Audio-Visual Kinship Verification in the Wild

Xiaoting Wu[1,2], Eric Granger[3], Tomi H. Kinnunen[4], Xiaoyi Feng[2], and Abdenour Hadid[1]

[1]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[2]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
[3]Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), Dept. of Systems Engineering, Ecole de technologie supérieure, Montreal, Canada
[4]School of Computing, University of Eastern Finland, Joensuu, Finland

Abstract

Kinship verification is a challenging problem, where recognition systems are trained to establish a kin relation between two individuals based on facial images or videos. However, due to variations in capture conditions (background, pose, expression, illumination and occlusion), state-of-the-art systems currently provide a low level of accuracy. As in many visual recognition and affective computing applications, kinship verification may benefit from a combination of discriminant information extracted from both video and audio signals. In this paper, we investigate for the first time the fusion audio-visual information from both face and voice modalities to improve kinship verification accuracy. First, we propose a new multi-modal kinship dataset called TALking KINship (TALKIN), that is comprised of several pairs of video sequences with subjects talking. State-of-the-art conventional and deep learning models are assessed and compared for kinship verification using this dataset. Finally, we propose a deep Siamese network for multi-modal fusion of kinship relations. Experiments with the TALKIN dataset indicate that the proposed Siamese network provides a significantly higher level of accuracy over baseline uni-modal and multi-modal fusion techniques for kinship verification. Results also indicate that audio (vocal) information is complementary and useful for kinship verification problem.

1. Introduction

Kinship verification from facial images has been an active research topic in computer vision and machine learning since 2010 [11]. It involves recognizing whether the facial image of two individuals have a kin relationship, and finds several applications, such as family album organization, creation of family trees and image annotation, etc. It can also be employed to find missing children even when their appearance changes over time as a person ages.

Researchers from the University of Nottingham carried out a pilot study on heritability of human voice parameters[1] to determine how the human voice is passed down through the generation, and which factors determine our voice. Inspired by this study, we investigate (for the first time) the used of audio signals as an additional source of information for kinship verification, to improve the verification accuracy. The use of kinship relation from voice has received very little attention in literature – some studies have addressed potential performance degradation of automatic speaker verification when tested with voice of persons with close kinship relation, such as identical twins [2, 15]. Fusing both face and voice modalities captured in video sequences may nonetheless help to improve the accuracy and robustness of kinship verification systems. Moreover, verifying kinship based on multiple (face and voice) modalities, rather than facial stills alone, can has several potential applications, such as social media analysis and detecting kidnapping victims in public places.

In this paper, we investigate kinship verification systems that allow for fusions of face and voice modalities to encode a discriminative kin information. We first propose a new kinship dataset called *TALking KINship* (TALKIN) that consists of visual (face) and audio (voice) information on several individuals talking in videos. Then a robust multi-modal fusion method for kinship verification is proposed and compared with uni-modal performance and several baseline methods. State-of-the-art conventional and

978-1-7281-3640-0/19/$31.00 ©2019 IEEE

[1]https://www.nottingham.ac.uk/news/pressreleases/2016/january/help-the-scientists-find-out-why-you-sound-like-your-parents.aspx

deep learning models are assessed and compared for kinship verification using this dataset. Finally, we propose a deep Siamese network for metric learning of multi-modal kinship verification, based on pair-wise similarities and contrastive loss. Siamese architectures contain identical sub-networks with the same configurations, parameters and weights. They are promising for multi-modal fusion because fewer parameters required optimization, and thereby limit over-fitting [28].

2. Related Work

2.1. Kinship verification

Methods proposed in literature for kinship verification from faces can be divided into two categories: feature-based and similarity-based (or metric-based) methods.

