outperform current state-of-the-art automated unsupervised approaches. Furthermore, we show that humans perform better when presented with videos instead of still images. Our findings contribute to the design of future human-in-the-loop kinship verification tasks, including time-critical use cases such as identifying missing persons.

Keywords: Kinship verification, worker characteristics, crowdsourcing

Introduction

Humans have the ability to recognise and verify family relationships (kinship) using cues such as facial resemblance (Alvergne et al. 2007; Bressan and Kramer 2015; Kaminski et al. 2009). The ability to recognise one's own relatives is known as *kinship recognition*, while the process of validating the kin relationship between two individuals through a certain method or a third unrelated individual is called *kinship verification*. Kinship verification in particular has several important potential applications. For example, when working with large volumes of multimedia (*e.g.*, as seen in social media platforms), kinship verification or estimation can be used to organise photos or videos into collections. Furthermore, kinship verification can be used to generate family trees using historical photos. Other than these classification tasks, kinship verification can be highly effective in emergency situations that involve, for example, missing children or seniors (Xia et al. 2011; Zhou et al. 2012).

Given the value of accurate identification of family relationships, an active research area called *automatic kinship verification* has emerged within the machine vision community (Fang et al. 2010). Recent developments in this area have produced a number of notable kinship verifying classifiers, with better overall accuracy than humans (Fang et al. 2010; Wu et al. 2019; Zhou et al. 2012), although the comparisons were conducted with small sample sizes. Furthermore, it is unclear if humans perform substantially better in certain conditions. In this paper, we set out to investigate whether humans can outperform machines at assessing kinship using a large sample and by investigating specific scenarios, such as verifying kinship of people of the same race or gender as the rater. We explore this hypothesis via the use of an online crowdsourcing platform, which enables near real-time kinship verification in practice.

We conducted our experiment on Amazon Mechanical Turk (MTurk) where 325 workers completed altogether 54,243 kinship verification tasks. The tasks originate from three datasets (KinFaceW-I, KinFaceW-II, UvA-NEMO Smile) which consist of pairs of images or video clips that each feature a person's face. Each pair was labelled as related (positive) or unrelated (negative), depending on whether the worker judged the depicted individuals to be related. Our datasets include all combinations of male and female pairs in a parent-child relationship (Father-son, Father-daughter, Mother-son, Mother-daughter), as well as sibling relationships (Brother-brother, Sister-sister, Sister-brother). Furthermore, the KinFaceW datasets include images of people of different races, such as East Asian, Black or African American, and White. We also recorded the demographics of participating workers.

We investigate the relationship between worker demographics (race and gender), nature of the task (gender and race of faces shown, use of video or still images), and the accuracy of the kinship verification. Our results indicate an own-race bias: humans are better at analysing image pairs of their own race, specifically when verifying East Asian and Black or African American image pairs. We also demonstrate that humans are more accurate when the two people appearing in the verification task are of the same gender, while the performance of male and female workers does not differ significantly. Finally, we show that human kinship verification accuracy is significantly higher when presented with videos instead of still images.

To the best of our knowledge, this is the first study to investigate the effect of gender, race and medium type on kinship verification in a crowdsourcing setting. Our work also opens avenues to explore how we can improve data quality in crowdsourcing by considering worker-specific characteristics during task assignment or routing.

Related Work

Kinship Recognition

Humans have an innate ability to recognise members of their own kin, which influences behaviour towards individuals based on the degree of relatedness. This ability has had several important benefits throughout our history, such as reducing the negative impact of inbreeding in mate selection (DeBruine 2004), as well as the trait of helping closely related kin over distant related individuals as suggested by the inclusive fitness theorem (Hamilton 1964). Furthermore, studies have found that self-resemblance plays a significant role in hypothetical investment decisions and the attachment of parents to their offspring (DeBruine 2004). Recognition of kin also increases attribution of trustworthiness (DeBruine et al. 2008), and influences pro-social behaviour (DeBruine 2002). In addition, there is contradicting literature on whether males or females exhibit more discriminated behaviour with regards to self-resemblance (Bressan and Zucchi 2009; DeBruine et al. 2008; Platek et al. 2004; Welling et al. 2011).

Humans are capable of this discriminated behaviour due to their ability to differentiate kinship cues (DeBruine et al. 2008). For instance, visual cues are the primary modality of kinship recognition, and can be categorised as either *contextual*: identifying kin based on a situation/location, or *phenotypic*: identifying kin based on externally observable elements such as hair colour. To a lesser extent, both olfactoric and acoustic modalities are also leveraged in kinship recognition (Mateo 2015). For instance, previous work has shown that mothers are able to recognise their offspring using visual as well as olfactoric cues (Porter et al. 1984; Russell et al. 1983). Similarly, breast-feeding infants are able to

recognise their mothers through the mother's axillary odour at two weeks of age (Cernoch and Porter 1985), whereas one-month old newborns can distinguish their mother's voice from other sounds (Jacques et al. 1978).

In addition, literature suggests that humans are better at recognising faces of people from their own race than from other, less familiar, races. This is known as the other-race effect, own-race bias, or cross-race effect (Meissner and Brigham 2001; Wu et al. 2012). Researchers demonstrated in a study involving 3, 6, and 9-month old infants that other-race effect develops as early as 9 months after birth, along with the development of the ability to capture facial input visually (Kelly et al. 2007). Similar studies have verified the presence of other-race effect in children and adults in different age categories (Pezdek et al. 2003). Interestingly, an experiment involving adults of Korean origin adopted by European Caucasian families has shown that the other-race effect could also be reversed when individuals are extensively exposed to a different race during their childhood (Sangrigoli et al. 2005). Furthermore, it has been shown that people also respond faster when they attempt to recognise faces of their own race when compared to the recognition of faces from other races (Wu et al. 2012). While kinship recognition is an active area of research, in our work we focus on kinship verification, which has several practical applications such as generating family trees using historical photos or for emergencies that involve, for example, missing children or seniors (Xia et al. 2011; Zhou et al. 2012).

