OULU 2022
C 844

ACTA UNIVERSITATIS OULUENSIS

Xiaoting Wu

MACHINE LEARNING FOR AUDIO-VISUAL KINSHIP VERIFICATION

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

ACTA UNIVERSITATIS OULUENSIS
C Technica 844

XIAOTING WU

MACHINE LEARNING FOR AUDIO-VISUAL KINSHIP VERIFICATION

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in Auditorium IT116, Linnanmaa, on 21 October 2022, at 12 noon

UNIVERSITY OF OULU, OULU 2022

Acta Univ. Oul. C 844, 2022

Supervised by
Assistant Professor Li Liu
Associate Professor Miguel Bordallo López

Reviewed by
Associate Professor Vishal M. Patel
Associate Professor Aythami Morales Moreno

Opponent
Professor Karen Eguiazarian

ISBN 978-952-62-3423-6 (Paperback)
ISBN 978-952-62-3424-3 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

PUNAMUSTA
TAMPERE 2022

Wu, Xiaoting, Machine learning for audio-visual kinship verification
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
Acta Univ. Oul. C 844, 2022
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Human faces implicitly indicate the family linkage, showing the perceived facial resemblance in people who are biologically related. Psychological studies found that humans have the ability to discriminate the parent-child pairs from unrelated pairs, just by observing facial images. Inspired by this finding, automatic facial kinship verification has emerged in the field of computer vision and pattern recognition, and many advanced computational models have been developed to assess the facial similarity between kinship pairs. Compared to human perception ability, automatic kinship verification methods can effectively and objectively capture subtle kin similarities such as shape and color. While many efforts have been devoted to improving the verification performance from human faces, multimodal exploration of kinship verification has not been properly addressed. This thesis proposes, for the first time, the combination of human faces and voices to verify kinship, which is referred to as audio-visual kinship verification, establishing the first comprehensive audio-visual kinship datasets, which consist of multiple videos of kin-related people speaking to the camera. Extensive experiments on these newly collected datasets are conducted, detailing the comparative performance of both audio and visual modalities and their combination using novel deep-learning fusion methods. The experimental results indicate the effectiveness of the proposed methods and that audio (voice) information is complementary and useful for the kinship verification problem.

Keywords: audio-visual fusion, datasets, deep learning, kinship verification, texture analysis

Wu, Xiaoting, Koneoppiminen audiovisuaalisen sukulaisuuden todentamiseen
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta
Acta Univ. Oul. C 844, 2022
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Ihmiskasvot osoittavat implisiittisesti perhesidonnaisuuden, mikä osoittaa biologisesti sukua olevien ihmisten koettua kasvojen samankaltaisuutta. Psykologiset tutkimukset havaitsivat, että ihmisillä on kyky erottaa vanhempi-lapsi-parit toisistaan riippumattomista pareista pelkästään kasvojen kuvien avulla. Tämän löydön innoittamana automaattinen kasvojen sukulaisuuden todentaminen on syntynyt tietokonenäön ja hahmontunnistuksen alalla, ja monia kehittyneitä laskennallisia malleja on kehitetty arvioimaan kasvojen samankaltaisuutta sukulaisparien välillä. Verrattuna ihmisen havainnointikykyyn automaattiset sukulaisuuden todentamismenetelmät voivat tehokkaasti ja objektiivisesti havaita hienovaraisia sukulaisyhteyksiä, kuten kasvojen muotoa ja ihonväriä. Vaikka monia ponnisteluja on tehty pyrkimyksenä parantaa ihmiskasvojen todentamista, sukulaisuuden todentamisen multimodaalista tutkimista ei ole käsitelty kunnolla. Tässä opinnäytetyössä ehdotetaan ensimmäistä kertaa ihmiskasvojen ja äänen yhdistämistä sukulaisuuden todentamiseksi tavalla, jota kutsutaan audiovisuaaliseksi sukulaisuustodentamiseksi. Näin luodaan ensimmäiset kattavat audiovisuaaliset sukulaisuustietojoukot, jotka koostuvat useista videoista, joissa esiintyy kameralle puhuvia sukulaisia. Näillä äskettäin kerätyillä tietojoukoilla tehdään laajoja kokeita, joissa kuvataan yksityiskohtaisesti sekä ääni että visuaalisten modaliteettien vertailevaa suorituskykyä ja niiden yhdistelmää käyttämällä uusia syvän oppimisen fuusiomenetelmiä. Kokeelliset tulokset osoittavat ehdotettujen menetelmien tehokkuuden ja sen, että ääni- (ääni)informaatio on täydentävää ja hyödyllistä sukulaisuuden todentamisongelmassa.

Asiasanat: audiovisuaalinen fuusio, sukulaisuuden todentaminen, syväoppiminen, tekstuurianalyysi, tietojoukot

To my parents and husband.

Acknowledgements

This thesis work was carried out at the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu. Here, I wish to express my sincere gratitude to all of them who have contributed to my thesis and supported me during these years.

First, I want to express my deep gratitude to my supervisors Dr. Li Liu and Dr. Miguel Bordallo López for supervising my thesis work, who have helped me a lot in my doctoral study and research. During this journey, my supervisors have guided me with patience and always encouraged me when I encountered obstacles. I also want to thank Prof. Olli Silvén for his kind and considerate support in completing my doctoral thesis.

I would also like to convey my appreciation to the members of my follow-up group, Dr. Xiaopeng Hong and Dr. Xiaobai Li, for their unsparing and concrete feedback and comments on my doctoral study progress. I am grateful to all my co-authors for their constructive comments and contributions. I also would like to extend my thanks to the participants of the TALKIN-Family dataset. The dataset collection contribution from Mr. Xueyi Zhang and Mr. An Huang is appreciated.

I would like to express my gratitude to Prof. Vishal M. Patel (Johns Hopkins University, USA) and Prof. Aythami Morales Moreno (Universidad Autónoma de Madrid, Spain) for reviewing this thesis and providing very useful comments. I would also like to thank Prof. Karen Eguiazarian (Tampere University, Finland) for acting as an opponent for the thesis.

The financial support from the CMVS, China Scholarship Council, and the Academy of Finland is gratefully acknowledged as well. I also want to acknowledge CSC-IT Center for Science, Finland, for the computational resources.

I thank all past and present colleagues at the CMVS for the joyful and encouraging working atmosphere. I also thank the administrative staff who have always been very kind and helpful, especially Mr. Hannu Rautio for his help with working facilities.

I would like to thank all my friends for their accompany and encouragement, who have brought me so many delightful times. Last but not the least, my warmest and affectionate thanks go to my mom and dad for their unconditional love and support. My deepest gratitude goes to my husband, P. Engr. Jinlei Pan, for his understanding and continuous support.

In Oulu, Finland, 9th of May 2022
Xiaoting Wu

List of abbreviations

FKV	*Facial kinship verification*
DNN	*Deep Neural Networks*
CNN	*Convolutional Neural Networks*
SS	*Sister-Sister*
BB	*Brother-Brother*
BS	*Brother-Sister*
FS	*Father-Son*
FD	*Father-Daughter*
MS	*Mother-Son*
MD	*Mother-Daughter*
GFGS	*Grandfather-Grandson*
GFGD	*Grandfather-Granddaughter*
GMGS	*Grandmother-Grandson*
GMGD	*Grandmother-Granddaughter*
MFCCs	*Mel-frequency cepstral coefficients*
KNN	*K-Nearest Neighbor*
SVM	*Support Vector Machine*
LBP	*Local Binary Pattern*
LPQ	*Local Phase Quantization*
HOG	*Histogram of Oriented Gradients*
ELM	*Extreme Learning Machines*
BSIF	*Binarized Statistical Image Features*
DoG	*Differences of Gaussians*
PML-COV	*Pyramid Multi-level covariance descriptor*
SP-DTCWT	*Selective Patch-based Dual-Tree Complex Wavelet Transform*
WLD	*Weber's Local Descriptor*
mRMR	*Max-Relevance and Min-Redundancy*
NRML	*Neighborhood Repulsed Metric Learning*
ESL	*Ensemble Similarity Learning*
SMCNN	*Similarity Metric based Convolutional Neural Networks*
AE	*Auto-Encoders*
GANs	*Generative Adversarial Networks*
LBP-TOP	*Local Binary Pattern histograms from Three Orthogonal Planes*
SMNAE	*Supervised Mixed Norm AutoEncoder*

FIW	*Families In the Wild*
FMD	*Father-Mother-Daughter*
FMS	*Father-Mother-Son*
FIW	*Families In the Wild*
KFVW	*Kinship Face Videos in the Wild*
TP	*True Positive*
TN	*True Negative*
FP	*False Positive*
FN	*False Negative*
A	*Accuracy*
UAAML	*Unified Adaptive Adversarial Multimodal Learning*
LSTM	*Long Short-Term Memory*
FC	*Fully Connected*
ReLU	*Rectified Linear Unit*
PCA	*Principal Component Analysis*
MLP	*Multiple Layer Perceptron*
ResNet	*Residual network*
GMM-UBM	*Gaussian Mixture Model-Universal Background Model*
MAP	*Maximum A Posteriori*
EM	*Expectation-Maximization*
LDA	*Linear Discriminant Analysis*
ROC	*Receiver Operating Characteristic*

List of original publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals (I–VI).

I Wu X, Feng X, Cao X, Xu X, Hu D, Bordallo López M & Liu L (2022) Facial kinship verification: A comprehensive review and outlook. International Journal of Computer Vision (IJCV), vol. 130, no. 6, pp. 1494–1525.

II Wu X, Boutellaa E, Feng X & Hadid A (2016) Kinship verification from faces: Methods, databases and challenges. Proc. IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2016), Hong Kong, China, pp. 1–6.

III Wu X, Boutellaa E, Bordallo López M, Feng X & Hadid A (2016) On the usefulness of color for kinship verification from face images. IEEE International Workshop on Information Forensics and Security (WIFS 2016), Abu Dhabi, UAE, pp. 1–6.

IV Wu X, Feng X, Boutellaa E & Hadid A (2018) Kinship verification using color features and extreme learning machine. Proc. 2018 IEEE International Conference on Signal and Image Processing (ICSIP 2018), Shenzhen, China, pp. 187–191.

V Wu X, Granger E, Kinnunen T H, Feng X & Hadid A (2019) Audio-visual kinship verification in the wild. Proc. IEEE International Conference on Biometrics (ICB 2019), Crete, Greece, pp. 1-8.

VI Wu X, Feng X, Zhang X, Bordallo López M & Liu L (2022) Audio-visual kinship verification: a new dataset and a unified adaptive adversarial multimodal learning approach. Submitted to IEEE Transactions on Cybernetics, (Major revision). `https://doi.org/10.36227/techrxiv.19776007.v1`.

The author of the dissertation is the first author in all the above articles (*i.e.,* I–VI). The main responsibility of defining the research questions, developing the ideas and methodologies, implementation, and experiments, along with writing, was carried out by the author of the dissertation, while valuable comments and suggestions were given by the coauthors.

Contents

1 Introduction

1.1 Background and motivation

Human faces, as a biometric trait, convey abundant information such as identity, gender, age, and ethnicity. Due to genetic heredity and lifestyle, children are usually more likely to "look" like their parents than other people. In our daily life, we always hear statements such as *"John has his father's nose"* or *"Joe has his mother's eyes"*. Extensive research has shown that human has an instinctive perception ability to indicate the familial genetic relatedness between individuals [1, 2]. This phenomenon has been the subject of a number of psychological studies [3, 4, 5, 6], aiming at understanding how humans visually perceive and identify kin signals from faces. DeBruine *et al*. [3] investigated which parts bear the most kin signals by showing partially occluded facial images to participants. Martello *et al*. [5] in 2006 and Alvergne *et al*. [6] in 2014 also found that the upper half of the face contains more kinship information than the lower face, which was believed to be due to morphological variations in the mouth area, which can produce noise effects. Other researchers [4, 7] studied the individual's biological attributes and pointed out that gender and age differences can significantly reduce the accuracy of kinship verification due to the variations introduced by them.

These psychological studies provided insights into how humans perceive kinship relations based on faces. Inspired by those findings, already in 2010 [8], the computer vision and pattern recognition research community proposed investigating the ability of machines to recognize kinship from facial images. Therefore, Facial Kinship Verification (FKV) emerged. FKV refers to automatically determining whether or not two individuals have a kin relationship from their given facial images or videos.

As a soft biometric trait, kinship information and Facial Kinship Verification can be used in various potential applications. In the anthropology and genetics domain, FKV can help in the study of the hereditary characteristics of close relatives in social relationships [9]. In the field of public social security, it can be applied to finding missing children, border control and customs, and criminal investigations [10, 11]. In the social media domain, FKV can be used for organizing family photo albums, improving the performance of face recognition systems and social media analysis [12]. Furthermore, FKV also has potential applications in smart homes, the Internet of Things (IoT) [13], and personalization.

As an emerging, important, and challenging problem in computer vision, FKV has attracted increasing attention [14, 15, 16]. Many facial kinship verification methods

have been proposed [17] in pursuit of improving verification performance. Already now, several studies [11, 18, 19, 20, 21, 22, 23] show that machine learning methods can outperform by a wide margin the human ability to recognize kinship. This could be expected, since human eyes have low sensory perception to quantify the similarity between two images of two different people [18]. Features such as distance, shape, and color are not easily judged at a glance, resulting in low recognition accuracy, especially for unknown faces never been seen before.

In this context, a remaining question is if, in addition to facial features, other biometric modalities (such as voice) can be used as additional cues for kinship verification. In practice, it is intuitively known that we can recognize certain traits of a parent's voice in their child's speech. Research on the genetics of the voice dates back to the early 1990s [24], which found that both genetic and environmental factors affect vocal characteristics. As a result, similarities in voice quality can be expected within a family. Human studies [25] indicated that listeners had the ability to identify kin voices among non-related ones. Specifically, the similarity of voices within families was [26, 27, 28, 29] quantitatively proven to be perceived by speech parameters such as the fundamental frequency (F0). Given the above evidence, this thesis explores for the first time the usage of acoustic features into solving the kinship verification problem.