Feature-based methods try to assemble the most discriminant kin features from facial images to improve the accuracy. The traditional face representations are handcrafted features, such as Binarized Statistical Image Feature (BSIF) [13], Local Phase Quantization (LPQ) [22], Local Binary Patterns (LBP) [1, 21] and LBP-TOP [39]. In the first attempt at kinship verification using facial images, a subset of local and global features containing the most kin information were extracted to represent faces [11]. Wu *et al.* [32] studied the utility of color information for this task, while Cui *et al.* [6] automatically selects discriminative feature patches to be processed using AdaBoost to improve the performance, because some facial regions do not represent useful information. Deep learning methods in literature provide a high level of accuracy. For instance, Zhang *et al.* [37] used a convolutional NNs (CNN) to learn discriminant feature representations to verify the kinship of an image pair. To reduce the gap across generations, a hierarchical representation learning method using a deep belief network (DBN) was proposed to encode deep embeddings for kinship verification [14].

Similarity-based methods have also provides good performance in kinship verification. Lu *et al.* [19] proposed a multi-view neighborhood repulsed metric learning (MNRML) method that learns a distance metric by projecting image pairs with kin relation as close as possible, and image pairs without kin relation as far as possible [19, 31]. Yan *et al.* [35] learning distance metrics for faces captured in videos. Lu *et al.* [18] proposed a discriminative deep metric learning (DDML) method for face and kinship verification, and Wang *et al.* [31] proposed the cross-generation method with sparse discriminative metric loss (SDN-Loss) to reduce the margin between both age and identity.

2.2. Multi-modal methods

Multi-modal fusion methods has successfully improved the recognition accuracy in many applications found in affective computing [29], person recognition [4], large-scale video classification [17] and gesture recognition [20] because they can exploit complementary sources of information. Different sources of information are typically integrated through early fusion (feature level) or through late fusion (score or decision levels) [3]. Feature-level fusion using concatenation or aggregation (e.g., canonical correlation analysis) is often considered to provide the high level of accuracy, although feature patterns may also be incompatible and increase system complexity. Techniques for score-level fusion using deterministic (e.g., average fusion) or learned functions are commonly employed, but are vulnerable to the impact of score normalization methods on the overall decision boundaries, and the availability of representative training samples. Despite reducing the information content about modalities, techniques for decision-level fusion (e.g., majority voting) can provide a simple framework for combination, although limitations are placed on decision boundaries due to the restricted operations that can be performed on binary decisions.

In deep learning literature, Neverova *et al.* [20] proposed a multi-scale and multi-modal early fusion method – called Multimodal Dropout (ModDrop) – for gesture recognition problems. First, the weights of each uni-modal are pretrained, and then the shared hidden and output layer allow combining many modalities. Liu *et al.* [17] introduced the multi-modal factorized bi-linear pooling (MFB) [36] method for fusing visual and audio representations for video-based classification. In affective computing applications, Tzirakis *et al.* [29] proposed an end-to-end multimodal deep NN for emotion recognition. For the visual modality and speech modality, they are first trained separately to speed up the fusion training phase. Then the fusion network is trained end-to-end for affective computing. For the late fusion methods, Chowdhury *et al.* [4] collected a new audio-visual dataset, called the MSU Audio-Video Indoor Surveillance (MSU-AVIS) dataset, for person recognition. They implemented the state-of-the-art methods for person recognition based on face and speech moralities.

From the work reviewed above, multi-modality methods has been shown to provide sufficient improvements in system accuracy and robustness compared with uni-modals. To improve the verification accuracy of kinship verification, we investigate algorithms for the fusion of face and voice modalities extracted from videos for accurate kinship verification. As far as we know, this is the first attempt to study the kinship verification from both visual and audio information.

The main contribution of this paper are the following. First, to study the multi-modal kinship verification based on videos, the new TALKIN dataset is established that consists of both visual (face) and audio (vocal) modalities recorded from several individuals. Then, uni-modal and multi-modal

Table 1. Main characteristics of publicly available kinship datasets.