Kinship Verification

Apart from being able to recognise own kin, humans are also capable of verifying kinship among strangers or people who are completely unrelated to themselves. Unrelated individuals have been able to match newborns to one of their parents (Alvergne et al. 2007; Kaminski et al. 2010; McLain et al. 2000), match children of different age categories to their parents (Alvergne et al. 2007, 2009; Nesse et al. 1990), verify sibling relationship among both children and adults (Dal Martello and Maloney 2006, 2010; DeBruine et al. 2009), and verify Father-son relationship among adults (Alvergne et al. 2014). When asked to assess facial resemblance, unrelated individuals have also given significantly higher resemblance ratings for children of different age categories and their parents in contrast to children and unrelated individuals (Bressan and Martello 2002; Oda et al. 2005). Furthermore, a number of studies have explored the ability of humans to verify kinship bonds beyond close family. Previous work has shown that people can successfully match faces of siblings to whom they are not related and assessed the relatedness of pairs of distant kin such as grandparents, grandchildren, aunts/uncles and nephews/nieces (Kaminski et al. 2009). Previous work also shows that the visible facial cues vary according to the degree of relatedness (Kaminski et al. 2009), and that the familiarity or the exposure to the face has limited impact on the ability to verify kinship (Alvergne et al. 2009).

Furthermore, variation in resemblance with regard to the gender of individual pairs has been explored in a number of studies (Alvergne et al. 2007; DeBruine et al. 2009; Kaminski et al. 2009). Previous work has shown that there is a higher detection rate for same-sex than opposite-sex between parent and offspring (Kaminski et al. 2010) and between adult siblings (DeBruine et al. 2009). Although there is a significant effect of the gender of individuals being verified, studies have shown that males and females have a similar capability in deciding on the kinship by assessing facial resemblance in photographs (Nesse et al. 1990) as well as recognising resemblance in pairs of related people (Oda et al. 2005).

While own-race bias has been extensively demonstrated in kinship recognition, it has not been verified in the domain of kinship verification. Limited literature (*i.e.*, small sample size) on the impact of race in kinship verification shows that there is no significant impact of the race or cultural background of humans on the performance of kinship verification (Alvergne et al. 2009). In our study we explore this further using participants representing three races to conduct a number of kinship verification tasks where the presented image pairs belong to participants' own race as well as other races.

Automated vs Human Performance

Several studies that investigate automated kinship verification have compared their results to human performance and have often reported that algorithms surpass human ability (*e.g.*, (Fang et al. 2010;

Zhou et al. 2012)). However, these studies have several limitations, such as a limited number of participants or present only a reduced subset of the original dataset.

A study by Fang et al. included an evaluation on human kinship verification ability using 16 participants and a limited random sample of 20 image pairs (Fang et al. 2010). Their results state that the resulting human accuracy of 67.19% was 4.9% lower than the algorithm they proposed. Zhou et al. used 100 pairs of randomly selected face samples following one of the following four categories (Father-daughter, Father-son, Mother-daughter and Mother-son), and verified them using 20 human observers (10 males and 10 females) of 20 to 30 years of age (Zhou et al. 2012). They utilised two approaches to display the images: showing only the image of cropped face region (as provided to the algorithm), or showing the whole face image, providing additional insights such as skin colour, hair and background. Their results show that when presented with cropped images, humans (Mean Accuracy 65.75%) perform worse than the automated approach (Mean Accuracy 69.75%). Lu et al. adopted a similar methodology with the KinFaceW-I and KinFaceW-II datasets (two datasets that we use in our study), although with an evaluation limited to 10 participants. They conclude that their automated approach outperforms humans for the KinFaceW-II dataset, whereas humans (Mean accuracy 71.0%) perform slightly better than the automated approach (Mean accuracy 69.9%) for the KinFaceW-I dataset when presented with the full image (Lu et al. 2014). While there exists more recent work on automated kinship verification in the domain of machine vision (e.g., (Wu et al. 2019)), they focus on comparing performance across methods but not with humans.

More generally, differences in performance between automated approaches and humans have been extensively studied in the crowdsourcing domain. For instance, previous work has shown that the accuracy of crowdsourced work can exceed the accuracy obtained by automated approaches in certain situations. Borromeo and Toyama showed that contributions originating from both paid and volunteer crowdsourcing systems could achieve better results than automated approaches when completing sentiment analysis tasks (Borromeo and Toyama 2014). In another study attempting to estimate the longitude/latitude coordinates of a video without using GPS data, Choi et al. showed that crowdsourcing can outperform algorithms in estimating the coordinates in cases where the training data includes a bias and/or videos have incorrect tags (Choi et al. 2013). Furthermore, in cases where automated approaches outperform crowd contributions, crowdsourcing can still offer additional benefits or even help train the algorithms (Cheng and Bernstein 2015). For instance, 'Legion:AR' (Lasecki et al. 2013) is a system for activity recognition that supplemented existing recognition systems with on-demand, real-time activity identification using input from the crowd. In another example, situated crowdsourcing input has been used to estimate queue lengths in real-time (Goncalves et al. 2016). While less accurate than camera-based methods, their crowdsourcing approach avoided other issues, such as occluding pillars or the need to purchase expensive equipment.

In our work, we aim to investigate how automated kinship verification techniques compare with human kinship verification through crowdsourcing in order to determine in which scenarios human intervention would be beneficial.