Furthermore, automatic kinship verification could also potentially benefit from a combination of discriminant information extracted from both face and voice signals. The fusion of audio-visual features has been shown to be an effective way to improve performance in various problems, including emotion recognition [30], speech recognition [31], event detection [32], and biometrics [33] such as speaker identification and speaker authentication.

Multi-modal fusion methods exploit complementary sources of information. Different sources of information are typically integrated through early fusion (feature level) or late fusion (score or decision level) [34]. Feature-level fusion using concatenation or aggregation is often considered to provide a high level of accuracy. Techniques for score-level fusion using deterministic (*e.g.*, average fusion) or learned functions are commonly employed but are sensitive to the impact of score normalization methods on the overall decision boundaries. However, the success of fusing different modalities depends on representing modal features in an effective manner. In this context, the input raw data should first be fed into a feature extraction module, where the features that capture the subject's identity information are extracted. Deep neural networks (DNNs) [35] provide an effective way to embed these data and a learnable way to automatically determine the correlation between two modalities. In this thesis, we set up

a benchmark study for audio-visual kinship verification and then explore novel deep frameworks for further improving the performance.

1.2 Objectives and contributions

This thesis aims at driving the development of FKV from faces, voices and multiple modalities. Given the theoretical and empirical evidence discussed above, we formulate the hypothesis that FKV can benefit from the fusion of different genetic features. In particular, we have set the following objectives: *(1) Developing a facial feature analysis method that can improve the discrimination of kinship features*, and *(2) Learning specific audio-visual features for kinship verification to further improve the system performance.*

In order to achieve the stated objectives, we set four research questions:

1. As there is no audio-visual kinship dataset, what are its the characteristics, so that the collected data is useful in the study of multimodal kinship verification?
2. How to extract the effective visual features for the FKV problem in order to encode kinship-related information.
3. Is it possible to use acoustic features to verify kinship, and if so, how should they be extracted effectively to contain discriminative information.
4. Given the facial and vocal features, how to effectively fuse the multimodal features to improve the performance of kinship verification?

As discussed in Chapter 1.1, the human voice could possibly show hereditary traits and hence be another useful biometric for verifying kinship. Furthermore, multiple modalities potentially provide the combination of multi-source discriminant information, thus improving system accuracy. Visual FKV systems encounter problems and fail in their predictions in certain circumstances, such as poor illumination. To mitigate this issue, multimodal learning provides an alternative solution by introducing other modality information and enhancing system robustness.

However, there is no available audio-visual kinship dataset. In order to perform this type of analysis, we first aimed at collecting the first set of representative data. To avoid the data bias issue, which might bring unwanted environmental clues, possible familial biases such as recording devices, recording conditions, and speech content are considered during the data collection procedure. We classify this work into two stages. In the first stage, we propose the TALKIN database as the first, yet simple, audio-visual kinship dataset. It consists of limited pairs of talking videos of parent-child relations only. Based on this first attempt at data collection and the subsequent analysis, we used the lessons learned to further establish a larger audio-visual kinship dataset named

TALKIN-Family, which consists of facial videos and synchronous speaking audio with properties that differ from the existing one. In TALKIN-Family, there are 246 unique family trees and 1012 individuals with rich annotations of family relationships, age, gender, and scene conditions. The size of each family tree ranges from 2 to 14 subjects between 5 and 81 years old. Each subject has multiple talking facial videos with a length of about 10 seconds, collected under different environmental conditions. Overall, the TALKIN-Family dataset contains more than 9.2 hours of video.

The general FKV framework includes two main steps: feature extraction and distance measurement. To evaluate the similarity of two facial images, the facial image first needs to be represented with features. In the early stage of FKV research (*e.g.,* before 2016), some common handcrafted descriptors were applied. They showed good verification performance with computational efficiency, especially in small datasets. Compared with naive enumeration features [8] and saliency features [36, 37] (which are generally affected by detection accuracy and facial variance), feature descriptor methods [38, 39, 40, 41, 42, 43, 44, 45] show robustness against noise and rotation.

Despite the relative success of these methods, most of the literature for automatic kinship verification has mainly focused on analyzing only gray-scale face images, hence discarding color information, which can be a useful additional cue for verifying kin relationships. From a biological point of view, the chromaticity of the face is tied to genetically expressed features, such as eye color or skin tone. These hereditary features are often present in kin-related persons in a similar way.

To address the issue, we propose to explore the usefulness of color for kinship verification. We consider different baseline methods used in their traditional gray-scale variants against their counterparts utilizing color information. Instead of analyzing images from the gray channel only, we introduce different image color spaces into feature extraction by combining the features extracted from different color channels. Three color spaces are considered: RGB, HSV, and YCbCr. More specifically, the joint color-texture features encode both the luminance and chrominance information in the color images, which enables color feature representations of kinship.

For kinship verification, we would need to find a proper distance measurement method to compute the distance between an image pair based on feature extraction methods. Recent deep convolutional neural networks (CNN) are a good alternative, but they usually benefit from large datasets that contain samples with a wide range of variations. However, most of the available kinship databases are only composed of limited data[1]. CNNs could then possibly overfit the limited training data and show problems in generalizing unseen test data, especially if collected in very different

[1] Until the publishing data of paper IV.

conditions. To mitigate this problem, we apply a new approach to kinship verification based on the extracted color features with extreme learning machines (ELM) that aims to deal with small sized training sets.

Lastly, based on the presented audiovisual kinship datasets and visual features, we propose and explore the problem of audio-visual kinship verification. We first evaluate the performance of the single modality and set the benchmarks for unimodal methods. Representing modalities, *i.e., audio* and *video*, in an appropriate way is crucial before fusion. Visual features have been widely studied for FKV [20]. Comparatively, very few acoustic features are specifically designed for kinship verification, mostly since these types of studies have been largely under-explored. The well-known acoustic representations such as Mel-frequency cepstral coefficients (MFCCs [46]) and data-driven features [47, 48] have been commonly applied in the speech research community. Similar to the correlation between facial similarity and FKV, we propose to compute the voice similarity and set new benchmark methods for FKV by using acoustic features.

When fusing audio-visual features for FKV, both early and late fusion methods are well-established baseline methods. Based on our benchmarks and investigation, we find that intermodal discrepancy and modal weighting are essential to exploit informative knowledge. Motivated by adversarial learning [49] strategies, and self-attention mechanisms [50], we propose a fusion method named Unified Adaptive Adversarial Multimodal Learning (UAAML), which is based on deep neural networks (DNN), and addresses the aforementioned challenges. UAAML jointly considers multimodal feature learning and kinship attention weights with similarity learning. Particularly, we introduce the L_2 norm layer [51] to generate the unified features before fusion and make the network training stable and efficient.

1.3 Summary of the original publications

Six papers are included in the thesis on the topic of visual kinship verification, of which papers I-II provide a comprehensive review of FKV works that trace back to 2016 and 2022, respectively. Papers III-IV focus on studying the visual-based FKV problem. Papers V-VI are related to audio-visual kinship verification. Six publications (forming the core of this doctoral thesis) are associated with the research questions described in Chapter 1.2, as shown in Table 1. The supplementary publications [52, 53] are related to the content of the dissertation but are not included in it.

Papers I and II review the FKV work in the literature. Paper I provides an in-depth introduction to FKV by identifying the problem definition and challenges. It also covers a larger scale of FKV works, with an improvement in Paper II that focuses on

Table 1. Research questions linked to publications. 'V' represents the marked paper.

Research Questions	Chapter	Paper I	Paper II	Paper III	Paper IV	Paper V	Paper VI
Literature Review	Chapter 2	V	V				
RQ1	Chapter 3					V	V
RQ2	Chapter 4			V	V		
RQ3	Chapter 5					V	V
RQ4	Chapter 5					V	V

summarizing works before the deep learning era. We build an intuitive taxonomy and situate past FKV research works in relation to each other. New ideas and insightful thoughts derived from the current review are provided for developing the next generation of kinship verification techniques.

Papers III and IV investigate how the visual features can be used for describing the kin resemblance in color textures. Paper III investigates for the first time the usefulness of color information for automatic kinship verification from face images. Paper IV proposes tackling the kinship verification challenge by combining color texture feature extraction and Extreme Learning Machines (ELM) for classification.

Papers V-VI study the new problem of *audio-visual kinship verification*. Two new and comprehensive audio-visual kinship databases (TALKIN & TALKIN-Family) were proposed. Paper V presents for the first time the audio-visual kinship verification by studying the problem with the TALKIN database, which results in identifying the fact that human voices can be adopted for kinship verification and are useful for further improving FKV performance. Based on Paper V, Paper VI extended the work from both the database and data fusion perspectives. Extensive benchmark evaluations are performed on the TALKIN-Family dataset with 11 kin relations. The proposed UAAML method achieves an overall competitive performance compared with other baseline methods. In addition to automatic methods, human performance was evaluated using a subset of the TALKIN-Family dataset. Generally, an important finding is that humans tend to have a better ability to verify kinship from the voice than from the face, while when given synchronous facial videos and voice, humans can make a much better judgment. Compared with human performance, machine learning methods can outperform human ability both efficiently and effectively.

1.4 Organization of the thesis

The thesis is organized into two parts. The first part is an introduction with six chapters. It gives an overview of different facial kinship verification problems and main ideas and findings of the original publications. The second part comprises the six original publications related to the technical content of the thesis. The rest of the chapters in the first part are the followings.

Chapter 1 gives an introduction to the thesis by presenting the background and rationale behind the study, pointing out the research problems studied in this thesis, identifying the contributions and a brief summary of the original papers. Chapter 2 gives a general overview on facial kinship verification, including the problem definition, a literature review of state-of-the-art methods, and the publicly available kinship databases that were used and compared in the experiments of this thesis. Chapter 3 presents the TALKIN and TALKIN-Family databases that aim at studying the problem of audio-visual kinship verification under unconstrained conditions. Chapter 4 investigates the visual features for the FKV problem. In Chapter 5, audio-visual feature fusion methods are presented. Specifically, it investigates the single model features, especially the study of the vocal kinship verification problem for the first time, and proposes improved fusion methods. Chapter 6 concludes the thesis by discussing the results, limitations, and future directions of the FKV problem.

2 Overview of facial kinship verification

In this chapter, we summarize the works on FKV in the literature. We first present the problem definition of kinship verification and provide an in-depth understanding of FKV by comparing it with the face verification problem. After that, we introduce the existing methods on FKV from images and videos. Then we present the publicly available kinship datasets that are used or compared in this thesis and the evaluation measures for assessing FKV performance. Finally, we summarize this chapter and discuss open issues.

2.1 Problem definition

Given a pair of facial images, the objective of kinship verification is to judge whether two people are biologically related (with a typical kin relation). Specifically, current kinship verification research uses a clear distinction between multiple kin relation types to study the verification problem. Only close family relationships are involved. These kin relations can be categorized into three levels of generation, *e.g.,* Siblings, Parent-Child, and Grandparent-Grandchild[2]. The four parent-child relations attract the most attention [12], mainly because of their application value. Kinship verification can be formulated as a binary classification problem (Kin vs. Non-kin). FKV primarily consists of two critical sub-problems: feature extraction and classifier designation. Formally, as shown in Figure 1, given a pair of faces $(\mathbf{X}, \mathbf{Y})$[3], appropriate feature representations $(\phi(\mathbf{X}), \phi(\mathbf{Y}))$ are extracted from both images, and then a classifier is used to determine or not if the two faces have a kin relationship.

In order to better understand the FKV problem, we would like to point out the relationship between two similar problems: the FKV problem and the face verification problem (face pair matching) [54] depicted in Figure 1. As can be seen from Figure 1, both problems share a similar algorithm pipeline. The classification at the end is used to judge whether or not if two faces are the same individual in the case of face verification, or whether or not they have a kin relation in the case of kinship verification. Intuitively, both problems depend on the existence of similar facial cues for making judgments

[2]Siblings are family members of the same generation: Brother-Brother (BB), Sister-Sister (SS), and Sister-Brother (SB). Parent-child relations are the first generation: Father-Son (FS), Father-Daughter (FD), Mother-Son (MS), and Mother-Daughter (MD). The Grandparent-Grandchild relation belongs to the second generation: Grandfather-Grandson (GFGS), Grandfather-Granddaughter (GFGD), Grandmother-Grandson (GMGS), and Grandmother-Granddaughter (GMGD).

[3]A face detection and normalization procedures are typically used to obtain the face in the images.

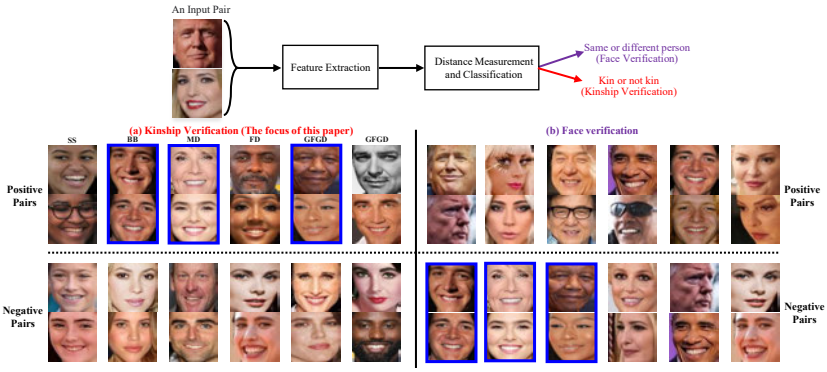


Fig. 1. General pipeline for face verification task and kinship verification task. Both tasks calculate the similarity of two facial images. While, positive pairs in the kinship verification task are negative pairs in the case of face verification. Reprinted, with permission, from Paper I ©2022 Springer.

