Kinship Verification from Facial Images and Videos: Human vs. Machine Supplemental Material

Miguel Bordallo López[1], Abdenour Hadid[1], Elhocine Boutellaa[1], Jorge Gonçalves[2], Vassilis Kostakos[2], Simo Hosio[2]

Index Terms—kinship verification, biometrics.

1 Appendix A: Available datasets

Automatic kinship verification methods are usually evaluated on publicly available data sets. Table 1 provides a list of publicly available data sets that can be used for kinship verification.

1.1 Kinship Faces in the Wild I and II

The *Kinship Face in the Wild* data sets are currently used as a benchmark for the evaluation of the best performing automatic kinship verification algorithms. These kinship face image data sets (known as *KinFaceW I & II*) are now available and widely distributed from www.kinfacew.com. The data sets have been utilized to evaluate the performance of kinship verification algorithms, and the results have been reported in several top forums and competitions such as TPAMI [1], CVPR [2] or ICCV [3].

The *KinFaceW* data sets contain face images collected from the internet depicting four classes of family relationships: Father-Son (F-S), Father-Daughter (F-D), Mother-Son (M-S) and Mother-Daughter (M-D). KinFaceW-I data set has 156, 134, 116 and 127 pairs of kinship face images for the aforementioned relationships while *KinFaceW-II* has 250 pairs for each one. Figure 1 show examples of images from the data sets.

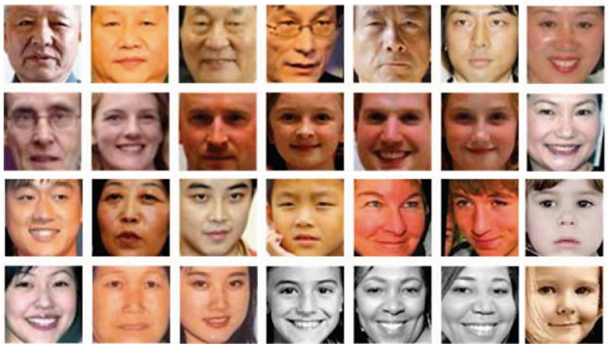

Fig. 1. Examples of images extracted from the *KinFaceW-I and II* data sets.

The datasets provide a description of the image collection process, describing images under uncontrolled environments with no restrictions in terms of pose, lighting, background, expression, age, ethnicity and partial occlusion. Both datasets include images of many public figures and their relatives.

The provided faces in both datasets are aligned using eyes coordinates and cropped into 64×64 pixels. A predefined training/testing split is provided for each dataset to serve as a benchmark for fair comparison between different methods. The splitting consists of randomly generated 5-fold cross validation.

1.2 UvA NEMO Smile dataset

The UvA-NEMO Smile Database (*SmileDB*) [4] was collected to analyze the dynamics of spontaneous/posed enjoyment smiles. This database is composed of 1240 videos recorded with a Panasonic HDC-HS700 3MOS camcorder, placed on a monitor, at approximately 1.5 meters away from the recorded subjects. Videos were recorded with a resolution of 1920×1080 pixels at a rate of 50 frames per second under controlled illumination conditions.

A subset of the database contains subjects who have a kin relationship, and composes the only available kinship verification database that uses video data and allows the study of the importance of spatiotemporal information. This kinship subset comprises 502 videos of 152 subjects with kin relationships. There are 15 subjects without spontaneous smile videos and there is no posed video for six subjects. The remaining subjects in the database has one or two posed/spontaneous smile videos. The ages of subjects vary from 8 to 74 years.

Table 2 depicts the relationship statistics of the UvA-NEMO smile database. It shows a total of 95 kinship relations defined between the subjects included in the dataset. The combination of different videos of each kin relation gives 228 pairs of spontaneous and 287 pairs of posed smile videos. These pairs consist of Sister-Sister (SS), Brother-Brother (B-B), Sister-Brother (S-B), Mother-Daughter (M-

TABLE 1
Kinship verification data sets available to the kinship research community as of June 2016.