Study

The study uses three datasets, Kinship Face in the Wild I and II (KinFaceW-I and II) (Lu et al. 2014) and the UvA-NEMO Smile dataset (Dibeklioglu et al. 2012). KinFaceW-I and II datasets contain four kin relationships: Father-son (F-S), Father-daughter (F-D), Mother-son (M-S), and Mother-daughter (M-D). In the KinFaceW-I dataset, there are 156, 134, 116, and 127 positive pairs of kinship images for these four relations respectively. For the KinFaceW-II dataset, there are 250 positive pairs of kinship images for each relationship type. Each image was cropped to contain only the facial region of each individual. Using these two datasets, we created 3066 image pairs (1533 positive pairs, 1533 negative pairs). The UvA-Nemo Smile database comprises of 480 video pairs of people appearing with two types of smiles; deliberate (posed) smiles and genuine (spontaneous) smiles. We scaled the videos to a resolution of 480 X 270 pixels and created two types of tasks; showing the original videos (515 positive pairs, 515 negative pairs) or showing just the first frame of the video (515 positive pairs, 515 negative pairs), for a total number of 2060 pairs across seven kin relationships: Father-son (376 pairs), Father-

daughter (232 pairs), Mother-son (328 pairs), Mother-daughter (532 pairs), Brother-brother (112 pairs), Brother-sister (264 pairs) and Sister-sister (216 pairs). Thus, in total we had 5126 unique kinship verification tasks. The study contained an equal number of positive and negative pairs.

We recruited U.S.-based workers from MTurk to label all image pairs. Each of these *Human Intelligent Tasks* (HITs) presented the worker with two images along with the question "*Are these two people related (i.e., part of the same family?)*". Workers were given the option to select either "Yes" or "No" (Figure 1). This binary classification is in line with results produced by automated kinship classifiers (*i.e.*, no scale-based classification) allowing for more straightforward comparisons.

Figure 1. Overview of one of the kinship verification tasks from the UvA-NEMO dataset.

Furthermore, we took a number of steps to ensure the overall reliability of the answers by excluding automated answers or non-serious answers. First, we designed the experiment such that each task is attempted by exactly 10 different workers. Each task was assigned the value 1 (correct) or 0 (incorrect) based on the response provided by the worker and the ground truth. Second, we established a primary entry criterion to ensure we have a genuine pool of workers. Only workers who had already completed at least 1000 HITs in MTurk and who had a 99% approval rate were allowed to participate in our study. Third, we excluded answers that were provided too quickly (less than 1 second). Finally, we created and included a set of explicitly verifiable questions (Goncalves et al. 2013; Kittur et al. 2008). The inclusion of these "fact-checking" questions has been shown to improve the quality of completed tasks as workers become aware of prompt response verification (Goncalves et al. 2013). In this case, we added images of cartoon and movie characters paired with humans as shown in Figure 2, which are obviously not related. Based on the results of these fact-checking tasks, we were able to remove the answers of workers that were deemed as not participating in the experiment in a serious manner.

Figure 2. Examples of explicitly verifiable questions.

Survey

At the end of the kinship verification data collection, we invited workers to complete a survey. The survey gathered demographics such as age, gender, and race.

Results

A total of 325 workers completed 54,243 HITs, of which we rejected 2,983 and retained 51,260 for further analysis. The accuracy for each dataset was 77.86% (KinFaceW-I), 82.85% (KinFaceW-II), 73.50% (UvA-NEMO Smile Images) and 78.54% (UvA-NEMO Smile Videos). Overall, the combined accuracies were 81.12% (KinFaceW datasets) and 76.02% (UvA-NEMO Smile). The average time

taken by workers to complete each task was 14.26 s (KinFaceW-I), 16.55 s (KinFaceW-II), 12.68 s (UvA-NEMO Smile Images) and 14.64 s (UvA-NEMO Smile Videos). We report Receiver Operating Characteristics (ROC) curves in our study. An ROC curve is a two-dimensional graph in which *true positive rate* is plotted against the *false positive rate*. It is a popular technique for visualising and organising binary categories based on their performance. We also present the area under the ROC curve (AUC) which serves as a single scalar value representing the expected performance (Fawcett 2004). ROC curves for each dataset regarding human performance are shown in Figure 3.

We conducted a chi-square test of independence using the data collected from the UvA-NEMO Smile dataset to investigate the impact of media type (videos and still images) on the likelihood of a task being answered correctly. The relation between these variables was significant, $\chi^2(1, N = 2060) = 6.93$, $p < .01$ indicating that humans are more likely to accurately verify videos (78.5%) as opposed to still images (73.5%). Further analysis using the UvA-NEMO Smile dataset revealed no significant impact of the nature of the smile posed (deliberate or spontaneous) on the likelihood of a task being answered correctly.

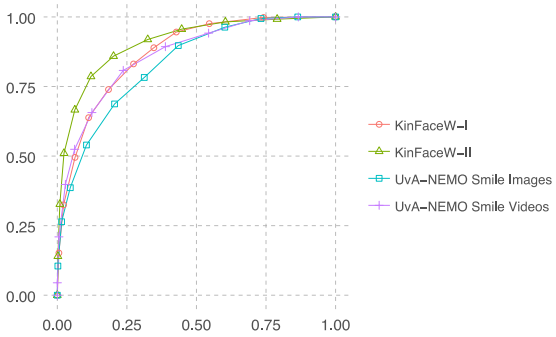

Figure 3. ROC - Datasets.

Survey Responses

A total of 81 workers completed our survey task. These 81 workers completed 25,825 tasks out of the 51,260 tasks (50.4%) in the experiment. The age of the workers ranged from 20 years to 59 years (M = 36.22, SD = 10.48). There were 42 female workers, 38 male workers and one worker who chose not to specify their gender. In terms of worker race, there were 18 East Asian, 19 Black or African American, and 44 White workers. This distribution is similar to the demographics of both MTurk (US marketplace) and the United States population (Ross et al. 2010; Vespa et al. 2018).

Impact of Gender

We conducted a two-way ANOVA test on the influence of two independent variables (gender of the worker, kin relationship of the image pair) on task accuracy. The analysis yielded a significant main effect for the kin relationship of the image pair, $F(6, 254) = 7.07$, $p < .001$. However, there was no main effect of gender of the worker, $F(1, 254) = 0.84$, $p > .05$, nor a significant interaction effect between the two variables, $F(6, 254) = 1.40$, $p > .05$. We further investigate the effect of kin relationship of the image pair using the complete dataset. We reduced these relationships into two categories (same gender and opposite gender pair). A one-way ANOVA showed a significant effect of kin relationship category on the accuracy, $F(1, 5124) = 15.78$, $p < 0.01$. Posthoc comparisons using a TukeyHSD test indicated that the mean accuracy for same gender pairs (M = 81.1, SD = 39.1) was significantly higher than opposite gender pairs (M = 76.6, SD = 42.4).