[7, 55], especially in the case of face verification where each positive pair represents the same individual (see Figure 1 (b)). In the case of kinship verification, each positive pair represents two different individuals with a kin relation (detecting kin clues in specific areas of the face rather than from the entire face [5]). Note that all positive pairs (including identical twins) in the case of kinship verification (see Figure 1 (a)) are negative pairs for face verification. It is interesting to ask: Do pairs of the same individuals (positive pairs in the case of face verification) belong to positives or negatives in the case of kinship verification[4]? This question has been overlooked as current FKV research assumes each input face pair belongs to two different individuals in their experimental setting. From an anthropological point of view, FKV is based on the degree of genetic similarity between the faces of two subjects. Thus, it is reasonable to expect that an FKV system will give high prediction accuracy for facial pairs of the same person. When facial images from one individual present an age variation, the study of age-invariant face verification can somehow be viewed as self-kinship verification, where the system verifies the same individual as related to himself [12, 56]. On the other hand, as facial aging and kinship are both genetically inherited [57], kinship is capable of providing guidance for age progression and boosting face verification. Conversely, a de-aging process can be performed to learn discriminative identity features for both face verification [58] and kinship verification [59].

[4]If one expects face verification to achieve age invariance, these pairs of images representing the same individual can also have an age gap and more significant differences than those of the same age.

Kinship is a well-established biological concept, but determining what kind of similar facial cues are critical for FKV is still an open question. According to recent psychology studies [7, 4], facial similarity and kinship judgments are highly correlated but not strictly synonymous. This makes FKV a difficult problem with various challenges which we discuss below.

2.2 Main challenges

As we defined above, FKV is formulated as a binary classification problem. The difficulty of FKV stems partially from the fact that the kinship facial pairs do not belong to the same identity and only show hidden genetic facial similarities that are more complex and less discriminative than similarities in other problems like facial verification. As discussed above, it is evident from psychology research [7, 4] that facial similarity and kinship judgments are not strictly synonymous though they are highly correlated, which makes the problem of FKV even harder. The main challenges of FKV are summarized in Figure 2, with visual examples for illustration.

(1) Large intraclass variations. As can be seen from Figure 2 (a), there are two types of intraclass variations: *intrapersonal* variations (facial appearance changes in the same identity) and *interpersonal* variations (facial appearance differences in different identities). The large intrapersonal variations come from uncooperative subjects such as changes in age pose, expression and accessories, unconstrained imaging environments like changes in illumination, imaging distance and angle, variations in image quality and resolution, blur, and even adversarial attacks (Figure 2 (b1)). All these pose great challenges for extracting discriminative features for kinship verification and greatly impact FKV performance. Many early approaches to FKV only considered facial images acquired in cooperative conditions. Therefore, it is more practical to build large-scale kinship datasets in the wild.

As the input of an FKV algorithm is a pair of facial images belonging to two individuals, the goal of FKV is to explore the hidden factors of visual similarity between the two input faces for kinship determination. Therefore, there are significant interpersonal variations that increase the intraclass distance between the positive class samples. Firstly, there can be a significant age gap between the kin pairs, particularly when verifying cross-generation kinship types. Figure 2 (b2) shows parent-child pairs with a similar age and a considerable age gap. It has been demonstrated that parent-child pairs with a similar age have more similarities [60, 7]. However, pairs of older parents and younger children can have significant textural differences between the two faces, which negatively influences similarity. Secondly, gender differences also negatively

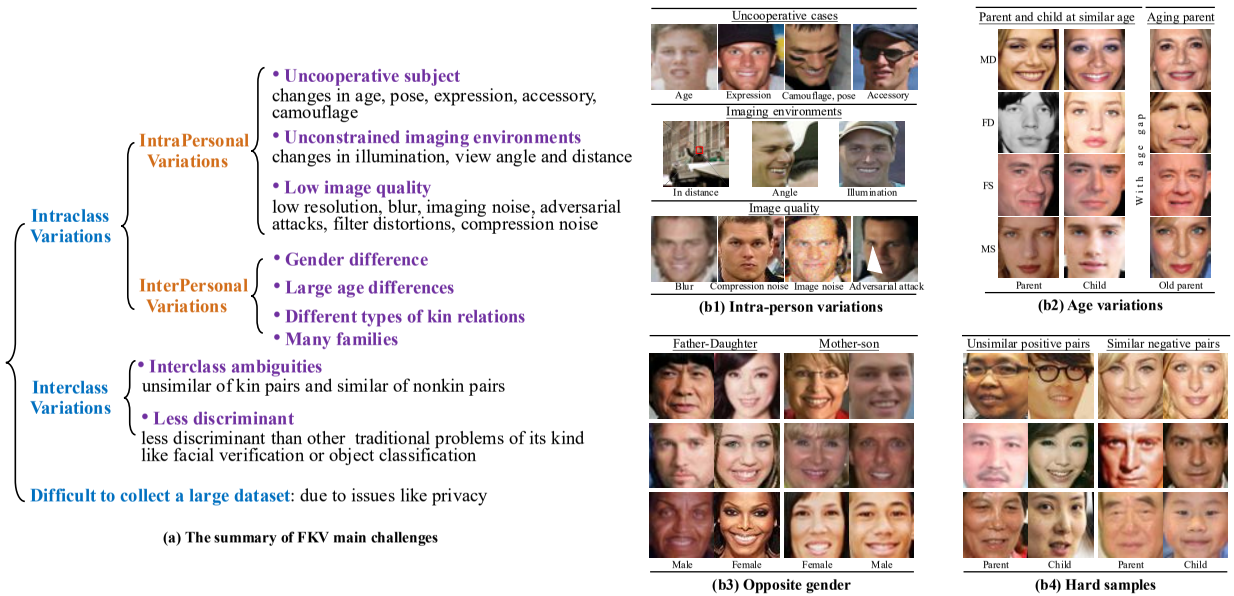


Fig. 2. The main challenges of facial kinship verification. Sub-figure (a) provides a taxonomy of these challenges brought by intraclass variations, interclass variations, and data establishment. The right sub-figure illustrates key scenarios with facial sample images. On the right, (b1), (b2) as well as (b3) show intraclass variations, in which (b1) contains the possible variations within one subject, with each image line demonstrating the influences of different factors. Then, (b2) and (b3) illustrate the facial similarity gap between kinship caused by age and gender differences, as well as variations among kin pairs and families. Figure (b4) demonstrates less discrimination of FKV that hard kin and non-kin samples exist when kin pairs have less similarity in appearance, while non-kin pairs inversely show significant similarity. Reprinted, with permission, from Paper I ©2022 Springer.

influence facial similarity. As shown in Figure 2 (b3), the kin pairs of mother-son, father-daughter, and brother-sister have different gender variations. It has been shown that non-kin pairs with the same gender have more similarities than those with a different gender [7]. Finally, in addition to the existing considered kinship types, facial similarities can also exist between some family members when one increases the height or width of the family tree (*e.g.,* by including cousins and nieces). One reason for this is that the inheritance among different kinship types is not deterministic [61]. It is tough to determine a mathematical inheritance model due to its randomness and required multidisciplinary knowledge [61, 62].

(2) Small interclass variations. As we defined above, FKV aims to learn a binary classifier by distinguishing a number of positive kinship pairs from a number of negative samples. The similarity among kin faces is attributed to hidden factors rather than the whole face. As illustrated in Figure 2 (b4), some positive examples may have small similarities, whereas negative examples may have high similarities. Therefore, small positive and negative variations decrease the interclass separation and pose significant challenges for learning the real decision boundary. In addition, there is a severe imbalance issue [15], *i.e.,* , the number of negatives is significantly more than the number of positive pairs.

(3) Difficulty in gathering large-scale kinship datasets. The lack of large kinship datasets impedes the development of FKV algorithms, especially the development of deep learning-based methods that are data-hungry. It is essential to collect a large kinship dataset that can represent the actual data distributions of families worldwide, reflecting the intraclass and interclass variations discussed above. However, due to security and privacy issues, it is challenging to meet this requirement.

2.3 The extended studies

FKV study is the widely explored and fundamental research problem of kinship recognition. Due to varying applications of kin-tasks, complementary kin research problems have emerged (illustrated in Figure 3).

Tri-subject kinship verification

A child's genetic inheritance comes from both parents (father and mother). This leads to tri-subject kinship verification [19], where the inputs are both parents' facial images and the child's facial image. Suppose that $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the father and mother's facial images (or videos), respectively, and $\mathbf{Y}$ formulates a child's facial image (or video). The

feature representations of parents and child are extracted, $\phi(\mathbf{X}_1)$, $\phi(\mathbf{X}_2)$ and $\phi(\mathbf{Y})$. The distances are computed between a child and his or her parents, $d(\langle \phi(\mathbf{X}_1), \phi(\mathbf{X}_2) \rangle, \phi(\mathbf{Y}))$, to verify whether they have a kin relation. Tri-subject kinship verification is also a binary classification problem.

Family classification

Family classification [63] is a multiclass classification problem, *i.e.,* the classification task contains multiple categories, and each category represents a family. Given a pending facial image, we need to determine which family it belongs to. A collection of k families is represented by $\chi = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k\}$. The corresponding multiclass label can be written as $\{y_1, y_2, \ldots, y_k\}$. By training a classifier, the system outputs the family label of an input facial image $\mathbf{x}$. The difficulty of family classification increases when family classes increase. In the FIW dataset [20], family classification accuracy is only 16.18% from a total of 564 families.

Family search and retrieval

Family search and retrieval [20] is designed to match family members to the input facial image, where the search is performed on a set consisting of members from all families. The input facial image is a query, and the output gives the most matched K family members. The difference between family classification and family retrieval is that family classification focuses on the training of family classification models, and family retrieval tries to retrieve face images that are more similar to the images to be queried through similarity metric learning and find the input's parents and other kinship members.

Other tasks

Other tasks include kin face synthesis [64, 65] and kin relation classification [66, 67]. The kin face synthesis study takes the facial images of parent(s) to synthesize the child's image. By synthesizing the child's facial image, kinship data are augmented for training and improving the model consistency, thus assisting FKV. Besides, it can also be applied for matching missing children. In the kin relation classification, the inputs are two facial images with a particular kin relation, and the system estimates which specific kin relation they have. This task has applications in family album organization and social media analysis.

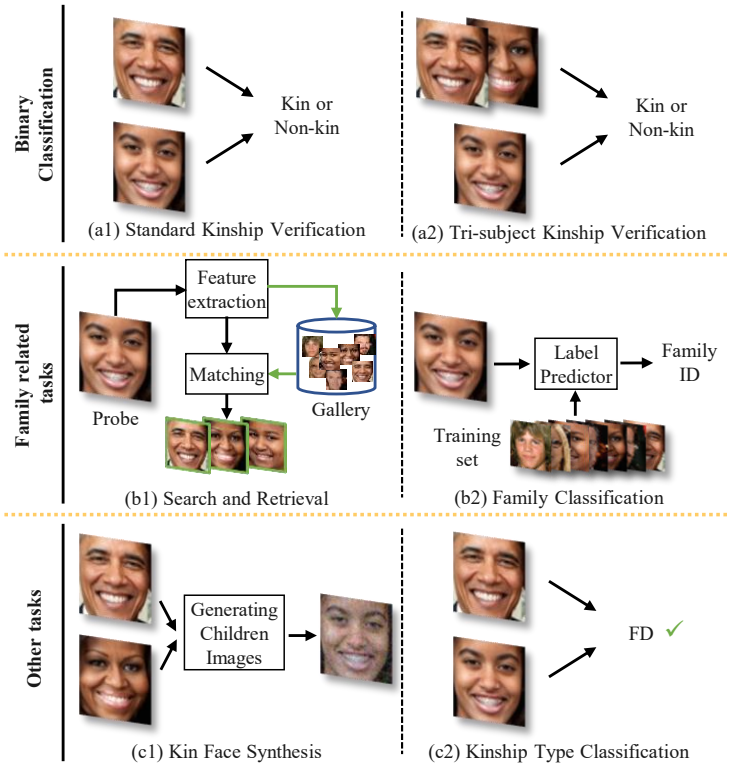


Fig. 3. Kinship analysis tasks. The main task is kinship verification. We categorize kinship-related research directions into binary classification tasks, family-related tasks, and other tasks. Reprinted, with permission, from Paper I ©2022 Springer.

Since the study of kinship analysis is still in its initial stages, facial kinship verification is the key and core of kinship research, which is also the focus of this thesis.

2.4 Facial kinship verification from images

Facial kinship verification from still images is popular, mainly due to the easily obtainable datasets and its wide range of applications. Generally, the kinship datasets contain pre-processed facial images. Facial images are cropped and resized to a normalized size. The main efforts are dedicated to kin feature extraction and distance measurement. Then the classifier is used for binary classification.

2.4.1 The key steps for facial kinship verification

(1) Face detection, alignment and segmentation. The goal of this step is to do face detection based on the input of raw facial images. After locating the face, the eyes' position is usually taken as the key feature to align the face. The purpose of face alignment and face adjustment is to reduce the influence of face scale and angle. The commonly used methods for face segmentation and alignment include MTCNN [68] and ERT [69]. Extensive research reviews on this sub-task have been carried out for example, the survey work of Wu *et al*. [70].

(2) Kin feature extraction. The two input facial images can be represented as **X**, **Y**. We extract features for these two facial images and denote them with the vectors **x** and **y**. Kin features will then be employed for distance measurement and classification in the next step. The kin feature extraction step is an important research topic, and it also affects performance. Before deep learning techniques are used in kinship verification, some common handcrafted descriptors are applied. With the implementation of deep learning in kinship verification problems, the traditional feature descriptors are gradually replaced by deep embeddings.