Dataset		Modalities	Size	Resolution ratio	Controlled environment
Cornell KinFace [11], 2010		Image	150 pairs	100×100	No
UB KinFace [34][33][27], 2011		Image	200 groups	89×96	No
UvA-NEMO Smile [8][9], 2012		Video	1240 videos	1920×1080	Yes
Family101 [10], 2013		Image	14816 images	120×150	No
KinFaceW [19], 2014	KinFaceW-I	Image	533 pairs	64×64	No
	KinFaceW-II	Image	1000 pairs	64×64	No
TSKinFace [24], 2015		Image	1015 tri-subjects	64×64	No
KFVW [35], 2017		Video	418 pairs of videos	about 900×500	No
FIW [26], 2018		Image	1000 family trees	224×224	No
TALKIN (ours)		Video & Audio	400 pairs of videos	about 1920×1080	No

fusion methods for kinship verification are compared with this datasets. Finally, we propose a new deep Siamese network that is suitable to assess pair-wise similarities for multi-modal kinship verification, based on backpropagation and contrastive loss. Siamese architectures perform metric learning using identical sub-networks with the same configurations, parameters and weights. They are promising for multi-modal fusion because fewer parameters required optimization, and thereby limit over-fitting [28].

3. TALKIN dataset

In this section, a new kinship dataset called TALKIN is described. It contains several videos of subjects talking in the wild environment (under unconstrained background, illumination and recording condition, *et al.*). The purpose of collecting it is to investigate the newly raised problem, multi-modal kinship verification in the wild. A comparison of TALKIN with existing kinship datasets is shown in Table 1.

3.1. Data collection pipeline

The overall collection pipeline for the TALKIN dataset is showed in Fig. 1.

Step 1. List of celebrities or family TV shows. The first step is to prepare a list of celebrities from which we intend to obtain videos. The target of the amount for each relation is 100 pairs of videos. Most of the list is formed by celebrities, such as musicians, actors, politician, etc., with the reminder from TV series involving family interactivity (non-celebrities).

Step 2. Downloading YouTube videos. Videos were downloaded from YouTube[2] by searching the name of celebrities or TV series. We collect parent's videos and child's videos from *different* video clips corresponding to different backgrounds or recording conditions.

Step 3. Data preparation. For the face detection and alignment, we use the MTCNN algorithm [38] to detect 5 face landmarks in every frame of the video. Finally, the videos are cropped according to the landmarks. The face frames are re-sized into 224×224. Both hand-crafted features and deep features are extracted to represent each individual. We directly extract audio from the video clips. Standard methods in the speech field, Mel-Frequency Cepstral Coefficients (MFCCs) [30] and Deep Neural Networks are used to embed the audio features.

3.2. Parameters of the dataset

The TALKIN dataset contains four kin relations: Father-Son (FS), Father-Daughter (FD), Mother-Son (MS) and Mother-Daughter (MD), with 100 pairs of videos (with audio) for each relation. As all the data originates from uncontrolled Internet resources, the speech contents vary from subject to subject and video to video, making the voice-related sub-task *text-independent kinship verification*, analogous with text-independent speaker verification. That is, the task is to verify kinship relations regardless of what was said between individuals.

TALKIN incorporates a wide range of backgrounds, recording environments, poses, occlusions and ethnicities. Table 2 shows the distribution of ethnicity in TALKIN. The distribution is count by kin pair rather than individuals, in case that one parent might appear multiple times with more than one kid. Note, however, that we exclude mixed-race trials, *i.e.* the parent and child in a trial has the same ethnicity. The dataset has two parts: video and audio. The

[2]YouTube is a popular US-based video-sharing website `https://www.youtube.com/`