Figures 4 shows the ROC curves by kin relationship of the image pair for KinFaceW (I and II aggregated) and UvA-NEMO Smile (Images and Videos aggregated). We further examined the AUC value of the ROC curve for each category which is summarised, along with mean accuracy, in Table 2. We observed that AUC values are higher for categories which involve people from the same gender (Mother-daughter, Father-son, Brother-brother and Sister-sister).

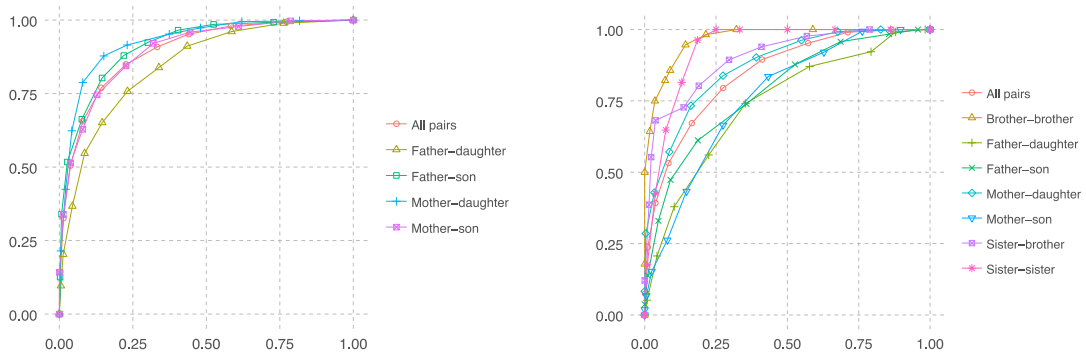

Figure 4. Aggregated ROC by kin relationship of the image pair for KinFaceW-I & II (left) and UvA-NEMO Smile Images and Videos (right).

Table 2. Summary of mean accuracy and AUC values by kin relationship of the image pair for KinFaceW (I and II) and UvA-NEMO Smile (Images and Videos).

	KinFaceW		UvA-NEMO	
Kin Relationship	**Acc.**	**AUC**	**Acc.**	**AUC**
Father-son	76.3	0.85	69.3	0.74
Father-daughter	83.0	0.91	69.1	0.79
Mother-son	84.2	0.93	78.2	0.87
Mother-daughter	80.8	0.90	69.5	0.76
Brother-brother	-	-	88.4	0.97
Brother-sister	-	-	87.5	0.93
Sister-sister	-	-	80.7	0.90
All-pairs	81.1	0.89	76.0	0.85

Impact of Race

Preprocessing

To evaluate the impact of worker race on kinship verification, two of the paper's authors of different race, carefully labelled all the images in the KinFaceW-I and II datasets shown in Figure 5. We excluded the UvA-NEMO Smile dataset from this process as it did not contain sufficient race variability within the images. An interrater reliability analysis using Cohen Kappa was performed to determine consistency among raters ($\kappa = 0.83$). Images with contradicting labels were re-evaluated in the presence of a third author to reach a consensus. We used four race categories for the labelling process based on the results of the survey, namely *East Asian*, *Black or African American*, *White* and *Other*. If both images in a task belong to the same race category, the image pair was classified under that race.

Figure 5. Sample image pairs with race labels.

To increase the reliability of the analysis based on the race of the worker and the race of the image pair, we selected a subset of workers who provided more than 5 labels (15 East Asian, 16 Black or African American, and 41 White workers). This subset of workers completed a total of 13,634 tasks. A two-way ANOVA test conducted on the influence of two independent variables (race of the worker, race of

the image pair) on the accuracy of the worker yielded no significant main effect for the race of the image pair, $F(2, 1631) = 2.99$, $p > 0.05$. Figure 6 (left) shows the ROC curves for different race categories of the image pair while Table 3 summarises the mean accuracy for task and AUC values.In addition, there was no main effect of the race of the worker, $F(2, 1631) = 0.36$, $p > 0.05$. However, there was a significant interaction effect between these two variables, $F(4, 1631) = 3.82$, $p < 0.01$. (Figure 6 right).

Table 3. Accuracy and AUC values by race of the image pair for KinFaceW (I and II).

Race of the image pair	Accuracy	AUC
East Asian	72.9	0.79
Black or African American	84.0	0.89
White	75.8	0.84
Cross-race	88.7	0.70

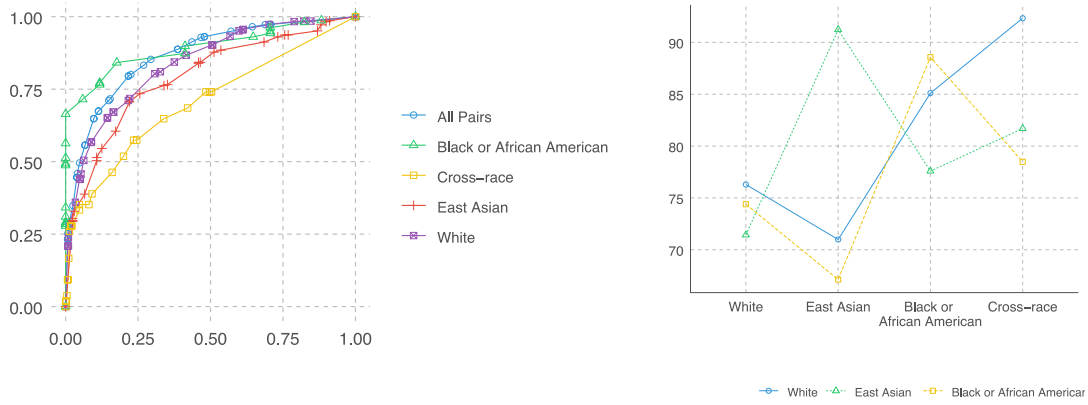

Figure 6. ROC curves by race of the image pair for KinFaceW (I and II aggregated) (left) and Interaction Effect between Race of the Worker and Race of the Image Pair (right).