(3) Distance measurement. By extracting facial image features, two inputs are represented as two vectors. Then a proper distance metric is used to calculate the distance of two inputs in the feature space and assess the similarity between two faces. Metric learning aims to learn a transform matrix to narrow the distance between kin pairs (positive pairs) and enlarge the distance between non-kin pairs (negative pairs). The extracted facial features can be mapped into a new feature space and improve the performance of kinship verification [11, 71, 72].

(4) Classification. The steps above produce a distance value between sample pairs. Kinship verification is a binary classification problem where commonly used classifiers are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and threshold classification.

2.4.2 Traditional methods

As kinship verification is a relatively new and challenging problem, many kinship verification methods were proposed during the last decades. At the beginning of the kinship verification research, traditional methods were proposed for solving the kinship verification problem. They showed good verification performance with computational efficiency, especially in small datasets. This subchapter reviews the traditional methods from aspects of feature extraction and metric measurement.

Feature extraction methods

To establish an automatic facial kinship verification system, the facial images should be first represented with features effectively. We categorized these methods into enumeration features, facial saliency features, hand-crafted features, feature transformation based on color spaces, and feature selection methods. Naive enumeration features started with the work of Fang [8] *et al*., which represented the facial traits from low-level features with different points of view, such as eye color, skin color, hair color, geometric characteristics between facial key points (eye, mouth, nose) and face shapes (size of the eyes, mouth or nose). Later, Xia [73, 74] included more descriptive information, such as age, gender, and race. These features are represented with binary features encoded as -1 and $+1$. Nevertheless, the enumeration of these features needs manual efforts to label the samples, while the resulting features are usually low-dimensional and not comprehensive enough.

Kinship verification based on saliency features. Methods of kinship verification based on saliency aim to verify kinship by comparing the similarity of salient facial parts, such as the nose, eyes, and mouth [36, 37, 75, 76]. Thus, we need to first locate the facial key points. Given a facial image, to find the salient parts, Guo *et al*. [37] proposed to use the eyes, mouth, and nose as the salient facial area. The DAISY descriptor [77] is applied to extract features and compute the similarity between the image pairs. Kohli *et al*. [36] proposed the Differences of Gaussians (DoG) method to locate the facial key parts. Then in 2014, Wang *et al*. [75] introduced the widely used 68 facial landmarks [78] extracted from facial images into kinship verification. Besides the methods that extract facial key points and facial landmarks, Goyal*et al*. [76] proposed an edge detection-based kinship feature extraction method. The Canny operator [79] was used for detecting the facial edges, and areas enclosed by them were considered salient parts.

Hand-crafted features. The previous subchapters introduced methods based on the shape of the face. These methods are usually affected by detection accuracy, facial expression variance, noise, and face rotation, resulting in low verification accuracy and low noise tolerance under complex conditions. To solve these problems, researchers proposed feature descriptor methods [38, 39, 40, 41, 42, 43, 44, 45] for kinship verification. Among them, Local Binary Pattern (LBP) [80] is a widely used hand-crafted feature extraction method. LBP is an operator that describes the image's local texture information. The resulting binary code describes the texture characteristics of an image block and is invariant to both rotation and gray-scale conversion[81].

Based on the basic hand-crafted features, many methods improve the performance in different ways. The Pyramid Multi-level covariance descriptor (PML-COV) [43] combined the LBP and HOG features extracted from multiple resolutions to establish the feature pyramid. Goyal *et al*. [45] proposed the Selective Patch-based Dual-Tree Complex Wavelet Transform (SP-DTCWT) method, which decomposes the facial image using six wavelet functions. By computing the similarity between corresponding patches of an image pair, they can get discriminative feature patches for kinship verification.

Feature selection. Unlike single feature extraction methods, feature selection aims to study fusion schemes by selecting among multiple features, enriching feature representations, and reducing feature redundancy [82, 83, 84, 85]. Usually, the inputs of feature selection methods are multiple feature representations. They can select the most effective representations by introducing a constraint as an objective function or directly as the classification accuracy. Alirezazadeh *et al*. [82] first proposed to fuse local and global features and select the valuable and discriminative features for kinship verification. Bottinok et al [83] extracted multiple features from images, including Local Phase Quantization (LPQ), Weber's Local Descriptor (WLD), and LBP. Before they classify the features, to improve the verification accuracy, they propose the Max-Relevance and Min-Redundancy (mRMR) method to select a subset of variables to best describe the data.

Metric learning methods

Metric learning was firstly proposed by Eric Xing *et al*. [86] at NIPS 2002. For the kinship verification problem, we would need to find a proper distance measurement method to compute the distance between an image pair based on feature extraction methods. Ideally, in this metric, the image pairs with kin relations (positive pairs) would have small distances, while those without kin relations (negative pairs) would have large distances. It maps the distance metric space into a new metric space [87]. The commonly used basic distance metrics in kinship verification are Euclidean distance [71], Mahalanobis distance [11, 72, 88, 89, 90, 91, 92], bilinear similarity [93, 94, 95, 96, 97, 98], graph learning [99, 100], cosine similarity [101, 102], CCA [103] and other metric patterns [104, 105, 106, 107, 108].

Neighborhood Repulsed Metric Learning. In 2014, Lu *et al*. [11] proposed the Neighborhood Repulsed Metric Learning (NRML) method for kinship verification (which is also the first try at metric learning in solving kinship verification), and provided the fundamental theory and protocol for the metric learning based kinship verification study. The motivation of NRML is that the negative neighbors of positive samples can

confuse the classifier. Based on that, NRML repulses the k negative neighbors and pulls the positive samples together, thus separating the positive samples and negative samples. The NRML method showed the best performance at that stage in 2014, achieving 73.8% and 69.9% verification accuracy on KinFaceW-I and KinFaceW-II datasets. The main idea of NRML is also used in other metric learning methods. Yan *et al*. [71] proposed to map the feature vectors into the hyperplane of SVM and applied the NRML method to optimize the distance metric. Xu *et al*. [94] concatenated multiple features into one vector and combined the NRML method with bilinear similarity to compute the distance between image pairs. Later, Yan et al [101] and Lei *et al*. [103] replaced the distance metric with cosine similarity and CCA. They also demonstrated the effectiveness of NRML.

Metric learning based on bilinear similarity. Besides the commonly used Mahalanobis distance measurement, bilinear similarity $S_{\mathbf{W}}(\mathbf{x}_i, \mathbf{y}_i) = \mathbf{x}_i^T \mathbf{W} \mathbf{y}_i$ is also used for the metric learning-based kinship verification studies, where $\mathbf{W}$ is the positive semidefinite matrix. When $\mathbf{W}$ is the identity matrix, bilinear similarity can be viewed as the cosine similarity without normalization. Bilinear similarity has shown good performance for image retrieval [109, 110] and it can effectively calculate the similarity between two sparse feature vectors. Zhou *et al*. [93, 95] proposed the Ensemble Similarity Learning (ESL) method to solve the kinship verification problem. The ESL method has superior computational efficiency and can be applied to high-dimensional data. Then the inputs of ESL are quadratic, which satisfies the inter- and intra- constraints on the similarity pattern for image pairs. Qin *et al*. [97] proposed a multitask-based bilinear similarity learning method. They combined the four kinship verification tasks to transfer knowledge from one task to other tasks. Fang [98] introduced the logistic loss to smooth the objective function and improve the efficiency of the optimization process.

Other metric learning methods. Besides the metric learning reviewed above, researchers also proposed methods from other points of view. Zhang *et al*. [104] proposed a generic metric. In the feature space, the distance between a child and two parents can be computed by the minimum length from the child feature vector to the feature vectors of two parents. Liu *et al*. [105, 106] introduced the angle θ between the parent's and child's feature vector to formulate the objective function. Wu *et al*. [107] introduced a low-rank metric learning method to learn the latent subspace and dig more discriminative representations adaptively. Zhao *et al*. [108] proposed the multi-kernel metric learning method, including linear and nonlinear distance metric methods. By weighted fusing them, they can obtain the final distance. The graph learning method is also studied for metric learning-based kinship verification. Liang *et al*. [100] built the Intrinsic Graph and Penalty Graph according to the relationship between the data.

They combined the NRML algorithm and graph learning to describe the intraclass compactness and interclass separability.

Age variance between parents and children can have an adverse effect on kinship verification. Shao *et al*. [111, 60] pointed out that children and their parents look more alike when the parents are young. The idea of reducing the divergence caused by the aging effect is to utilize the young parent's facial images as a bridge between children and elder parents. The module takes images of young parents, old parents, and children as the source, intermediate, and target, which can be denoted as $\mathbf{X_{yp}}$, $\mathbf{X_{op}}$ and $\mathbf{Y}$. The subspace projector matrix $\mathbf{W}$ is learned to project the intermediate domain and the other two domains to have the same distribution. One drawback of this study is that it requires manual efforts to collect the images of parents both when they are old and young.

Metric learning methods project the feature vectors into a new feature space that pulls the kin image pairs together and pushes the non-kin image pairs further away. In this sub-chapter, we reviewed and summarized the existing metric learning-based kinship verification methods. Traditional metric learning methods are based on the feature extraction module. Besides that, deep metric learning methods integrate the feature extraction and metric learning loss to guide the deep network in learning comprehensive feature extraction strategies. We will review these methods in the following sub-chapter.

2.4.3 Deep learning methods

Traditional hand-crafted feature extraction methods have a limited ability in feature description. While the CNN-based deep learning methods have a strong capability of non-linear expression, they can learn the effective feature embeddings from the original raw data by applying task-related constraints, thus avoiding the traditional hand-crafted feature extraction rules [112, 113, 114].

With the fast development of deep learning in computer vision and the emergence of large-scale kinship datasets, researchers started to study the deep learning methods for kinship analysis in 2016 [53]. The existing facial kinship verification algorithms have used multiple novel deep architectures, including basic neural networks [115], deep metric learning [53], architectures based on auto-encoders [115, 65, 10] and attention networks [116].

The very first method proposed by Wang *et al*. [115] in 2015 has two stages: feature extraction and deep metric learning. The facial features are extracted by traditional methods. Features are fed into nonlinear AutoEncoders followed by the Mahalanobis distance metric to project the features into a non-linear space. The drawback of the method is that the input is the LBP feature, and the detailed information of the original

image is missing. The first End-to-End deep learning method for kinship verification was proposed by Zhang *et al*. [117]. The network inputs two stacked facial images, and then outputs the final result. The architecture of the network is simple yet effective.

Deep metric learning methods. To optimize the distance between two input facial images, researchers proposed to add a distance metric to the network training, which we call Deep Metric Learning methods [53, 118, 119, 120, 121, 122]. The typical network architecture is the Siamese network. Different from one-stream networks, Siamese networks have two streams that share the same weights and use the distance metric as the loss function to learn an optimal feature space so that positive pairs (pairs with a kin relation) have small distances and negative pairs (pairs without a kin relation) have a large distance.

Li *et al*. [53] proposed the similarity metric based convolutional neural networks (SMCNN) method. The inputs of the network are two facial images, $\mathbf{X}$ and $\mathbf{Y}$. $G(\cdot)$ indicates the FC layer output of the network. They employed the l_1-norm to compute the distance of two output embeddings. During the training, Li *et al*. added a threshold τ to further partition the positive and negative samples. The labels for the positive samples and negative samples are denoted as $y = 1$ and $y = -1$. We can then have the cost function of the network: $L_{SMCNN} = f(1 - y(\tau - D(\mathbf{X}, \mathbf{Y})))$, where $f(\cdot)$ is the generalized logistic loss. To minimize the cost function, the gradient descent algorithm is adopted to optimize the convolutional neural networks.

Moreover, the commonly used metric-based loss functions include contrastive loss and triplet loss [123, 124]. These two loss functions are based on distance measurement, such as the Euclidean distance. The contrastive loss takes positive pairs and negative pairs as inputs. Different from contrastive loss, triplet loss has three inputs, including the Anchor (a), Positive (p), and Negative (n). The positive and negative pairs refer to the anchor sample. Thus, positive sample pairs are clustered, and the positive and negative samples are separated.

Regarding the deep metric learning techniques, the selection of the sample pairs/tuples can directly affect the efficiency and performance of the network. Researchers proposed the hard sample mining methods [125, 15, 126]. Hard sample mining methods are designed to find positive sample pairs with large distances and negative sample pairs with small distances from training batches, which can produce large backward losses and effectively train the network. Li *et al*. [15] proposed a discriminative sample mining approach using meta-learning in kinship verification. They abandoned the easy negative ones and kept the hard samples to dominate the gradient.

Architectures based on auto-encoders. Another deep kinship verification architecture is based on Auto-Encoders (AE) [115, 65, 10, 127, 128, 64, 129, 130, 123]. The

very first autoencoders to be applied to kinship verification aim to train a model for facial feature extraction [115]. The encoded feature is the reduced feature representation of the input. Many auto-encoder methods were motivated by the correlation between inputs and outputs. They can be categorized into two classes, traditional autoencoders [10, 130, 127] and NN-based autoencoders [128, 123, 129]. Traditional autoencoders learn the relation mapping representation by minimizing the loss function formulated to fit two input images. The NN-based autoencoders use multiple layers of projection and optimize the network by back-propagation. Liang *et al*. [128] proposed to use the intermediate layer to describe the relationship between inputs and outputs. They first extracted the features of two facial images by a pre-trained CNN. The obtained features are the inputs of the autoencoders. By minimizing the difference between the encoded feature and the child's feature, the autoencoders can be optimized, and the output of the intermediate layer shows the relational feature of the two facial features. The method proposed by Liang *et al*. requires learning the rational feature every time a new input pair arrives.

Dibeklioglu *et al*. [123] improved it by encoding both inputs into a dual network and defined comprehensive losses to learn kin-related features in an End-to-End fashion. They took a pair of kin images as the inputs of dual autoencoders. They made the output of each decoder to be similar not only to the input facial image but also to its kin facial image. At last, they adopted the encoded features as the kin feature representations. Moreover, some researchers applied image synthesis and generative techniques to synthesize a child's facial image given the parent's facial image by Generative Adversarial Networks (GANs) [65, 131]. Besides, GANs can also be use to learn disentangled images or representations when facing challenges such as age and gender. Wang *et al*. [59] applied GANs as a cross-generation framework for generating young parents. The old parents were transformed to their young ages to mitigate the age gap.