Data Set	Contact point	Number of kinship pairs	Types of relationships	Controlled environment	Pairs from same photo	Public figures
Smile DB	http://www.uva-nemo.org/	95	7	Yes	No	No
KinFaceW-I	http://www.kinfacew.com/	500+	4	No	Partially	Partially
KinFaceW-II	http://www.kinfacew.com/	1000+	4	No	Yes	Partially
UB KinFace	http://www.computervisiononline.com/dataset/ub-kinface	400	4	No	No	Partially
Family101	http://chenlab.ece.cornell.edu/projects/KinshipClassification	206	4	No	Partially	Yes
CornellKin	http://chenlab.ece.cornell.edu/projects/KinshipVerification	150	4	No	Partially	Yes
TS-KinFace	http://parnec.nuaa.edu.cn/xtan/data/TSKinFace.html	1000+	5	No	Yes	Yes

D), Mother-Son (M-S), Father-Daughter (FD), and Father-Son (F-S) relationships.

For evaluation of the automatic approach, the available kinship relations in the database can be used as positive samples. However, the protocol defined for kinship verification proposes to randomly generate negative pairs by associating images of persons with no kinship relation. The same number of negative pairs as positive should be defined for each subset. Therefore, for each positive pair, the first image is retained, replacing the second one by that of a person from the same subset who has no kinship relation with the person in the first image. As the number of pairs in the subsets is limited, the database description suggests a leave-one-out cross-validation evaluation scheme.

1.3 Tri-subject Kinship Face database

The Tri-subject Kinship Face Database, (*TSKinFace*) [5] is a recently published dataset that contains two groups of family-based kinship relations: Father-Mother-Daughter (FM-D), Father-Mother-Son (FM-S). FM-D and FM-S relations have 513 and 502 tri-subject groups respectively, which are collected from wild circumstances from Internet of public figures. Face images are cropped from the same original image into a resolution of 64×64 pixels.

The TSKinFace dataset includes a testing protocol with two groups composed of three subjects: Father-Mother-Daughter and Father-Mother-Son. However, the dataset can be easily reorganized into four different relationships, and use it as a bi-subject setup.

TABLE 2
Kinship statistics of UvA-NEMO Smile database.

	Spontaneous		Posed	
Relation	Subj. #	Vid. #	Sub. #	Vid. #
S-S	7	22	9	32
B-B	7	15	6	13
S-B	12	32	10	34
M-D	16	57	20	76
M-S	12	36	14	46
F-D	9	28	9	30
F-S	12	38	19	56
All	75	228	87	287

1.4 UBKinFace dataset

The *UB KinFace* dataset [6] consists of 600 face images of 400 different people corresponding to 200 kin relationships. Each relationship is composed of a child, a young parent and an old parent image, allowing the comparison of kinship verification with images of similar and different ages. The database is also divided in two ethnicity groups: Asian and not Asian. The images are not cropped from the same photographs an the are obtained in an uncontrolled environment with different cameras, at different times. The subjects depicted in the database are famous and well known public figures and their relatives, which makes it less suitable for human assessment.

1.5 Family101

The *Family101* dataset [7] is a large-scale dataset of families composed by well-known public figures across several generations. It contains 101 different families with distinct family names, including 206 groups, 607 individuals and a total of 14,816 images. The resolution of the face images is normalized to 150×120 pixels.

The dataset was assembled using crowdsourcing techniques and identity post-verification. In total the, 206 family groups contain images from at least both parents and one child, so they are componsed by a number of subjects that varies between 3 and 9. The final dataset contains different ethnicities with a proportion of 72% Caucasians, 23% Asians, and 5% African Americans. The database includes a bi-subject testing protocol containing 213 father-son relations, 147 father-daughter relations, 184 mother-son relations, and 148 mother-daughter relations.