We observe that when presented with an image pair of East Asians, East Asian workers are significantly more accurate than Black or African American workers. East Asian workers also outperform White workers in the same situation. Similarly, when verifying an image pair of Blacks or African Americans, Black or African American workers perform better than East Asian workers as well as White workers. With regards to image pairs of White people, White workers tend to outperform both East Asian workers and Black or African American workers although the difference in accuracy is slightly lower when compared to image pairs other races.

A two-way ANOVA conducted on the influence of two independent variables (race of the worker, race of the image pair) on the mean time (seconds) taken by the worker for the verification yielded a significant main effect for the race of the worker, $F(2,1631) = 124.62$, $p < 0.01$. Mean time taken by East Asian workers ($M = 28.5$, $SD = 20.2$) was significantly higher than time taken by Black or African American workers ($M = 14.5$, $SD = 8.3$) and White workers ($M = 15.7$, $SD = 8.1$). In addition, there was a significant interaction effect $F(4,1631) = 2.83$, $p < 0.01$ between the two variables. However, the main effect for the race of the image pair $F(2,1631) = 0.89$, $p > 0.05$ was not significant.

Humans vs Automated Approaches

Lu et al. introduced a protocol to evaluate the performance of different kinship verification algorithms. It includes three different settings (Lu et al. 2015):

- *Unsupervised*: No prior kin relationship information (labelled image pairs) is utilised and from the perspective of automated approaches, it is deemed as the most challenging setting. This setting is also considered as the most comparable to humans performing kinship verification.

- *Image-restricted*: A set of pre-labelled image pairs with kin relationship information is utilised to train the model. The protocol recommends using a pre-specified training and testing split, which has been generated randomly and independently for 5-fold cross validation.
- *Image-unrestricted*: Similar to Image-restricted, a set of pre-labelled image pairs with kin relation information is utilised to train the model. However, it also contains identity information (*i.e.*, additional negative pairs can be created).

Table 4 compares the accuracy of automatic approaches (unsupervised, image-restricted and image-unrestricted) against human performance in various scenarios based on the rater's race, whereas Table 5 compares human and machine accuracy based on the kin relationship of the image pair.

Table 4. Comparison of human and machine performance for KinFaceW-I & II.

Method	Accuracy
SILD (LBP) (Lu et al. 2015) [a]	69.4
SILD (HOG) (Lu et al. 2015) [a]	71.2
Polito (Lu et al. 2015) [b]	84.7
LIRS (Lu et al. 2015) [b]	84.4
BIU (LBP) (Lu et al. 2015) [c]	77.9
BIU (HOG) (Lu et al. 2015) [c]	79.7
Humans	81.1
East Asian workers verifying own-race pairs	**91.2**
East Asian workers other-race pairs	74.1
Black or African American workers verifying own-race pairs	**88.6**
Black or African American workers verifying other-race pairs	71.0
White workers verifying own-race pairs	76.3
White workers verifying other-race pairs	72.3

[a] Unsupervised [b] Image-restricted [c] Image-unrestricted

Table 5. Classification accuracy (%) of the KinFaceW-I and II datasets based on kin relation of the image pair.

Method	Same Gender	Opposite Gender	Mean
SILD (LBP) (Lu et al. 2015) [a]	72.0	66.9	69.4
SILD (HOG) (Lu et al. 2015) [a]	72.8	69.5	71.2
Polito (Lu et al. 2015) [b]	84.3	85.1	84.7
LIRS (Lu et al. 2015) [b]	85.6	83.2	84.4
BIU (LBP) (Lu et al. 2015) [c]	80.3	75.5	77.9
BIU (HOG) (Lu et al. 2015) [c]	82.4	76.9	79.7
Humans	83.6	78.5	81.1

[a] Unsupervised [b] Image-restricted [c] Image-unrestricted

Discussion

Impact of Gender and Race on Kinship Verification

In our study we explore the impact of gender and race of both the rater and the people appearing on the images or videos on human kinship verification accuracy in a crowdsourcing setting. Our results indicate that both males and females are equally competent at assessing kinship, which is in line with

previous work on sex differences in ability to recognise family resemblance (Nesse et al. 1990). However, there was a significant main effect of the gender of the people appearing in the tasks, as also reported in previous work (Kaminski et al. 2010). Our results show that accuracy is significantly higher for same gender image pairs (Father-son, Mother-daughter, Brother-brother, Sister-sister) in contrast to opposite gender image pairs (Father-daughter, Mother-son, Sister-brother). We also show that there is no interaction effect between the gender of the worker and the gender of the people appearing in the task. In other words, our results indicate that there is no advantage in assigning workers to only perform kinship verification of image pairs of their own gender.

Regarding race, our results contradict the findings from previous work (Alvergne et al. 2007), which suggest that there is no impact of race in kinship verification. In contrast to their study, which only used two race categories and a small sample of 114 child-parent image sets, we examined the impact of race using three race categories across 3066 image pairs. Our finding of the presence of own-race bias in kinship verification is well-supported by literature in kinship recognition, *i.e.*, when considering own family members (Kelly et al. 2007; Pezdek et al. 2003; Wu et al. 2012). We show that East Asian, Black or African American and White workers were better at assessing image pairs of their own race. The fact that workers from different races do not demonstrate any significant deviations in accuracy when presented with an image pair of White people could be explained through the notion of reversibility of the other-race effect (Sangrigoli et al. 2005). In their study, Sangrigoli et al. showed that extensive exposure to a particular race could reverse the effect of own-race bias in face recognition. Since our participants were crowdworkers from the US (around 75% of the population is White (Vespa et al. 2018)), it is likely that our workers are more accustomed to the White population, which results in an absence of the own-race bias when presented with image pairs of Whites.