Architectures based on the attention scheme. Psychological research indicates that kin clues are located in specific areas of the face rather than the entire face [5]. The method discussed above take the whole face as a clue for verifying kinship while ignoring the facial kin feature distribution. In order to learn an effective kin feature embedder, multiple attention mechanisms can be applied to guide the network to pay attention to genetic regions. One possible and widely used method is the channel-wise attention mechanism [132]. It learns an adapting weight for different feature channels, as it is assumed that channel-wise features reflect variant information over space. By training deep networks with kin-constrained loss function, the kinship-interested feature is generated.

Yan *et al*. learned the facial geometric weights directly from the transformation of the intermediate feature map. They also applied the residual learning idea to retain original information by summing the weighted feature map with the original feature map. Specifically, the feature map passes through a pooling operation and convolutional layer. To restore the feature map to the same size as the original feature map, they used an up-sampling method followed by a sigmoid function to map the weights to a 0 to 1 scale. The original feature map is formulated as $C(\mathbf{X})$ and $F(\mathbf{X})$ denotes the attention weights. The weighted feature is denoted as $P(\mathbf{X}) = F(\mathbf{X}) * C(\mathbf{X})$. To avoid the loss of information, Yan *et al*.applied the residual method $P(\mathbf{X}) = (1 + F(\mathbf{X})) * C(\mathbf{X})$. The attention network shows good performance on KinFaceW-I and KinFaceW-II datasets. They reached 82.6% and 92.0% accuracy, respectively, which is superior to basic CNN.

2.5 Facial kinship verification from videos

Compared to still images, facial videos can provide more information. A video-based kinship verification system indicates the kin or non-kin relation between subjects present in video sequences containing faces. This is an important research problem for some use cases, such as surveillance systems and social media broadcasting. The first video-based kinship verification study dates back to 2013 [22], when Dibeklioglu [22] combined appearance and dynamic features to depict kin characteristics. Although video-based kinship verification is an extension of image-based kinship verification research, it contains additional spatio-temporal information that can be useful for FKV. However, due to the significant challenges listed below, video-based kinship verification has still not reached its full potential.

(1) Low quality of facial videos. Typical facial videos are usually recorded with subjects who do not always cooperate with the recorder. Hence, the facial quality shows more variability, especially in pose and illumination, which can fluctuate across subjects and frames of the same video. In addition, occlusion and target loss are also possible. Eliminating the noise while adaptively extracting helpful information is still an unsolved problem, which is usually mitigated in current datasets by simplifying the recording conditions [22].

(2) Blurry video frames. The understanding of moving faces in sequences is frequently hindered by frame blurring due to motion. This is especially evident with slow shutter speeds and long exposure times [133]. Advanced devices can address this issue by collecting data at higher frame rates, with high-quality optics and short exposure times. However, this can cause an unnecessary waste of resources [134]. Deblurring video frames for kinship analysis still remains a challenge.

(3) Integration of faces, audio, and body information. Videos provide rich behavior information and dynamic cues besides facial appearance. Voice [124] and gait [135] could act as complementary modalities that provide kin clues. The main challenge is to devise how to properly fuse multiple modalities to learn the complementary features for kinship verification.

Video-based kinship verification systems are similar to image-based kinship verification and follow the similar approach introduced in Chapter 2.4.1. The distinct difference is in modeling kin features from sequences. We review the existing video-based methods from constrained video-based kinship verification in Chapter 2.5.1 and unconstrained video-based kinship verification in Chapter 2.5.2.

2.5.1 Constrained video-based kinship verification

Constrained video-based kinship verification refers to verifying kinship from facial videos where there is no variance in the shooting environment and subject actions. A representative constrained dataset is the UVA-NEMO Smile dataset [22]. It is hypothesized that people with kin relations might also share similar facial expression dynamic features that could be present in a smiling style, for example. This hypothesis was corroborated by the original authors in 2013.

Dibeklioglu *et al*. [22] extracted the dynamic and facial spatio-temporal features for kinship verification. They localized 17 facial landmarks to track facial movement and extracted the dynamic features based on them. Together with the spatio-temporal feature CLBP-TOP, they demonstrated the family resemblance of smiling faces. Boutellaa [136] combined deep features and spatio-temporal features (*e.g.,* LBP-TOP) to study constrained video-based kinship verification. Experimental results showed that deep features have complementary information regarding spatio-temporal features. In 2017, Dibeklioglu *et al*. [123] proposed to measure the similarity of kin facial smile videos by matching affective intensity. They decomposed the smile video into frames and aligned the sub-sequence according to the smile intensity of the face. The matched sequence pair is the input of dual auto-encoders.

Constrained video-based kinship verification studies indicate that people with kinship have both a similar appearance and smiling expressions. However, it requires strict collection conditions that hinder its applicability. To answer this limitation, researchers formulated unconstrained video-based kinship verification, which we will review in the following sub-chapter.

2.5.2 Unconstrained video-based kinship verification

Compared with constrained videos, unconstrained videos are collected in the wild. Relaxing the restriction of the collection conditions makes it easier to enlarge the scale of the datasets. The collection of large video frames provides a larger number of individual frames to be used in training, but at the same time, it severely increases the burden of computation. On the other hand, the variability of the collected videos also provides for additional multimodal cues that could be exploited in a complementary manner.

Yan *et al*. [23] investigate the problem of video-based kinship verification with several metric learning methods. Compared to a single image, a face video provides more information to describe the appearance of a human face. It can capture the face of the person of interest from different poses, expressions, and illuminations. Kohli *et al*. [10] proposed a three-stage autoencoder to learn the relation between two facial videos, called Supervised Mixed Norm AutoEncoder (SMNAE). First, every video was decomposed into a sub-sequence with a specific number of frames, called a *vidlet*. The vidlet pair is the input of the three-stage autoencoder. In the first stage, the relation of the corresponding video frame was learned as the facial resemblance. The second stage concatenated the spatio-temporal representations. In the end, the third stage fused the spatio-temporal information and learned the final score of kin probability. This method has a common drawback, since the learning procedure needs to be repeated for each input pair.

Besides video-based kinship verification is still a relatively new research topic, only limited research was found in the literature. Since it shows the potential capability of describing more comprehensive features related to kinship when compared to facial images, kinship verification from videos deserves more study in the future.

2.6 Kinship datasets

Databases play an important role in the study of kinship verification. In the era of big data, the collection of large databases is becoming more and more important. On one hand, an open standard database provides researchers with experimental data and unified evaluation standards. On the other hand, the construction and development of the database further promotes the development of the research problem. Before researchers first raised the kinship verification problem, there was no relevant kin face database. Fang *et al*. [8] from Cornell University established the first kinship database, named Cornell KinFace, in 2010. It consists on 300 facial images (*i.e.,* 150 parent-child pairs) and includes four kin relations: FS, FD, MS, and MD. During the next ten years,

(a) Sample images from the KinFaceW-I dataset.

(b) Sample images from the KinFaceW-II dataset.

Fig. 4. Sample images from the KinFaceW-I & II datasets. From top to bottom are the FS, FD, MS, and MD kinship relations, and the neighboring two images in each row are with the kinship relation, respectively. Reprinted, with permission, from [11] ©2014 IEEE.

many scholars established a variety of kinship databases. Compared with the very first database, new ones have been developed and enriched in terms of size, structure, kin relation types and data modality. Next, we briefly review the kinship datasets with images and videos, which are primarily and commonly used in kinship verification.

Fig. 5. Sample images from the TSKinFace dataset. Each group consists of father-mother-child. From top to bottom are Father-Mother-Daughter (FMD) and Father-Mother-Son (FMS) kinship relations, respectively. Reprinted, with permission, from [19] ©2015 IEEE.

2.6.1 Image datasets

KinFaceW

The KinFaceW[5]. [11] database has two subsets: that are KinFaceW-I and KinFaceW-II. They are all collected from the Internet and there is no constrain on the recording environment. KinFaceW-I and KinFaceW-II proposed by Lu *et al*. [11] are composed of facial images with kin pairs. The resolution of each image is 64×64. These two subsets have the same kin relations. The difference between them is the result of kin images in KinFaceW-I being cropped from different photos and the kin images of KinFaceW-II being from the same photo. In KinFaceW-I, there are 134 pairs of FS, 156 pairs of FD, 127 pairs of MS, and 116 pairs of MD. In KinFaceW-II, every kin relation has 250 pairs of facial images. The KinFaceW database is widely used in research on kinship verification. Figure 4 shows samples of the KinFaceW-I and KinFaceW-II datasets.

TSKinFace

The TSKinFace[6] [19] database is a facial image database proposed by Qin *et al*.. Different from the KinFaceW database, it is mainly used for the study of tri-subject kinship verification. TSKinFace has two types of kinship relations: Father-Mother-Son and Father-Mother-Daughter, for which there are 513 groups and 502 groups, respectively. Figure 5 shows sample images in TSKinFace that are organized with triplets. The facial images are all downloaded from the Internet and there are no constraints on the recording environment. The resolution of each facial image is 64×64.

[5] http://www.kinfacew.com/

[6] http://parnec.nuaa.edu.cn/xtan/data/TSKinFace.html

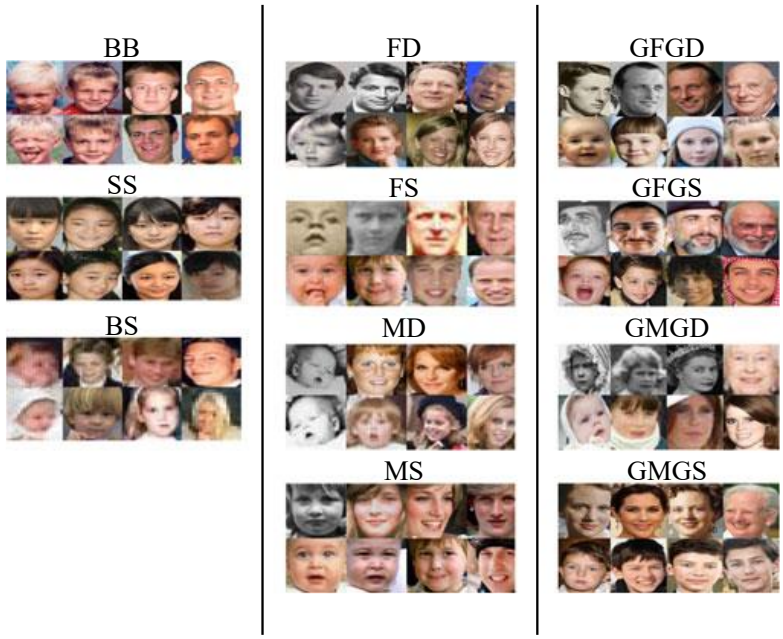


Fig. 6. Sample images from the FIW dataset. Modified, with permission, from [20] ©2018 IEEE.

FIW

FIW[7] (Families In the Wild) [20] is the largest and most comprehensive kinship database for now. It consists of facial images and was established by Robinson *et al.*. The FIW database is organized by family. It has over 13,000 facial images from 1,000 families. Each facial image is resized to 224×224. FIW consists of multiple facial images from different periods for each family member. FIW covers the kin relations of siblings, four parent-child kin relations, and four second-generation kin relations. Figure 6 provides some samples with 11 kin relations from the FIW dataset. FIW is similar to Family 101 [63], but it is much superior in family structure, data volume, and data variants.

2.6.2 Video datasets

UVA-NEMO Smile

The UVA-NEMO Smile[8] [21, 22] database was proposed by Dibeklioglu *et al.*. The database was first established to classify spontaneous smiles and deliberate smiles.

[7]https://web.northeastern.edu/smilelab/fiw/
[8]http://www.uva-nemo.org/

Because the participants in the database are family-related, it is also considered to be the first video-based kinship database. The UVA-NEMO Smile database has 1240 clips of smiling videos of 400 subjects (597 spontaneous smiling videos and 643 deliberate smiling videos). The background of the video is black and the illumination environment is fixed for all videos. The age range is 8 to 76. The resolution of the video frame is 1920×1080. This database has four main parent-child kin relations and three sibling relations. According to gender, sibling relations include Sister-Sister, Sister-Brother and Brother-Brother. The shortcoming of this database is that it only has 95 kin pairs and the subjects are mostly Caucasian.

KFVW

Yan *et al*. [23] proposed the KFVW[9] (Kinship Face Videos in the Wild) database. KFVW is made up of facial videos. The difference from the video-based kinship database (UVA-NEMO Smile database) is that KFVW is collected in the natural environment. Videos have no constrain in illumination, pose, occlusion, background, expression, age, etc. KFVW has 418 pairs of facial videos. Each clip of a facial video has 100 to 500 frames. The resolution of the video frame is 900×500. These videos are all from the Internet. KFVW has four main kin relations. Compared with the UVA-NEMO Smile database, KFVW contains more data, but it doesn't have a family structure and each subject has only one clip of facial video.

KIVI

The KIVI[10] [137] database was collected by Kohli *et al*.. It is organized with the family structure containing facial videos of 503 subjects from 211 families. It has 355 kin pairs in total. The database is downloaded from the Internet. The average length of the video is 18.78 seconds, and the frame rate is 26.79 frames per second. There are a total of 250,000 frames. The KIVI has four main kin relations and three sibling relations. KIVI dataset [10, 138] is organized with the family structure and contains facial videos of 503 subjects from 211 families. The dataset is downloaded from the Internet.