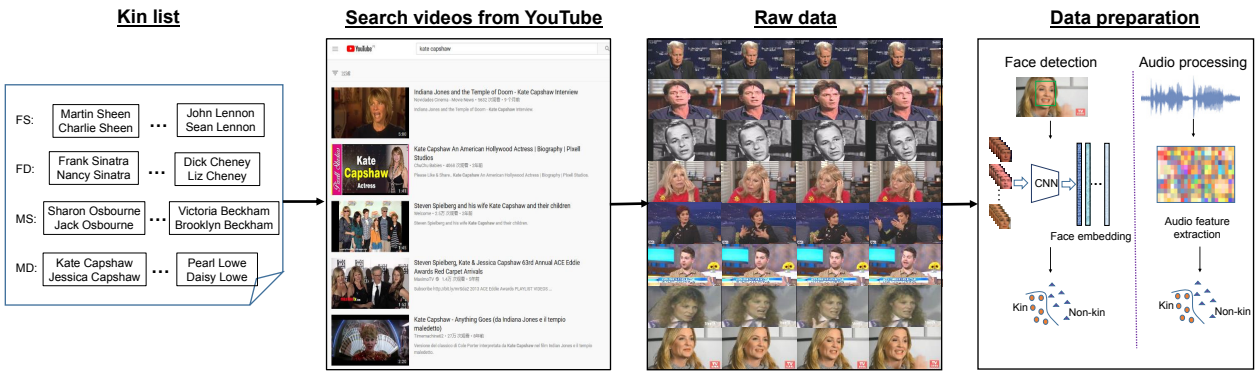


Figure 1. The collection pipeline TALKIN dataset.

Table 2. The ethnicity distribution (%) of TALKIN dataset.

British	American	French	Australian	Chinese	Dutch	Italian	Swedish	Turkish
56.50	33.50	6.50	2.00	0.50	0.25	0.25	0.25	0.25

length of the video varies from 4.032 seconds to 15 seconds with a resolution of about 1920×1080. Audio is extracted from video files. The sample rate are all set with 44.1 kHz. Besides the varied text content, the audio files contain substantial channel variations (*e.g.* due to differing recording devices). Some of them also contain reverberation and additive noise.

4. A Siamese Network for A-V Fusion

This section presents a new deep Siamese network for the fusion of face and voice modalities for accurate multimodal kinship verification. It is trained to evaluate pair-wise similarities based on face and voice modalities. In a particular implementation, we fine-tune the VGG-Face [23] CNN cascaded with an Long Short-Term Memory (LSTM) [12] network for the face modality. For the voice modality extracted from videos, we fine-tune a ResNet-50 pre-trained on VoxCeleb2 [5]. Finally, a fully connected (FC) layer is added to fuse the audio and visual information. During the training procedure, our system is trained on our dataset, using backpropagation and contrastive loss to learn the correlation between parent and child based on audio visual modalities.

4.1. Face network

We implement the VGG-face [23] CNN cascaded with an LSTM [12] network for the facial representations. VGG-Face network is trained on a large face dataset with 2.6 million images of over 2662 people. The input of the network is a RGB image with the size of $224 \times 224 \times 3$. As shown at the top of Fig. 2, it is comprised of 13 convolution layers, each followed by Rectified Linear Unit (ReLU). Some of them are also followed by max pooling operator. The last three layers are FC layers. The first two FC layers have 4096 outputs and last FC layer have N outputs as N-class prediction. We feed the facial frames one by one and collect the deep features from layer fc7. To integrate both spatio and temporal information, a layer LSTM with 4096 cells is stack on the top of it.

4.2. Voice network

In the previous research, the acoustic features are first extracted and machine learning methods such as I-vector [7] and Gaussian mixture model - universal background model (GMM-UBM) [25] are used to analysis features. In this work (see the top of Fig. 2), we use the ResNet-50 pre-trained with a large speaker verification dataset called VoxCeleb2 [5], and then fine-tuned with TALKIN data to get feature embedding from it for audio based kinship verification.

The audio samples are converted into single-channel and down sampled into 16 kHz to have the consistence with VoxCeleb2 dataset. Then the audio samples are segmented into 3 seconds. A hamming window with 25ms width and 10ms step is applied on the audio. Following the same manner of [5], spectrograms with the size of 512×300 can be computed, where 512 is the size of the spectrum and 300 is the number of frames. After performing mean and variance normalization, the spectrograms are fed into the ResNet-50.