Finally, while literature suggests that people respond faster when they attempt to recognise faces of their own race when compared to the recognition of faces from other races (Wu et al. 2012), we did not observe this effect in our data. Mean time taken by workers to verify image pairs of people who belong to their own race did not significantly differ from the mean time taken to verify image pairs of people from other races.

Human Kinship Verification

In our study, we observe that humans are capable of outperforming state-of-the-art automated approaches in kinship verification, particularly when compared to unsupervised techniques (*i.e.*, no labelled data available). When considering specific applications of kinship verification, such as finding relatives of a missing person, we identify several critical factors. The images presented for comparison could be highly inconsistent unlike the images used for bench-marking algorithms due to variations in terms of lighting condition, image quality, angle of the face, etc. Such circumstances could result in poor performance of automated verification approaches, while humans are more flexible with varying conditions (Lopez et al. 2018; López et al. 2016). Automated approaches based on supervised learning, such as the Image-restricted and unrestricted settings shown in Tables 4 and 5, could also suffer from lack of relevant training data (based on the properties of the image or the demographics of the person appearing). Thus, we argue that human-in-the-loop systems for kinship verification has substantial potential, as shown with our results, with crowdsourcing platforms allowing us to easily and quickly recruit the required workforce. For instance, humans could be asked to provide their judgement in certain scenarios, such as images of people of certain race or when the image characteristics are highly inconsistent, which diminishes the effectiveness of automated approaches (López et al. 2016). On the other hand, for applications such as organising multimedia via kinship verification, purely automated approaches would likely be preferred due to economic and practical reasons. Nevertheless, in such scenarios data collected from the crowd could be used to train and further improve automated algorithms.

Finally, several studies have used video-based media for kinship verification (Dibeklioglu et al. 2013; Yan and Hu 2018). However, there has been no previous study on the impact of media type on human kinship verification performance. Our findings show that humans are more accurate in verifying kinship when presented with videos instead of still images. Thus, we recommend the use of dynamic media types, if available, when asking humans to complete kinship verification tasks.

Task Assignment and Routing in Crowdsourcing

In their discussion on the future of crowdwork, Kittur at al. identified task assignment as one of the research foci that could increase the value and meaning of the contributed data (Kittur et al. 2013). While task assignment typically follows a first-come/first-serve model (*e.g.*, (von Ahn and Dabbish 2004)) or a market model (*e.g.*, MTurk, (Hosio et al. 2014)), the assignment of tasks in relation to individuals' abilities has been increasingly researched with promising results (Hettiachchi et al. 2019; Shen et al. 2003). In our work we show that it is possible to attain better results in kinship verification by assigning tasks based on certain worker characteristics (*i.e.*, race). This could be further applied to different crowdsourcing applications where a certain subset of workers could be more capable and suited for tasks, reducing the likelihood of the task being too difficult, which often leads to worker dissatisfaction and task abandonment (Kittur et al. 2013).

Popular crowdsourcing platforms, such as MTurk, could provide automated recommendations of possible next tasks based on workers' characteristics, such as specific cognitive abilities (Goncalves et al. 2017; Hettiachchi et al. 2019) or demographics as shown in our study and in previous work (Li et al. 2014), while still maintaining the market model. One way to achieve this is to offer workers the possibility to provide more information on their profile, while at the same time emphasising that this would not reduce in any way the number of available tasks to them, but instead simply help with worker-task matching. This is of particular importance as person-job misfit has been shown to substantially affect performance and places increasing strain on both workers and requesters over time (Chilton et al. 2005). Such an approach would be particularly useful when considering new workers of a platform, as there is no past performance to predict how well they will do on similar or relevant tasks, so collecting this information upon or shortly after registration would be ideal.

Limitations

We acknowledge several limitations in our study. First, our comparison of human kinship verification with automated approaches is limited to the KinFaceW-I and KinFaceW-II datasets. Including the UvA-NEMO Smile dataset would have allowed us to draw comparisons with regard to sibling relationships in addition to the presented parent-child relationships. However, results for this dataset are not publicly available. Second, there were no age labels available in the datasets used for the study, so we were unable to investigate the impact of age on kinship verification accuracy. We refrained from labelling the dataset ourselves as it would most likely have high variability and lead to a bias in the results. Finally, the subset of workers that completed the survey reported being from three races. Future work could investigate the impact of more race combinations (rater and ratees) on human kinship verification accuracy. Third, we acknowledge that asking participants to recall what facial features they used in their judgement is not optimal. Since information about what facial features were considered by the workers was not the main focus of our study, we opted for this method in order to minimise any possible disruption to the main task.

Conclusion

In this paper we provide an in-depth analysis on the effect of gender, race and media type on human kinship verification accuracy. We conducted an online crowdsourcing study using Amazon Mechanical Turk in which participants assessed the kinship of two people based on images or videos of their faces. In our analysis, we consider gender and race of both the people shown in the images as well as the person doing the actual assessment. Our results show that East Asians and Black or African Americans are better at verifying kinship of their own race, indicating an own-race bias in kinship verification. Our results also confirm that humans are better at assessing image pairs of people of the same gender as opposed to image pairs of opposite genders. We further establish that in both these scenarios, human accuracy surpasses the accuracy of unsupervised automated kinship verification approaches. However, we show that there is no significant effect of gender of the worker nor an interaction effect between gender of the worker and the gender of the people appearing in the task on accuracy. In addition, we establish that humans are more accurate in verifying kinship in videos than in still images.

Our results show that there are significant differences between demographic groups in their assessment of human kinship. These results have implications for future kinship studies and establish a new baseline for automated verification approaches. For instance, in a crowdsourcing setting, rather than presenting kinship verification tasks to an arbitrary audience, selection of participants based on race can significantly improve verification results.

References

von Ahn, L., and Dabbish, L. 2004. "Labeling Images with a Computer Game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, New York, NY, USA: ACM, pp. 319–326.

Alvergne, A., Faurie, C., and Raymond, M. 2007. "Differential Facial Resemblance of Young Children to Their Parents: Who Do Children Look like More?," *Evolution and Human Behavior* (28:2), pp. 135–144.