[9]https://www.kinfacew.com/datasets.html
[10]http://iab-rubric.org/resources/KIVI.html

2.6.3 Evaluation metrics

In kinship verification experiments, the data is usually divided into positive pairs and negative pairs. The positive pairs are all pairs with kin relations in the dataset, while the negative pairs are most often randomly generated among the image pairs without a kin relation. Generally, when establishing a protocol, the number of positive pairs and negative pairs is balanced, although the creation of additional negative pairs has also been explored [15]. The most typical evaluation protocols are based on N-fold cross-validation with the intent to reduce overfitting. In the most typical 5-fold configuration, four folds are used as training data, while the remaining one is used for testing. After repeating the process through all five testing folds, we can compute the final result with the average accuracy of each one of the five. Notably, in this configuration, the positive and negative pairs should only be generated within each fold.

Verification accuracy is the typical assessment criteria in kinship verification studies. Given True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the accuracy A is defined as:

$$A = \frac{TP + TN}{P + N}. \tag{1}$$

2.7 Conclusion

Automatic facial kinship verification essentially use image feature extraction methods and machine learning methods to analyze the similarity between two different facial images to verify whether or not they have some kin relationship. Comparatively, human eyes have difficulties in quantifying the similarity of two images from different people solely by the sensory perception of the human eye. Features such as the distance between the eyes, the shape, and size of the facial parts are not easily judged at a glance. In addition to facial shapes, the human eye's ability to distinguish color is low. Thus during the procedure of kinship verification, the human brain is using fuzzy judgment, resulting in low recognition accuracy. Complementing to the human eye's judgment on kinship verification, computer vision methods can accurately capture the similarity between parents and children's faces in terms of shape and color, as well as distinguish the difference from non-kinship images. Machine learning methods are based on mathematical measurement. Through optimizing the feature extraction methods and classifier, machine learning methods are more accurate in inferring whether or not two facial images have kin relations.

Over the last decade, image-based kinship verification techniques have been developed to a great extent. In the early research, the first attempts demonstrated the

possibility of automatically verifying kinship with computer vision methods. With machine learning methods showing great potential, kinship verification got increasing attention in the field. Advances were made with many methods proposed. Brand new ideas of problem formulation [11, 91, 115] and algorithms from different disciplines [116] were raised. More recently, deep learning methods [15, 14] have emerged and have shown powerful learning capability on large datasets. Different methods have shown different levels of progress in specific tasks. However, most current research focuses on kinship verification from facial images, while only a few works consider facial videos. Videos provide an abundance of information that can be leveraged to compensate for the limited temporal information of individual still images. Compared with kinship verification from images, research on video analysis is found on a limited scale, especially for the unconstrained video-based kinship verification. How to learn kin features from videos (*e.g.,* dynamic facial features, multi-modal features such as voice, gait, and gestures) is a key research direction.

Although the publicly available kinship databases, including both image and video datasets, have been an important kick-off for developing useful kinship approaches and applications, these datasets are limited in presenting multiple modalities of human biometrics, such as speaking voice. Until now, kinship video datasets are generally very small and consist of the mono face modality. Generally, only hundreds of subjects are included in the dataset due to the difficulty of collecting subjects' facial videos. Unlike facial images that depict single individuals and can be obtained automatically by web crawlers (since facial images sometimes have identity tags), facial videos usually come with multiple non-kin subjects and scene transitions and require careful curation. Thus, to study and evaluate the vocal usefulness and audio-visual fusion on the kinship verification problem, new multimodal datasets are needed.

3 Audio-visual kinship datasets

Visual kinship verification using facial information has been studied extensively. However, the related research neglects the usage and benefits of other biometric modalities, such as voices. This is mainly because the existing kinship datasets cover a variety of visual facial data but lack human talking data. Based on papers V and VI, in this chapter we present two new audio-visual kinship datasets. Unlike the existing kinship datasets, we include both human faces and speaking voices in our datasets called, TALKIN and TALKIN-Family, respectively.

3.1 Introduction

The last decade has witnessed great advances in facial kinship verification, where the research community devoted many efforts to advance the study of facial kinship verification by developing methods, databases, and applications [139]. Due to the various applications of kin-related tasks, complementary kin research problems have emerged with the publication of new relevant datasets. We identify those studies as Tri-subject kinship verification [19], Family classification [63], and Family search and retrieval [20]. To carry out research on the above-mentioned problems, various kinship databases have been designed, such as the UB Kinface dataset [111, 60], the TSKinFace dataset [19], the Family 101 dataset [63] and the FIW dataset [20]. In recent years, the image-based kinship datasets have extended to facial videos, including UvA-NEMO Smile [21, 22], KFVW [23], FFVW [140] and KIVI [10]. Table 2 compares the main characteristics of existing kinship datasets. We categorize those datasets based on data modality.

Even though certain progress has been achieved, the current kinship recognition research still focuses mainly on analyzing facial features only, mainly because the existing kinship datasets consist only of single-modality data (*i.e.,* faces). In the real world, the data usually comes from multiple modalities. Multiple modalities provide complementary information about the data and potentially help to enhance the robustness of the verification system. As for the kinship verification problem, besides the faces that perceive heritage features, voices have also been proven to contain generic features [141, 27, 26, 28, 29].

To promote the kinship study to a new stage and bridge the gap between current kinship datasets and real-world usage, in this thesis, we introduce the multiple modality analysis (*i.e.,* Faces and Voices) into FKV research, as *audio-visual kinship verification*.

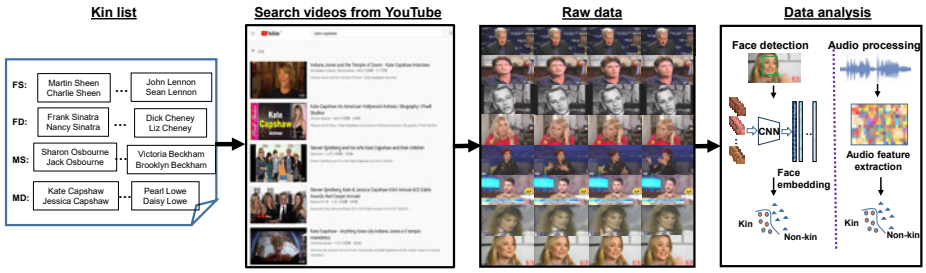


Fig. 7. The collection pipeline TALKIN dataset. Reprinted, with permission, from Paper V ©2019 IEEE.

Thus, a carefully designed multimodal kinship dataset is desired. In our first attempt, we propose a new kinship dataset composed of facial videos with speaking voices, called the **TALK**ing **KIN**ship (**TALKIN**) dataset. The TALKIN dataset is organized with a pairwise structure where the video data comes in kin pairs. Based on the TALKIN dataset, to enrich and extend the data in aspects of volume, organization, and data variety, the TALKIN-Family is then proposed. In the following chapters, we introduce those two datasets in detail.

3.2 The TALKIN dataset

We first introduce a new kinship dataset called TALKIN. It contains several videos of subjects talking in the wild environment (under unconstrained background, illumination and recording condition, etc.). The purpose of collecting this is to investigate the newly raised problem, audio-visual kinship verification in the wild.

3.2.1 Data collection pipeline

The overall collection pipeline for the TALKIN dataset is shown in Figure 7.

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 number for each relation is 100 pairs of videos. Most of the list is formed of celebrities, such as musicians, actors, and politicians, with the remaining pairs obtained from reality TV series that involve family interactions of noncelebrity individuals.

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

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

Table 2. Main characteristics of existing kinship datasets. We sort those datasets by the data modality. In the early years, many image kinship datasets were proposed. Then some video datasets with aligned facial information were proposed.

Modality	Dataset	Year	Size	Family structure	Multiple samples	Controlled environment
Image	CornellKin [8]	2010	150 pairs	No	No	No
Image	UB Kinface [111, 60]	2011	200 groups	No	No	No
Image	Family 101 [63]	2013	101 families	Yes	Yes	No
Image	KinFaceW-I [11]	2014	533 pairs	No	No	No
Image	KinFaceW-II [11]	2014	1000 pairs	No	No	No
Image	TSKinFace [19]	2015	1015 groups	No	No	No
Image	FIW [20]	2016	1000 families	Yes	Yes	No
Image	WVU [142]	2017	113 pairs	No	Yes	No
Video	UvA-NEMO Smile [21, 22]	2012	1240 videos	No	Yes	Yes
Video	KFVW [23]	2018	418 pairs	No	No	No
Video	FFVW [140]	2018	100 groups	Yes	No	No
Video	KIVI [10]	2019	211 families	Yes	No	No
Video, audio	TALKIN [124]	2019	400 pairs	No	No	No
Video, audio	TALKIN-Family [143]	2022	246 families	Yes	Yes	Both

videos from *different* video clips corresponding to different backgrounds or recording conditions.

Step 3. Data preparation. For face detection and alignment, we use the MTCNN algorithm [144] to detect five face landmarks in every frame of the video. Finally, the videos are cropped according to the landmarks. The face frames are resized to 224 $\times$ 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) [46], and Deep Neural Networks are used to embed the audio features.

3.2.2 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 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 to 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 3 shows the distribution of ethnicity in TALKIN. The distribution is counted by kin pairs rather than individuals, since the image of one parent might appear multiple times with more than one kid. Note, however, that we exclude mixed-race samples, *i.e.* the parent and child in a sample pair always have the same ethnicity. The dataset has two parts: video and audio. The length of the video varies from 4.032 seconds to 15 seconds with a resolution of about 1920 $\times$ 1080. Audio is extracted from video files. The sample rates are all set at 44.1 kHz. Besides the varied text content, the audio files contain substantial channel variations (due to different recording devices, for example). Some of them also contain reverberation and additive noise.

Table 3. Ethnicity distribution (%) of the TALKIN dataset. Reprinted, with permission, from Paper V ©2019 IEEE.

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

3.3 The TALKIN-Family dataset

In our previous preliminary attempt, we collected the TALKIN dataset. However, the TALKIN dataset has some obvious limitations, *i.e.,* a limited number of training samples, limited diversity in terms of environmental conditions, kinship categories, and mono-annotation with *binary kinship labels* only. To address some of these identified limitations, we aim to establish a new audio-visual kinship dataset named **TALKIN-Family** that consists of facial videos and synchronous speaking audio with properties that differ from the existing one. Specifically, we consider improving the TALKIN dataset in three aspects. First, to make the audio-visual kinship dataset more applicable in recognizing kinship from different tasks, we propose to arrange the data with the family structure instead of simple kin pairs. Within the family structure, beyond limited parent-child relations, more kinship categories could be generated, *e.g.,* siblings and grandparent-grandchild relations. Then, besides enlarging the dataset by increasing the number of subjects, we also intend to collect multiple samples for each subject under different conditions of environment background (white and non-white ones) and speaking content (fixed and free text). Lastly, we consider providing more biometric labels for our data, such as age, gender, family relation labels and scene conditions. This could help in promoting future deeper kinship recognition studies, since the related factors could possibly affect the recognition performance as has been discussed in [139]. We found that video-sharing websites such as YouTube usually contain free-style speaking videos while lacking fixed-text speech. To fill in the blansk, we chose to collect the TALKIN-Family offline. The video recording task is distributed to the participating families, and family members record the qualified videos by following the provided instructions. We will introduce the collection steps in detail in the following section.

3.3.1 Collection pipeline

The overall collection pipeline is shown in Figure 8. TALKIN-Family was collected offline by recruiting several participants. The participants were asked to record frontal talking facial videos of themselves and biologically related family members. To eliminate family-related biases (*e.g.,* recording conditions, recording devices, speech contents), we set up several recording protocols.

Participants. The subjects involved in the recording mission within one family should be biologically related. The number of subjects within one family should be more than two, including collateral relatives and direct relatives across generations. This

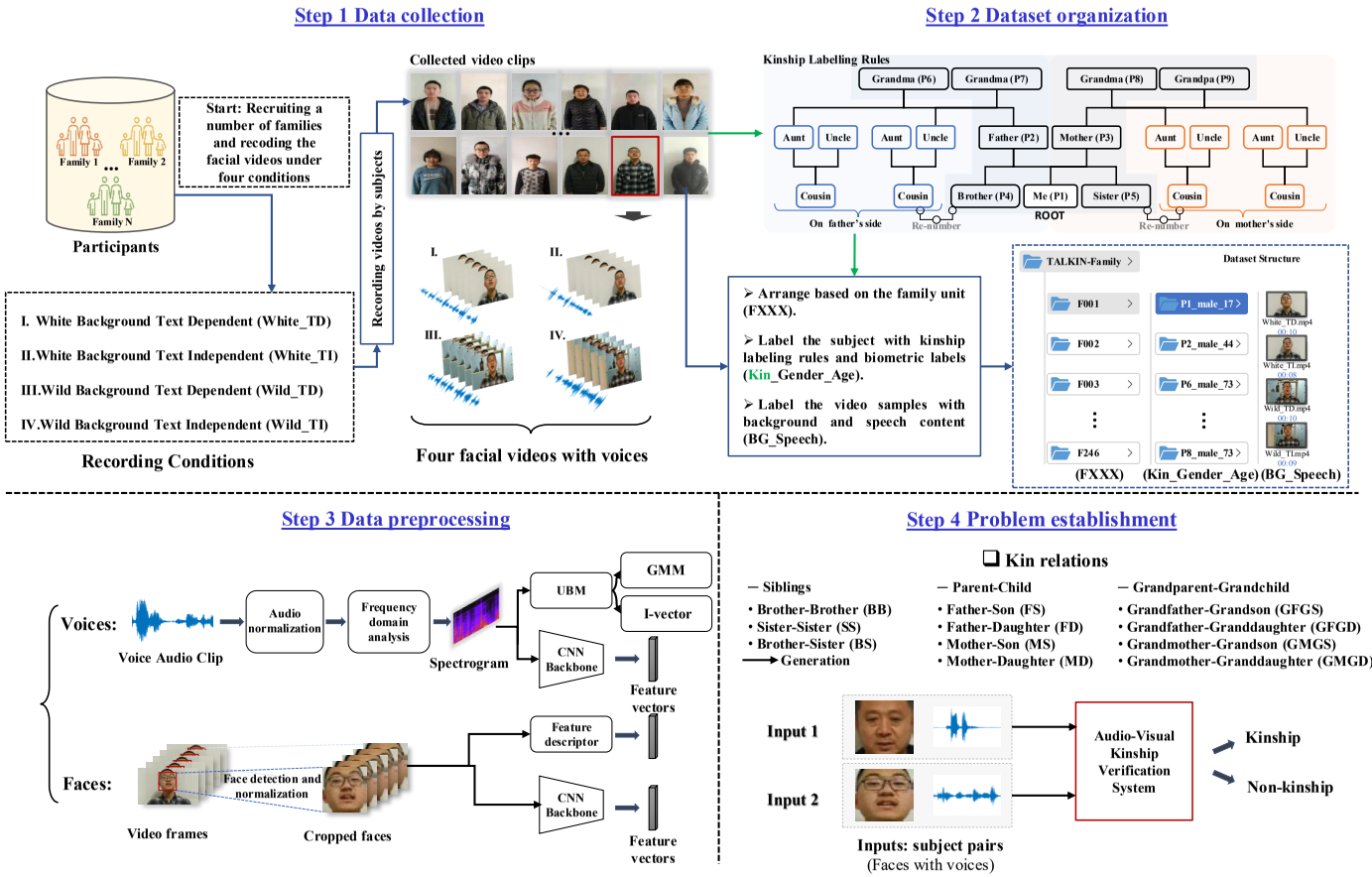


Fig. 8. The overall collection pipeline for the TALKIN-Family dataset. The data in TALKIN-family is collected offline by recruiting a number of families. Subjects participate the data collection with their family. Each subject has four facial talking videos under two background and two speech conditions. The TALKIN-Family is organized with family structure, and in each family, people are labeled according to our kinship labelling rules. Then, we do data pre-processing with audio and facial video separately. To study the audio-visual kinship verification, we define the problem with different kin relation types.

means that collateral relatives cannot be considered an isolated family. Subjects across different families should have no biological relation.