4.3. Fusion network

We propose a deep Siamese network with contrastive loss [16] for kinship verification based on fusing videos and audio. The whole architecture is shown in Fig. 2. For each voice and face network, we use contrastive loss to learn the intra-class similarity and inter-class dissimilarity among

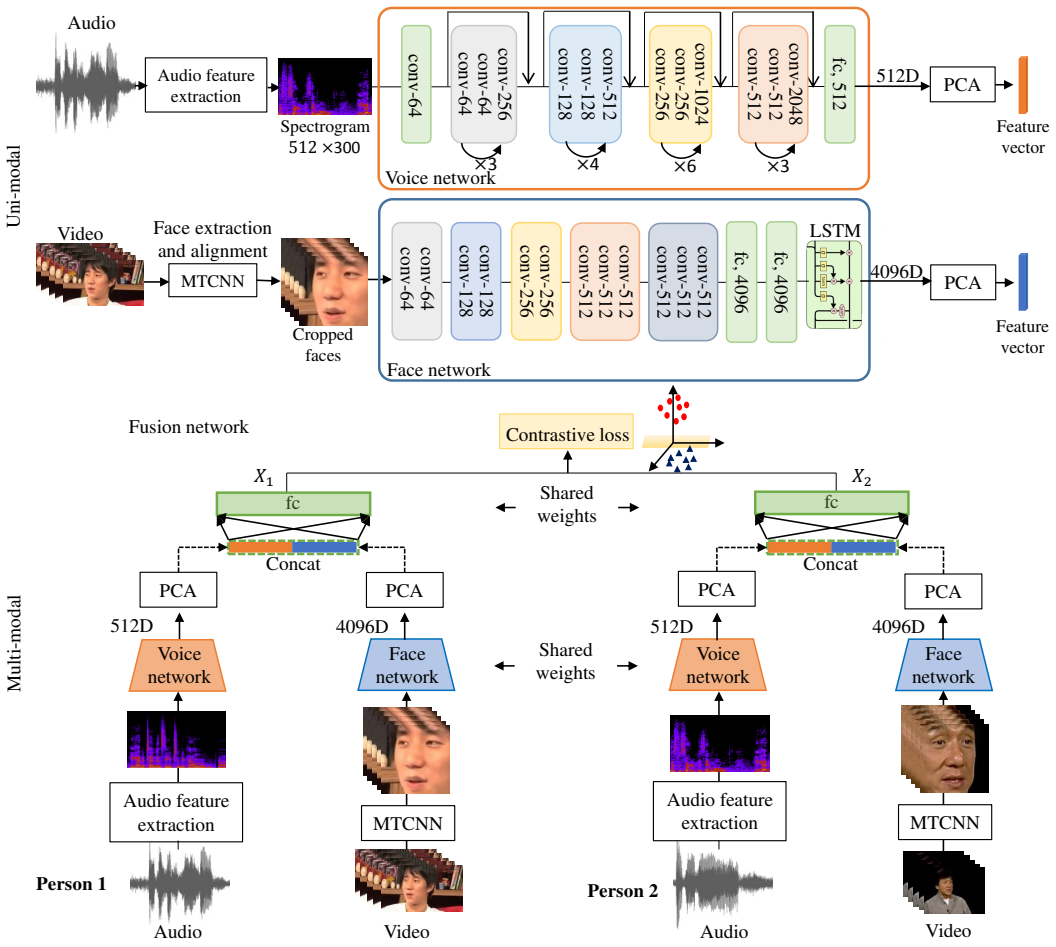


Figure 2. Architecture of the proposed fusion method.

subjects. The contrastive loss is defined as:

$$L = \frac{1}{2N}\sum_{n=1}^{N}(y_n d^2 + (1-y_n)\max(M-d,0)^2) \quad (1)$$

where threshold M is the margin, N is the batch size, $d = \|a_n - b_n\|^2$, a_n and b_n denote two sample features, y_n is the label of the sample pair. y_n equals 1 when the inputs have the kin relation and y_n equals 0 the otherwise.

After training the face and voice networks, we can collect their features – 4096D features are extracted from face network and 512D features are extracted from voice network. Then, after performing PCA on them to reduce the dimension into 130, they are concatenated into a 260D feature and followed by a FC layer with 260 nodes. By adding contrastive loss during fusion part, we can automatically learn the fusion rule for kinship verification to narrow the distance between pairs with a kin relation, and enlarge the distance between the negative pairs. After training the network, the feature extracted from the added FC layer is viewed as fusion feature of one facial video and audio signal. The cosine similarity $sim(x_1, x_2) = \frac{x_1 \cdot x_2}{\|x_1\| \cdot \|x_2\|}$ is calculated to represent the distance between two inputs (*e.g.* parent and child represented by feature vectors x_1 and x_2). A threshold applied to sim allow to determine whether two inputs have a kin relation.