1.6 Cornell-Kin

The *Cornell-Kin* dataset [8] is a small subset of this *Family101* dataset published prior to it, that contains only 143 relationship pairs cropped to a resolution of 100×100 pixels.

2 Appendix B: Kinship verification by machines

Based on our prior work [9], we propose a hybrid methodology for kinship verification from facial images and videos that exploits the complementarity of deep and shallow features. Our proposed approach consists on five main steps. It starts with detecting, segmenting and aligning the face images based on eye coordinates and other facial landmarks. Then, two types of descriptors are extracted: shallow spatio-temporal texture features and deep features. As spatio-temporal features, we extract local binary patterns (LBP) [10], local phase quantization (LPQ) [11] and binarized statistical image features (BSIF) [12]. These features are all extracted from Three Orthogonal Planes (TOP) of the videos. Deep features are extracted by convolutional neural networks (CNNs) [13]. 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, the four different kin relations are treated differently during the model training.

2.1 Preprocessing and detection of facial landmarks

To mitigate the influence of possible inconsistent color and pose across the images and videos included in the database, the first step of out approach consists in segmenting the face region from each video sequence. For that purpose, we have employed an active shape model (ASM) based approach that detects 68 facial landmarks and is able to track them along the video. The regions containing faces are then cropped from every frame in the video using the detected landmarks. Finally, The face-regions are aligned using key landmark points, registering them to a predefined template that preserves the interpupillary distance.

2.2 Feature extraction

Most of the existing work proposed to solve the automatic kinship verification problem focus on the extraction of shallow handcrafted features from still face images. However, our methodology approaches this problem from a hybrid perspective that takes advantage of both the spatio-temporal information contained in videos and of the facial similarity features learnt by deep learning approaches.

2.2.1 Spatio-temporal features

Spatio-temporal texture features have been shown to be efficient for describing faces in various face analysis tasks, such as face recognition and facial expression classification. In this work, we extract three local texture descriptors: LBP, LPQ and BSIF.

These three features are able to describe an image using a histogram of decimal values. The code corresponding to each pixel in the image is computed from a series of binary responses of the pixel neighborhood to a filter bank. In LBP and LPQ the filters are handcrafted while the filters of BSIF are learned from natural images. Specifically, the binary code of a pixel in LBP is computed by thresholding its value with the circularly symmetric P neighboring pixels (on a circle of radius R). LPQ encodes the local phase information of four frequencies of the short term Fourier transform (STFT) over a local window of size $W \times W$ surrounding the pixel. BSIF binarizes the responses of f independent filters of size $W \times W$ learnt by independent component analysis (ICA).

The spatio-temporal textural dynamics of the face in a video are extracted from three orthogonal planes XY, XT, and YT, separately. X and Y are the horizontal and vertical spatial axes of the video, and T refers to the time. The texture features of each plane are aggregated into a separate histogram. Then the three histograms are concatenated into a single feature vector.

To take benefit of the multi-resolution representation [14], the three features are extracted at multiple scales, varying their parameters. For the LBP descriptor, the selected parameters are are $P = \{8, 16, 24\}$ and $R = \{1, 2, 3\}$. For LPQ and BSIF descriptors, the filter sizes were selected as $W = \{3, 5, 7, 9, 11, 13, 15, 17\}$.

2.2.2 Deep learning features

Deep neural networks have been recently outperforming the state of the art in various classification tasks. Particularly, convolutional neural networks (CNNs) demonstrated impressive performance in object classification in general and face recognition in particular. However, deep neural networks require a huge amount of training data to learn efficient features.

Currently available kinship databases do not contain enough data samples to learn meaningful features. Preliminary experiments [9] using a Siamese CNN architecture as well as deep architecture [15] resulted in lower performance than using simple shallow features, probably due to the lack of enough training data.