Alvergne, A., Oda, R., Faurie, C., Matsumoto-Oda, A., Durand, V., and Raymond, M. 2009. "Cross-Cultural Perceptions of Facial Resemblance between Kin," *Journal of Vision* (9:6), p. 23.

Alvergne, A., Perreau, F., Mazur, A., Mueller, U., and Raymond, M. 2014. "Identification of Visual Paternity Cues in Humans," *Biology Letters* (10:4), The Royal Society.

Borromeo, R. M., and Toyama, M. 2014. "Automatic vs. Crowdsourced Sentiment Analysis," in *Proceedings of the 19th International Database Engineering & Applications Symposium*, IDEAS '15, New York, NY, USA: ACM, pp. 90–95.

Bressan, P., and Kramer, P. 2015. "Human Kin Detection," *Wiley Interdisciplinary Reviews: Cognitive Science* (6:3), Wiley-Blackwell, pp. 299–311.

Bressan, P., and Martello, M. F. D. 2002. "Talis Pater, Talis Filius: Perceived Resemblance and the Belief in Genetic Relatedness," *Psychological Science* (13:3), pp. 213–218.

Bressan, P., and Zucchi, G. 2009. "Human Kin Recognition Is Self- Rather than Family-Referential," *Biology Letters* (5:3), pp. 336–338.

Cernoch, J. M., and Porter, R. H. 1985. "Recognition of Maternal Axillary Odors by Infants," *Child Development* (56:6), pp. 1593–1598.

Cheng, J., and Bernstein, M. S. 2015. "Flock: Hybrid Crowd-Machine Learning Classifiers," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, New York, NY, USA: ACM, pp. 600–611.

Chilton, M. A., Hardgrave, B. C., and Armstrong, D. J. 2005. "Person-Job Cognitive Style Fit for Software Developers: The Effect on Strain and Performance," *Journal of Management Information Systems* (22:2), Routledge, pp. 193–226.

Choi, J., Lei, H., Ekambaram, V., Kelm, P., Gottlieb, L., Sikora, T., Ramchandran, K., and Friedland, G. 2013. "Human vs Machine: Establishing a Human Baseline for Multimodal Location Estimation," in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, New York, NY, USA: ACM, pp. 867–876.

Dal Martello, M. F., and Maloney, L. T. 2006. "Where Are Kin Recognition Signals in the Human Face?," *Journal of Vision* (6:12), p. 2.

Dal Martello, M. F., and Maloney, L. T. 2010. "Lateralization of Kin Recognition Signals in the Human Face," *Journal of Vision* (10:8), p. 9.

DeBruine, L. M. 2002. "Facial Resemblance Enhances Trust," *Proceedings of the Royal Society of London B: Biological Sciences* (269:1498), The Royal Society, pp. 1307–1312.

DeBruine, L. M. 2004. "Resemblance to Self Increases the Appeal of Child Faces to Both Men and Women," *Evolution and Human Behavior* (25:3), pp. 142–154.

DeBruine, L. M., Jones, B. C., Little, A. C., and Perrett, D. I. 2008. "Social Perception of Facial Resemblance in Humans," *Archives of Sexual Behavior* (37:1), pp. 64–77.

DeBruine, L. M., Smith, F. G., Jones, B. C., Craig Roberts, S., Petrie, M., and Spector, T. D. 2009. "Kin Recognition Signals in Adult Faces," *Vision Research* (49:1), pp. 38–43.

Dibeklioglu, H., Salah, A. A., and Gevers, T. 2012. "Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles," in *Computer Vision -- ECCV 2012*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 525–538.

Dibeklioglu, H., Salah, A. A., and Gevers, T. 2013. "Like Father, Like Son: Facial Expression Dynamics for Kinship Verification," in *2013 IEEE International Conference on Computer Vision*, IEEE, pp. 1497–1504.

Fang, R., Tang, K. D., Snavely, N., and Chen, T. 2010. "Towards Computational Models of Kinship Verification," in *Proceedings - International Conference on Image Processing, ICIP*, IEEE, pp. 1577–1580.

Fawcett, T. 2004. "ROC Graphs: Notes and Practical Considerations for Researchers," *Machine Learning* (31:1), pp. 1–38.

Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., and Kostakos, V. 2013. "Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, New York, NY, USA: ACM, pp. 753–762.

Goncalves, J., Kukka, H., Sánchez, I., and Kostakos, V. 2016. "Crowdsourcing Queue Estimations in Situ," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, New York, NY, USA: ACM, pp. 1040–1051.

Goncalves, J., Feldman, M., Hu, S., Kostakos, V., and Bernstein, A. 2017. "Task Routing and Assignment in Crowdsourcing based on Cognitive Abilities," *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17, pp. 1023–1031.

Hamilton, W. D. 1964. "The Genetical Evolution of Social Behaviour." *Journal of Theoretical Biology* (7:1), pp. 1–16.

Hettiachchi, D., van Berkel, N., Hosio, S., Kostakos, V., and Goncalves, J. 2019. "Effect of Cognitive Abilities on Crowdsourcing Task Performance," in *Human-Computer Interaction -- INTERACT 2019*, Cham: Springer International Publishing, pp. 442–464.

Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., and Kostakos, V. 2014. "Situated Crowdsourcing Using a Market Model," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, New York, NY, USA: ACM, pp. 55–64.

Jacques, M., Josiane, B., Michele, B., and Dora, J.-G. 1978. "Infant Recognition of Mother's Voice," *Perception* (7:5), pp. 491–497.

Kaminski, G., Dridi, S., Graff, C., and Gentaz, E. 2009. "Human Ability to Detect Kinship in Strangers' Faces: Effects of the Degree of Relatedness," *Proceedings of the Royal Society of London B: Biological Sciences* (276:1670), The Royal Society, pp. 3193–3200.

Kaminski, G., Méary, D., Mermillod, M., and Gentaz, E. 2010. "Perceptual Factors Affecting the Ability to Assess Facial Resemblance between Parents and Newborns in Humans," *Perception* (39:6), pp. 807–818.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., and Pascalis, O. 2007. "The Other-Race Effect Develops During Infancy: Evidence of Perceptual Narrowing," *Psychological Science* (18:12), pp. 1084–1089.