5. Experimental results and analysis

5.1. Experimental setup and baseline methods

The TALKIN dataset was used to evaluate the performance of the baseline and proposed methods for uni-modal and multi-modal kinship verification. For each relation – FS, FD, MS and MD – there are 100 pairs of videos. We randomly generate 100 pairs of videos without kin relation as the negative pairs. Thus there are 100 pairs of positive pairs in total with kin relation, and 100 pairs of negative pairs without kin relation. Then 5-fold cross-validation is performed in our experiment. For both proposed uni-modal and multi-modal methods, we reserve 100% energy with the PCA operation.

Baseline kinship verification with face modality. We employed the following image-based feature representations: BSIF, LPQ and LBP. We averaged these frame-by-frame features to represent each video by a single feature vector. The facial frames are first converted into HSV color space [32] with size of 64 $\times$ 64 $\times$ 3. For BSIF feature extraction, images are divided into non-overlapping 32 $\times$ 32 blocks in each color channel. Each block is represented using 256 features and the whole face with 256 $\times$ 4 $\times$ 3 = 3072 features. For LPQ feature extraction, images are divided into non-overlapping 32 $\times$ 32 blocks in each color channel. Each block is represented using 256 features, leading to 3072-dimensional (256 $\times$ 4 $\times$ 3) feature representation for the whole face. For LBP feature extraction, the images are divided into non-overlapping 16 $\times$ 16 blocks in each color channel. The parameters of LBP is that the radius is set as 1, the sampling number is 8. 59 histogram values are used to represent each block. Thus, each facial image is represented using 59 $\times$ 16 $\times$ 3 = 2832 features. Furthermore, we also evaluate the video representation, LBP-TOP. In the experiment, the frames are converted into gray scale. Then the face frames are divided into 56 $\times$ 56 non-overlapping blocks. All features extracted from each block volume are connected to represent the appearance and motion of the kinship video. The radius is 1. For each block volume, we extract 59 histogram features in XY, XT and YT planes, respectively. Thus, one video can be represented as a 59 $\times$ 3 $\times$ 16 = 2832 face features. At last, we compute the cosine similarity between two facial features.

Baseline kinship verification with voice modality. We employ two baseline GMM-UBM and I-vector for audio processing. We extracted Mel-frequency cepstral coefficients (MFCCs) with 12 cepstral coefficients from the audio samples. The UBM with 128 mixture components of GMM is trained with training set. For GMM-UBM based method, the kin pair model is created from UBM using the Maximum A Posteriori (MAP) estimation. The verification likelihood is the log-likelihood ratios between speaker models and registered speaker's GMM. In the I-vector framework, UBM is trained using expectation-maximization (EM) with MFCCs. The GMM super-vector can be represented as Equation 2, where $\mathbf{s}$ is the super-vector of the input utterance, $\mathbf{m}$ is the UBM mean vector, $\mathbf{T}$ is the total-variability matrix, $\mathbf{w}$ is the I-vector. Then the dimension of I-vector is reduced by linear discriminant analysis (LDA). The similarity between two speakers is represented by cosine similarity score of their I-vectors.

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{2}$$

Baseline for multi-modal kinship verification. Two baseline methods for multi-modal kinship verification, early (feature) level and late (score) level fusion methods, are applied. For the early fusion method, after extracting features from face and voice network, Principal Component Analysis (PCA) is used to make it consistent size for video and audio. Then the video and audio features are concatenated together into one feature vector as the fused feature. For the late fusion method, the evaluation for the video based and audio based kinship verification are performed separately. Then, the averaged score is selected as the fused score. When reduce feature dimension with PCA, we preserve 100% energy during both feature fusion and score fusion procedure.