An alternative for extracting deep face features is to use a pre-trained network. A number of very deep pre-trained architectures has already been made available to the research community. Motivated by the biological similarities between the face recognition and kinship verification problems, were the goal is to compute common features in two facial representations, we use a deep-learnt feature representation designed for face recognition, the VGG-face [13] network.

VGG-face has been initially trained for face recognition on a reasonably large dataset of 2.6 million images of over 2622 people. This network has been evaluated for face verification from both pairs of images and videos showing state of the art performance. The detailed parameters of the VGG-face CNN are provided by Table 3.

In VGG-face, the input of the network is an RGB face image of size 224×224 pixels. The network is composed of

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name		conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support		3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim		3		64			64		128			128		256		256			256
num filts		64		64			128		128			256		256		256			512
stride		1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad		1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim		512		512			512		512		512			512		4096		4096	
num filts		512		512			512		512		512			4096		4096		2622	
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

TABLE 3
VGG-face CNN architecture.

13 linear convolution layers *(conv)*, each followed by a non-linear rectification layer *(relu)*. Some of these rectification layers are followed by a non-linear max pooling layer *(mpool)*. Following are two fully connected layers *(fc)* both outputting a vector of size 4096. At the top of the initial network are a *fully connected* layer with the size of classes to predict (2622) and a *softmax* layer for computing the class posterior probabilities.

In this context, 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.

2.3 Classification

To classify a pair of face features as positive (the two persons have a kinship relation) or negative (no kinship relation between the two persons), we use a bi-class linear Support Vector Machine classifier (SVM). SVM imposes that each pair of features is transformed into a single feature vector. We use the normalized absolute difference transformation, where a pair of feature vectors $X = \{x_1, \ldots, x_d\}$ and $Y = \{y_1, \ldots, y_d\}$ is represented by the vector $F = \{f_1, \ldots, f_d\}$ as follows:

$$f_i = \sum_j \frac{|x_j - y_j|}{\sum_j (x_j + y_j)} \quad (1)$$

2.4 Score-level fusion

In order to check their complementarity, we have fused both spatio-temporal and deep facial similarity features. Preliminary experiments lead us to empirically find that simple score level fusion performs better than feature fusion. In this context, we have opted for training separate SVM classifiers for the different types of features extracted from each face video, performing the fusion by using a simple sum of the score-level results.

REFERENCES

[1] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 2, pp. 331–345, 2014.

[2] A. Dehghan, E. Ortiz, R. Villegas, and M. Shah, "Who do i look like? determining parent-offspring resemblance via gated autoencoders," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 1757–1764.

[3] H. Dibeklioglu, A. Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 1497–1504.

[4] H. Dibeklioglu, A. Salah, and T. Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *Computer Vision ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7574, pp. 525–538. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33712-3_38

[5] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: understanding the core of a family," *Multimedia, IEEE Transactions on*, vol. 17, pp. 1855–1867, 2015.

[6] S. X. M. Shao and Y. Fu, "Genealogical face recognition based on ub kinface database," in *Proc. IEEE CVPR Workshop on Biometrics (BIOM)*, 2011.

[7] R. Fang, A. Gallagher, T. Chen, and A. Loui, "Kinship classification by modeling facial feature heredity," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 2983–2987.

[8] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1577–1580.

[9] E. Boutellaa, B. López, S. M. Ait-Aoudia, X. Feng, and A. Hadid, "Kinship verification from videos using texture spatio-temporal features and deep learning features," in *International Conference on Biometrics (ICB'16)*, 2016.

[10] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Dec 2006.

[11] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[12] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *International Conference on Pattern Recognition (ICPR)*, 2012, pp. 1363–1366.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[14] C. H. Chan, M. Tahir, J. Kittler, and M. Pietikainen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 5, pp. 1164–1177, May 2013.

[15] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang, "Kinship verification with deep convolutional neural networks," in *Proc. British Machine Vision Conference (BMVC)*, September 2015, pp. 148.1–148.12.