Kittur, A., Chi, E. H., and Suh, B. 2008. "Crowdsourcing User Studies with Mechanical Turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, New York, NY, USA: ACM, pp. 453–456.

Kittur, A., Nickerson, J. V, Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. 2013. "The Future of Crowd Work," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, New York, NY, USA: ACM, pp. 1301–1318.

Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. 2013. "Real-Time Crowd Labeling for Deployable Activity Recognition," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, New York, NY, USA: ACM, pp. 1203–1212.

Li, H., Zhao, B., and Fuxman, A. 2014. "The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, New York, NY, USA: ACM, pp. 165–176.

López, M. B., Boutellaa, E., and Hadid, A. 2016. "Comments on the 'Kinship Face in the Wild' Data Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (38:11), IEEE, pp. 2342–2344.

Lopez, M. B., Hadid, A., Boutellaa, E., Goncalves, J., Kostakos, V., and Hosio, S. 2018. "Kinship Verification from Facial Images and Videos: Human versus Machine," *Machine Vision and Applications* (29:5), pp. 873–890.

Lu, J., Hu, J., Liong, V. E., Zhou, X., Bottino, A., Islam, I. U., Vieira, T. F., Qin, X., Tan, X., Chen, S., Mahpod, S., Keller, Y., Zheng, L., Idrissi, K., Garcia, C., Duffner, S., Baskurt, A., Castrillón-Santana, M., and Lorenzo-Navarro, J. 2015. "The FG 2015 Kinship Verification in the Wild Evaluation," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 1), IEEE, pp. 1–7.

Lu, J., Zhou, X., Tan, Y. P., Shang, Y., and Zhou, J. 2014. "Neighborhood Repulsed Metric Learning for Kinship Verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (36:2), pp. 331–345.

Mateo, J. M. 2015. "Perspectives: Hamilton's Legacy: Mechanisms of Kin Recognition in Humans," *Ethology* (121:5), pp. 419–427.

McLain, D. K., Setters, D., Moulton, M. P., and Pratt, A. E. 2000. "Ascription of Resemblance of Newborns by Parents and Nonrelatives," *Evolution and Human Behavior* (21:1), pp. 11–23.

Meissner, C. A., and Brigham, J. C. 2001. "Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review," *Psychology, Public Policy, and Law* (7:1), pp. 3–35.

Nesse, R. M., Silverman, A., and Bortz, A. 1990. "Sex Differences in Ability to Recognize Family Resemblance," *Ethology and Sociobiology* (11:1), pp. 11–21.

Oda, R., Matsumoto, A., and Kurashima, O. 2005. "Effects of Belief in Genetic Relatedness on Resemblance Judgments by Japanese Raters," *Evolution & Human Behavior* (26:5), pp. 441–450.

Pezdek, K., Blandon-Gitlin, I., and Moore, C. 2003. "Children's Face Recognition Memory: More Evidence for the Cross-Race Effect," *Journal of Applied Psychology* (88:4), pp. 760–763.

Platek, S. M., Raines, D. M., Gallup, G. G., Mohamed, F. B., Thomson, J. W., Myers, T. E., Panyavin, I. S., Levin, S. L., Davis, J. A., Fonteyn, L. C. M., and Arigo, D. R. 2004. "Reactions to Children's Faces: Males Are More Affected by Resemblance than Females Are, and so Are Their Brains," *Evolution and Human Behavior* (25:6), pp. 394–405.

Porter, R. H., Cernoch, J. M., and Balogh, R. D. 1984. "Recognition of Neonates by Facial-Visual Characteristics," *Pediatrics* (74:4), pp. 501–504.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. 2010. "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, New York, NY, USA: ACM, pp. 2863–2872.

Russell, M. J., Mendelson, T., and Peeke, H. V. S. 1983. "Mother's Identification of Their Infant's Odors," *Ethology and Sociobiology* (4:1), pp. 29–31.

Sangrigoli, S., Pallier, C., Argenti, A., Ventureyra, V., and de Schonen, S. 2005. "Reversibility of the Other-Race Effect in Face Recognition during Childhood," *Psychological Science* (16:6), pp. 440–444.

Shen, M., Tzeng, G.-H., and Liu, D.-R. 2003. "Multi-Criteria Task Assignment in Workflow Management Systems," in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of The*, IEEE, p. 9.

Vespa, J., Armstrong, D. M., and Medina, L. 2018. "Demographic Turning Points for the United States: Population Projections for 2020 to 2060 Population Estimates and Projections," Washington, DC.

Welling, L. L. M., Burriss, R. P., and Puts, D. A. 2011. "Mate Retention Behavior Modulates Men's Preferences for Self-Resemblance in Infant Faces," *Evolution and Human Behavior* (32:2), Elsevier, pp. 118–126.

Wu, E. X. W., Laeng, B., and Magnussen, S. 2012. "Through the Eyes of the Own-Race Bias: Eye-Tracking and Pupillometry during Face Recognition," *Social Neuroscience* (7:2), pp. 202–216.

Wu, X., Feng, X., Boutellaa, E., and Hadid, A. 2019. "Kinship Verification Using Color Features and Extreme Learning Machine," *2018 IEEE 3rd International Conference on Signal and Image Processing, ICSIP 2018*, IEEE, pp. 187–191.

Xia, S., Shao, M., and Fu, Y. 2011. "Kinship Verification through Transfer Learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (Vol. 22), Menlo Park, California: AAAI Press, p. 2539.

Yan, H., and Hu, J. 2018. "Video-Based Kinship Verification Using Distance Metric Learning," *Pattern Recognition* (75), pp. 15–24.

Zhou, X., Lu, J., Hu, J., and Shang, Y. 2012. "Gabor-Based Gradient Orientation Pyramid for Kinship Verification Under Uncontrolled Environments," in *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, New York, NY, USA: ACM, pp. 725–728.