Table 3. Verification accuracy (%) for the face modality on TALKIN dataset

Techniques	FS	FD	MS	MD	Average
BSIF-Average [13]	61.5	58.5	61.0	59.5	60.1
LPQ-Average [22]	62.5	58.0	60.5	59.0	60.0
LBP-Average [1, 21]	61.5	60.0	59.5	61.5	60.6
LBP-TOP [39]	64.5	60.0	67.0	59.5	62.8
VGG + LSTM	**76.5**	**69.5**	**70.0**	**71.5**	**71.9**

Table 4. Verification accuracy (%) for the voice modality on TALKIN dataset.

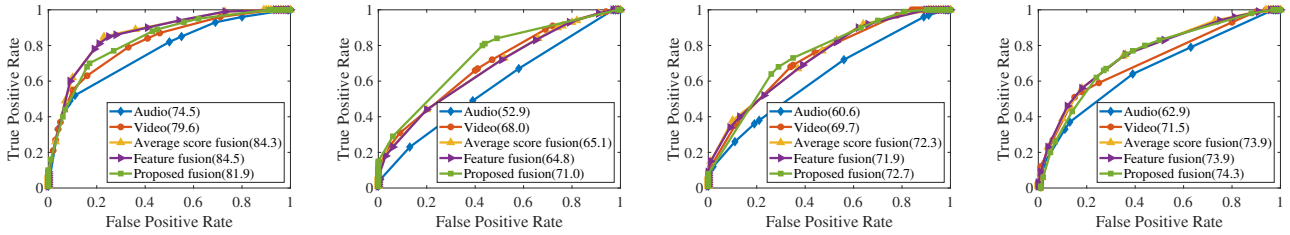
Techniques	FS	FD	MS	MD	Average
I-vector [7]	63.5	60.0	63.0	63.0	62.4
GMM-UBM [25]	59.5	59.5	**66.5**	60.0	61.4
Resnet-50	**73.0**	**60.0**	63.5	**66.5**	**65.8**

Table 5. Verification accuracy (%) from uni-modal and multi-modal techniques on TALKIN dataset.

Techniques	FS	FD	MS	MD	Average
Resnet-50 (audio)	73.0	60.0	63.5	66.5	65.8
VGG+LSTM (video)	76.5	69.5	70.0	71.5	71.9
Late fusion	82.5	67.0	69.0	**73.0**	73.1
Early fusion	**83.0**	67.5	69.5	**73.0**	73.3
Deep Siamese Network (ours)	80.0	**70.5**	**73.5**	72.5	**74.1**

5.2. Results and analysis

Uni-modal kinship verification. Table 3 and Table 4 present the results of experiments for uni-modal kinship verification from voice and face modalities, respectively. VGG-Face with LSTM shows better performance compared with traditional hand-crafted features. For voice based kinship verification, Resnet-50 has better performance, except for mother-son relation, which has a small drop compared



(a) FS relation (b) FD relation (c) MS relation (d) MD relation

Figure 3. ROC curves uni- and multi-modal techniques for kinship verification on TALKIN dataset. The numbers in parentheses are the Area Under the ROC Curve (%) for each method.

with GMM-UBM method. Overall, the proposed methods for uni-modal based kinship verification show the efficiency with contrast of baseline methods.

Multi-modal kinship verification. Table 5 compares uni-modal based kinship verification with two baseline fusion methods and proposed deep Siamese network method. The corresponding ROC curves are shown in Fig. 3. Compared with uni-modals, feature fusion method and score fusion method improve the accuracy from the average aspect, where both has comparable accuracy as video modal in father-daughter and mother-son relations. The proposed Siamese network shows a higher level of accuracy compared with uni-modals and baseline fusion methods. The average accuracy is improved about 3.8% from uni-modals and 1.0% from baseline fusion methods, where feature fusion method has best performance in father-son relation and both feature fusion and score fusion method has highest accuracy in mother-daughter relation.

6. Conclusion

In this paper, we study for the first time the fusion of audio-visual information from both face and voice modalities for kinship verification. A new TALKIN dataset is proposed for multi-modal kinship verification, and used to compare the proposed and baseline models in this paper. A deep Siamese network for multi-modal fusion is also proposed for metric learning of kinship verification. Experiments indicate that the proposed Siamese network improves accuracy over baseline uni-modal and multi-modal fusion techniques for kinship verification. Additionally, the audio (vocal) information is shown to be complementary and useful for kinship verification problem.