Environment conditions. The background should be quiet without noise or voices from other people. There is only one subject that appears in the video. To further ensure that videos within one family do not have only one background, we ask the subject to record videos against both the white and the non-white background. We refer to the white background as "white" and the non-white background as "wild", as shown in Fig. 8. This could eliminate the familial background bias [145] by generating kin pairs across different backgrounds.

Speech content. In speaker verification studies, *text-dependent speaker verification* and *text-independent speaker verification* are considered different tasks. Text-dependent speaker verification is characterized by using fixed speaking content [146]. On the other hand, in text-independent speaker verification, subjects talk freely without explicit cooperation [147]. In our dataset, to facilitate an extensive usage of the TALKIN-Family dataset we consider both scenarios. In addition this could facilitate avoiding bias due to family-wise spoken utterances. The participants were provided with the specific content (that is the Mandarin new year greeting). After the fixed speech content, in separate videos, they were asked to speak freely, using any desired sentences different from the provided content. The abbreviations for text-dependent and text-independent tasks are TD and TI. Therefore, for each subject, there are four talking videos, referred as *BACKGROUND_CONTENT* (*i.e., White_TD*, *White_TI*, *Wild_TD* and *Wild_TI*), as shown in Fig. 8.

Shooting device. The videos were recorded by a smartphone camera. The phone should be held still during recording, and the retouching and video editing functions were turned off. Within one family, the videos were recorded using multiple (more than one) phones, in order to minimize possible bias due to the device characteristics. Each video lasts for about 10 seconds.

Data packing. We set the principle subject as ROOT ("me"), who is one of the young generations. Family members are backtracked based on the root, and the family tree is generated and labeled as in Fig. 8. Every involved single-family has a family folder as FXXX (*i.e.,* F001-F246). In addition, the gender and age labels were also collected. Under the family folder, each subject has a sub-folder, *ID_GENDER_AGE* (*e.g.,* P1_female_6), where *ID* refers to the subject's family role defined by the family tree. *GENDER* is male or female, and *AGE* is an integer referring to the subject's age. In the subject's folder, four facial videos are stored.

3.3.2 Data preparation

In the TALKIN-Family dataset, each video clip was recorded with the cooperation of the participants, and only one subject appears in each video. Therefore, Speaker Diarization [148] is not required to determine "who spoke when" before data pre-processing. We do the pre-processing from visual data and audio data separately, as described below.

- **Visual data.** We first extract facial frames from each video, and the faces are automatically detected, cropped, and aligned as done in [69]. Note that some recorded videos are shot in landscape mode or upside down. Therefore, in such cases, face orientation and image rotation are needed during face detection. Then the facial frames are resized to 224×224 and encoded by face-image descriptors. Chapter 5.5 details the face descriptors we employed in the experiments, including traditional methods and deep encoders.
- **Audio data.** Since the subject starts to talk and ends right after the subject stops, we extract the audio directly from the videos and save them as WAV files. The signal is converted and normalized to the single channel at a 44.1 kHz sample rate. Standard methods in the speech field, MFCCs [46] and Deep Neural Networks are used to embed the audio features.

3.3.3 Dataset statistics

Familial information

TALKIN-Family is organized with family structure, and it contains 246 families. Each family has 2 to 14 family members. Three levels of generation (Siblings, Parent-Child and Grandparent-Grandchild) are involved with 1012 subjects and 4048 clips of videos in the dataset. The age of the subjects varies between 5 years and 81 years old.

Data details

Most videos in TALKIN-Family were recorded indoors. The length of each video clip is about 10 seconds. In total, TALKIN-Family has 9.2 hours of videos. There are about 1 million facial frames in TALKIN-Family. All subjects are from China and speak Mandarin Chinese, although some of them have different accents.

3.3.4 Problem statement

We address the audio-visual kinship verification as a binary classification problem: given a pair of signals (a pair of video sequences with speech utterances, *e.g.,* $(\mathbf{X}, \mathbf{Y})$), the objective is to automatically determine whether they have a kin relation. In practice, we represent $\mathbf{X}$ and $\mathbf{Y}$ using recording-level representations. The kinship score, a numerical indicator associated with higher values for kin relation pairs, is obtained by computing similarity score between the feature representations. Three levels of generation (Siblings, Parent-Child and Grandparent-Grandchild) are considered in our experiments.

3.4 Conclusion

In this chapter, we introduced our newly established audio-visual kinship datasets: TALKIN and TALKIN-Family. These two datasets are mainly presented for the research on audio-visual kinship verification. The TALKIN dataset consists of 100 pairs of facial videos with synchronous subject's speaking voices for each kind of parent-child relation. Though the TALKIN dataset is small, it opens an era of the kinship verification study of audio-visual modalities. Then, to extend and develop a more comprehensive dataset, we propose the TALKIN-Family dataset, which is organized by family. In total, it consists of 246 families and 9.2 hours of videos. The TALKIN-Family dataset is not only superior in terms of dataset size, it also contains diverse samples with respect to recording background and speaking content. In the next chapters, we explore the methods of FKV from visual features, vocal features and the fusion of both.

4 Kinship verification from visual features

Kinship verification from faces is a challenging task that has been attracting increasing attention in recent years. Most of the traditional methods have mainly focused on analyzing only the luminance of the face images, hence discarding the chrominance (i.e. color) information which can be a useful additional cue for verifying kin relationships. Moreover, the small datasets suffer from limited training data. To mitigate those problems, we propose to combine color features and extreme learning machines. In this chapter, we summarize our findings in publications III-IV, which demonstrate the importance of color texture features and ELM classifiers.

4.1 Introduction

A key question in kinship verification from faces is about which facial parts exhibit the kin relation the most. In other words, what are the most shared facial features between family members? This question has been studied from psychological perspectives (e.g. [5, 3]), suggesting that the eyes may bear more kin information than other facial parts. Most of the proposed traditional methods in the literature for automatic kinship verification have mainly focused on analyzing only gray-scale face images, hence discarding color information that can be a useful additional cue for verifying kin relationships. From a biological point of view, the chromaticity of the face is tied to genetically expressed features, such as eye color and skin tone. These inherited features are many times present in kin-related persons in a similar manner. This chapter investigates the usefulness of color information in the verification of kinship relationships from facial images. We aim at answering the question of whether or not color helps to improve kinship verification. For this purpose, we compare the performance of several baseline methods used in their traditional gray-scale variants against their counterparts using color information. More specifically, we extract joint color-texture features to encode both the luminance and chrominance information in the color images. The kinship verification performance using joint color-texture analysis is then compared against their counterpart approaches that use only gray-scale information.

The most recent studies using deep CNNs have not shown their full potential due to limited training data. This makes the topic of kinship verification from facial appearances an exciting and still open research problem. We propose to tackle the kinship verification challenge by extracting color texture features and using ELM for classification. Our approach is motivated by two observations: (1) color texture features

are proven to provide significant enhancement over gray-scale counterparts and (2) ELM seems to deal better than deep neural architectures when facing small training sets.

4.2 Kinship verification based on color texture analysis

To study the usefulness of color in automatic kinship verification:

1. We considered three baseline methods successfully used in automatic kinship verification, namely, Local Phase Quantization (LPQ) [133], Binarized Statistical Image Features (BSIF) [149] and Neighborhood Repulsed Metric Learning (NRML) [11].
2. We considered three color spaces, RGB, HSV and YCbCr.
3. As the baseline methods were originally designed for grayscale images, we extended the methods to include color information by considering a joint texture-color analysis and combining the features extracted from different color channels (e.g. R, G and B).
4. We compared the performance of these three baseline methods (LPQ, BSIF, and NRML) in the three color spaces (RGB, HSV, and YCbCr) against their performance in the gray-scale space (Gray).
5. We experimented with two different benchmark kinship face databases, the Tri-subject Kinship Face Database (TSKinFace) [19] and the Kinship Face in the Wild dataset (KinFaceW-I & II) [11].

Our kinship verification scheme is depicted in Figure 9. In a nutshell, a target pair of face images is given as input. These two images are first converted into different color spaces (e.g. HSV). Then, the features (e.g., BSIF) are separately extracted from each channel of the considered color space (e.g., H, S, and V). The features are then concatenated to form one enhanced feature vector. Finally, we apply cosine similarity between the feature vectors of the pair of face images. The cosine similarity between the two vectors X and Y is defined as below:

$$\mathrm{sim}(\mathrm{X},\mathrm{Y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm, and X,Y refer to parent and child feature vectors.

4.3 Classification using extreme learning machines

Figure 10 depicts an overview of our proposed approach. The input is a pair of two color face images, e.g., a parent and a child. We convert these images into different color spaces and encode the facial texture in each channel. We compute the cosine similarities between the texture features in each color channel. These similarities are

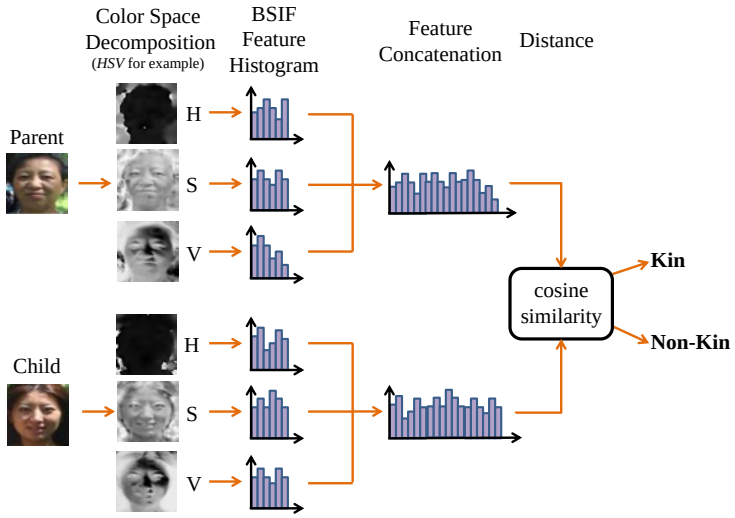


Fig. 9. An illustration of the proposed classification method. Reprinted, with permission, from Paper III ©2016 IEEE.

fed to an extreme learning machine (ELM) classifier. The ELM classifier is trained to predict whether or not the two persons are kin related.