In the future work, we plan to extend deep learning architectures to implement the uni-modal models (face and voice networks), and the Siamese network to further improving verification accuracy. In order to train more discriminant and robust deep networks, we will enlarge the TALKIN dataset. Other loss functions (such as triplet and magnate loss) should be investigated for the Siamese metric learning, where the loss seeks to discriminate between the positive pair of matching person from the negative non-matching person.

Acknowledgment

This work is partially supported by the China Scholarship Council (grant 201706290103), the Academy of Finland, and the Natural Sciences and Engineering Research Council of Canada. The authors wish to acknowledge CSC-IT Center for Science, Finland, for the computational resources. The initial help from Dr. Miguel Bordallo López and Dr. Elhocine Boutellaa is also acknowledged.

References

[1] 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*, 28(12):2037–2041, 2006.

[2] A. Ariyaeeinia, C. Morrison, A. Malegaonkar, and S. Black. A test of the effectiveness of speaker verification for differentiating between identical twins. *Science & Justice: Journal of the Forensic Science Society*, 48(4):182–186, Dec. 2008.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[4] A. Chowdhury, Y. Atoum, L. Tran, X. Liu, and A. Ross. Msu-avis dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos. In *ICPR 2018*.

[5] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*.

[6] L. Cui and B. Ma. Adaptive feature selection for kinship verification. In *ICME 20187*.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[8] H. Dibeklioğlu, A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *ECCV 2012*.

[9] H. Dibeklioğlu, A. Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *ICCV 2013*.

[10] R. Fang, A. Gallagher, T. Chen, and A. Loui. Kinship classification by modeling facial feature heredity. In *ICIP 2013*.

[11] R. Fang, K. D. Tang, N. Snavely, and T. Chen. Towards computational models of kinship verification. In *ICIP 2010*.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *ICPR 2012*.

[14] N. Kohli, M. Vatsa, R. Singh, A. Noore, and A. Majumdar. Hierarchical representation learning for kinship verification. *IEEE Transactions on Image Processing*, 26(1):289–302, 2017.

[15] H. Künzel. Automatic Speaker Recognition of Identical Twins. *International Journal of Speech Language and the Law*, 17(2), Feb. 2011.

[16] L. Li, X. Feng, X. Wu, Z. Xia, and A. Hadid. Kinship verification from faces via similarity metric based convolutional neural network. In *International Conference Image Analysis and Recognition*, pages 539–548. Springer, 2016.

[17] J. Liu, Z. Yuan, and C. Wang. Towards good practices for multi-modal fusion in large-scale video classification. In *European Conference on Computer Vision*, pages 287–296. Springer, 2018.

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

[19] 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*, 36(2):331–345, 2014.

[20] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[21] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[22] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and Signal Processing*, volume 5099, pages 236–243, 2008.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conf.*, 2015.

[24] X. Qin, X. Tan, and S. Chen. Tri-subject kinship verification: Understanding the core of a family. *arXiv preprint arXiv:1501.02555*, 2015.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1):19–41, Jan. 2000.

[26] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[27] M. Shao, S. Xia, and Y. Fu. Genealogical face recognition based on ub kinface database. In *CVPRw 2011*.

[28] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV 2015*.

[29] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[30] R. Vergin, D. O'shaughnessy, and A. Farhat. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on speech and audio processing*, 7(5):525–532, 1999.

[31] S. Wang, Z. Ding, and Y. Fu. Cross-generation kinship verification with sparse discriminative metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[32] X. Wu, E. Boutellaa, M. B. López, X. Feng, and A. Hadid. On the usefulness of color for kinship verification from face images. In *WIFS 2016*.

[33] S. Xia, M. Shao, and Y. Fu. Kinship verification through transfer learning. In *IJCAI 2011*.

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

[35] H. Yan and J. Hu. Video-based kinship verification using distance metric learning. *Pattern Recognition*, 2017.

[36] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV 2017*.

[37] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang. Kinship verification with deep convolutional neural networks. In *BMVC 2015*.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[39] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.