An Extreme Learning Machine (ELM) [150] is a single hidden layer network that has been shown to perform better and faster than SVM in some classification problems [151]. The output of an ELM network with L hidden neurons can be represented as:

$$\sum_{i=1}^{L} \beta_i g\left(W_i \cdot X_j + b_i\right) = o_j, j = 1, \cdots, N, \tag{3}$$

where β_i is the weight between the hidden layer and output layer, and $g(x)$ is the activation function. $X_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T$ is the input vector with the ground truth $t_i = [t_{i1}, t_{i2}, \cdots, t_{im}]^T$. W_i and b_i are the weight and bias of the hidden layer. One key feature of ELM is to randomly set both W_i and b_i to speed up the training process. The distances between the ground truth and the actual output $\sum_{j=1}^{N} \left\| o_j - t_j \right\|$ should be minimized. The output weights are optimized by minimizing the approximation in:

$$H \cdot \beta = T, \tag{4}$$

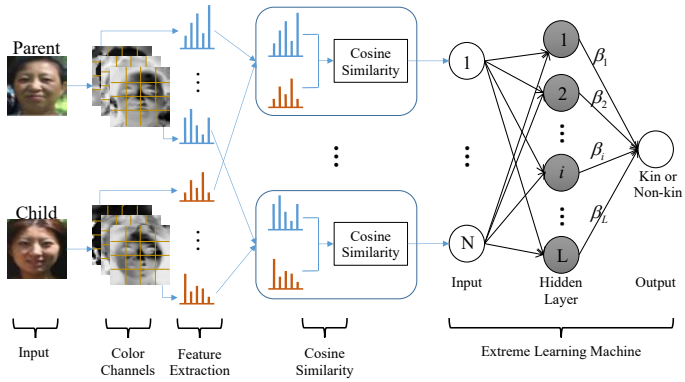


Fig. 10. Overview of the proposed color texture features with ELM for kinship verification. Reprinted, with permission, from Paper IV ©2018 IEEE.

where H is the randomly generated hidden layer output matrix,

$$H(W_1,\cdots,W_L,b_1,\cdots,b_L,X_1,\cdots,X_L)=\begin{bmatrix} g(W_1\cdot X_1+b_1) & \cdots & g(W_L\cdot X_1+b_L) \\ \vdots & \cdots & \vdots \\ g(W_1\cdot X_N+b_1) & \cdots & g(W_L\cdot X_N+b_L) \end{bmatrix}_{N\times L}, \tag{5}$$

and T is the target output:

$$T=\begin{bmatrix} T_1^T \\ \vdots \\ T_N^T \end{bmatrix}_{N\times m}. \tag{6}$$

The optimization procedure in ELM can be reduced to computing the Moore-Penrose inverse of H, determined at the beginning of the training, rather than optimizing β using a gradient descent algorithm by tuning the parameters in an iterative algorithm as in deep architectures. Thus, $\hat{\beta}$ can be calculated as:

$$\hat{\beta}=H^{-1}\cdot T. \tag{7}$$

4.4 Experimental results and analysis

For experimental evaluation, we considered three commonly used kinship databases: KinFaceW-I, KinFaceW-II [11], and TSKinFace [19]. These three databases are composed of four types of kin relations, namely father-son (FS), father-daughter (FD), mother-son (MS), and mother-daughter (MD). As TSKinFace was originally formed

Table 4. Kinship verification accuracies (in %) on the TSKinFace database. Reprinted, with permission, from Paper III ©2016 IEEE.

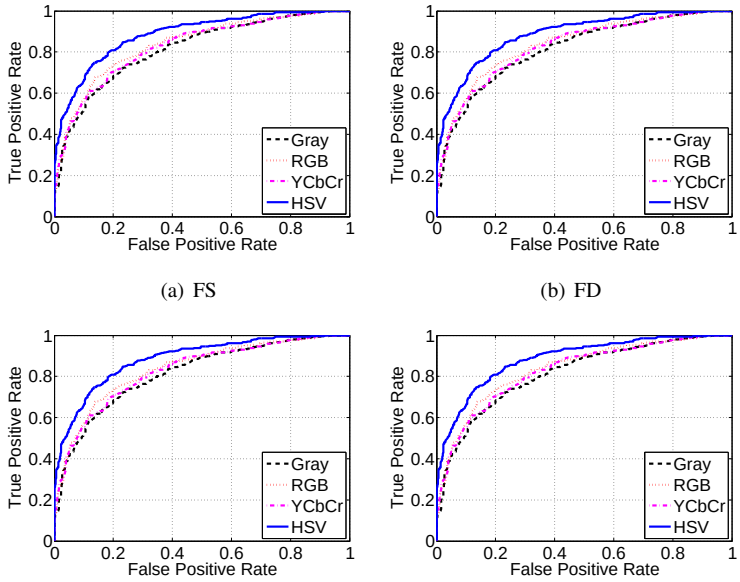
Method	F-S				F-D			
	Gray	RGB	YCbCr	HSV	Gray	RGB	YCbCr	HSV
NRML [11]	73.5	74.5	77.8	81.3	73.1	74.1	76.0	79.2
LPQ [133]	73.5	76.7	76.6	80.1	70.7	73.1	73.6	79.3
BSIF [149]	75.8	77.9	77.0	81.5	73.1	76.4	76.2	81.4

Table 5. Kinship verification accuracies (in %) on the TSKinFace database. Reprinted, with permission, from Paper III ©2016 IEEE.

Method	M-S				M-D			
	Gray	RGB	YCbCr	HSV	Gray	RGB	YCbCr	HSV
NRML [11]	72.4	74.0	75.0	78.8	70.1	71.7	77.0	77.5
LPQ [133]	72.8	74.4	74.0	80.7	71.5	73.8	77.0	80.3
BSIF [149]	75.6	76.9	77.6	79.9	73.4	76.1	76.6	82.0

of triple relations (two parents and one child), we convert each triplet into two pairs: father-child and mother-child. The number of positive and negative pairs used in the experiments is the same for each relation in the three databases. We use a five-fold cross-validation strategy for the evaluation. We report the mean accuracy over the five folds. The negative pairs and folds are predefined for KinFaceW-I and KinFaceW-II. In the case of the TSKinFace database, we randomly generate negative pairs and folds. For the color texture features, we extracted color-BSIF [149] as this has been shown to perform better than color-LBP and color-LPQ [152]. The dimensionality of each face block feature is reduced using PCA before computing the cosine similarities. For ELM, the number of neurons in the hidden layer is an important parameter. The number is determined empirically and set to 40.

Tables 4 and Table 5 show the obtained results on the TSKinFace database, comparing the performance of the three baseline methods (LPQ, BSIF, and NRML) in the three color spaces (RGB, HSV and Y-CbCr) against the performance of these baseline methods in the gray-scale space (Gray). From these results, we can clearly see that the use of color consistently improves the kinship performance compared to the corresponding gray-scale approaches. When comparing the different baseline methods, we can see that BSIF yields the best overall performance, outperforming the NRML-based metric learning method. When comparing the color spaces, although improvement can be seen in all of them, HSV seems to give the best overall performance. This superiority can be noted for all individual folds. A statistical ANOVA test performed across all five folds



(a) FS (b) FD

(c) MS (d) MD

Fig. 11. ROC curves of different color spaces using the best performing baseline feature (BSIF) on the TSKinFace database obtained on the 11(a) F-S set, 11(b) F-D set, 11(c) M-S set, and 11(d) M-D set, respectively. Reprinted, with permission, from Paper III ©2016 IEEE.

and all four relationships showed that the better performance of HSV is statistically significant. Hence, we can conclude that color does provide some discriminative information that can help in boosting the kinship verification performance. Figure 11 shows the obtained results, validating our findings and pointing out the usefulness of color information for kinship verification. The ROC curves are shown for the best performing feature (BSIF) for grayscale and three different color spaces.

Our proposed method is compared against some recent state-of-the-art methods on KinFaceW-I and KinFaceW-II databases in Table 6. Note that some of these methods, such as MultiviewSSL, use a combination of different features to describe a face image. Some other methods are based on deep learning. On the KinFaceW-I database, our method gives the best performance on the MS subset. For KinFaceW-II, our approach gives the best results for two subsets: MS and MD. On the larger TSKinFace database, our approach yields the best results for all four kinship subsets. These results are promising, and they demonstrate that our proposed approach is competitive compared to recent methods for kinship verification.

Table 6. Kinship verification accuracies (in %) of the proposed approach and the state-of-the-art methods on the KinFaceW-I and KinFaceW-II databases. Reprinted, with permission, from Paper IV ©2018 IEEE.

Method	KinFaceW-I				KinFaceW-II			
	FS	FD	MS	MD	FS	FD	MS	MD
PDFL [71]	73.5	67.5	66.1	73.1	77.3	74.7	77.8	78.0
DMML [72]	74.5	69.5	69.5	75.5	78.5	76.5	78.5	79.5
NRML [11]	72.5	66.5	66.2	72.0	76.9	74.3	77.4	77.6
MultiviewSSL [95]	82.8	75.4	72.6	81.3	81.8	74.0	75.3	72.5
SSML [98]	81.7	75.3	71.4	77.9	82.4	78.6	79.8	77.9
SPML-P [106]	81.1	75.7	73.2	75.7	82.4	77.6	76.6	76.2
Proposed	70.0	64.2	73.0	77.2	78.6	73.6	81.0	79.6

4.5 Conclusion

In this chapter, we explored the FKV problem from visual features using color texture features and extreme learning machines. Most of the proposed traditional methods in the literature for automatic kinship verification have mainly focused on analyzing only gray-scale face images, hence discarding color information. When considering the color information, the problem usually consists in learning a discriminating color space where the classification (e.g. kinship verification in our case) becomes more affordable than with the gray-scale space. HSV has uncorrelated information in every channel. This is in contrast to RGB color space in which its R, G and B channels may contain redundant color information. Therefore, we first convert the RGB images into the HSV color space. Then, the binarized statistical image features (BSIF) [149] are extracted from each channel separately. Extensive experiments on three kinship databases (TSKinFace, KinFaceW I & II) showed good performance when color texture features are extracted for kinship verification. The proposed ELM approach is shown to perform well with the limited number of training data sets. The results obtained are comparable with those of recent state-of-the-art methods.

The presented methods show promising performance for FKV from visual features. It is also worth exploring whether or not the other features from modalities, such as voice, are useful for FKV. In the next chapter, we discuss leveraging the acoustic features that are learned from human speaking voices. Then, we systemically investigate the audio-visual kinship verification problem on the basis of visual and vocal features. It fuses the multimodal data to recognize the kinship with the aim of improving performance.

5 Audio-visual kinship verification based on deep learning

Over the past decade, many efforts have been devoted to improving the kinship verification performance only from human faces while lacking other biometric information, *e.g.,* speaking voice. 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 chapter, we propose the audio-visual fusion method for kinship verification by summarizing the findings in papers V and VI, which interpret and benefit from multiple modalities.

5.1 Introduction

The existing research in kinship verification has extensively focused on exploring kinship features from the visual modality of the facial images/videos [11, 14, 15, 123, 10]. Certainly, facial similarity plays an important role in FKV, as facial similarity and kinship judgments are highly correlated according to recent psychology research [7, 55]. However, there have been studies demonstrating that voice similarity is also related to kinship judgments [141, 27, 26, 28, 29]. For instance, according to [141], the vocal tract shape that affects voice properties is genetically determined. Consequently, subjects with kinship have a similar voice. In addition, the study of human perception of kin voice indicates that humans have the ability to judge kinship by listening to the speaking voice [25, 153]. Despite this evidence, voice modality has not been explored for FKV yet.

In recent years, audio-visual fusion has been shown to be an effective way of improving performance in various problems including emotion recognition [30], speech recognition [31], event detection [32], and biometrics [33] such as speaker identification and speaker authentication. Based on the aforementioned discussion, it is natural to ask: **In addition to the visual modality, is it beneficial to explore other modalities (specifically the voice channel) for the problem of FKV?** Therefore, in order to answer this question, in this chapter we carry out the first study that aims to build an audio-visual kinship verification framework, in an attempt to further improve the FKV performance.

We consider the design of the framework for the task of audio-visual FKV. It encompasses two main steps, *i.e.,* extracting appropriate features and integrating modality

information [154]. Representing modalities, *i.e., audio* and *video*, in an appropriate way is crucial before fusion. Visual features have been widely studied for FKV [20]. Comparatively, very few acoustic features are designed specially for kinship verification, because the study has been largely under explored. However, well-known acoustic representations such as Mel-frequency cepstral coefficients (MFCCs [46]) and data driven features [47, 48] have been commonly applied in the speech community. Similar to the correlation between facial similarity and FKV, we propose then computation of the voice similarity and set new benchmark methods for FKV by using acoustic features. In the second step, the idea is to fuse facial and vocal features, while at the same time learning the closeness of kinship and the discrepancy of non-kinship pairs in these feature spaces. 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 are required for their optimization [155]. Based on the Siamese architecture, we find that inter-modal discrepancy and modal weighting are essential to exploit informative knowledge. Motivated by the adversarial learning [49] strategy and the self-attention mechanism [50], we propose the fusion method named Unified Adaptive Adversarial Multimodal Learning (UAAML) based on deep neural networks (DNN), which addresses the aforementioned challenges. The UAAML jointly considers multimodal feature learning and kinship attention weights with similarity learning. Particularly, we introduce the L_2 norm layer [51] to generate the unified features before fusion and make the network training stable and efficient. Using multimodal learning, we conduct experiments on the proposed audio-visual kinship datasets to improve the performance of a single modality.

5.2 Related work

5.2.1 Acoustical study of kinship

In our daily lives, people with a kin relation can have similar voices. For instance, it is sometimes hard to distinguish between father and son over the phone. This phenomenon has attracted the attention of a few researchers to this field. In this context, we review works related to speaker verification and relate them to kinship acoustic analysis.

Speaker verification of identical twins

Voices from two different persons with a close kinship relation might be confusing, given the latent voice print similarity. One special case is the voices of identical twins, a case that was addressed almost five decades ago [156]. Ariyaeeinia *et al*. [157] studied *automatic speaker verification* (ASV) performance using the voices of identical twins. They reported that performance dropped when tested with twin voices, when compared to other speakers. Künzel *et al*. [158] studied the performance of commercial forensic automatic speaker recognition with identical twin data. The performance of the speaker verification system drops when tested with twins due to voice similarity, especially in the case of female twin sisters. The length and content of speaking also affect the performance to some extent. This phenomenon is an observation that poses the reverse question of how voice could contribute to the assessment of kinship verification for different relationships.

Kinship acoustic analysis

Based on the above findings, researchers explicitly studied the vocal similarity of kin people. The earliest genetics of voice research was found in the 1990s. Sataloff [141] demonstrated that voice function is related to the phonatory organ structures. The physical features are genetically determined, which intuitively indicates that the human voice is also genetically determined. Later, psychological studies assessed human perception in recognizing the kin voice. Studies by [25, 153] showed that humans could verify kinship from voices by providing listeners with the voice of specific sentences. Motivated by the above research, acoustic studies quantitatively confirmed voice similarity within kinship by measuring and comparing various acoustic characteristics [26, 29, 27]. Although many works have been carried out on studying the vocal similarity of kinship, voice has not been directly applied in automatic kinship verification.

5.2.2 Multi-modal learning

Multi-modal fusion methods can exploit complementary sources of information. Different sources of information are typically integrated through early fusion (feature level) or late fusion (score or decision levels) [34]. Feature-level fusion using concatenation or aggregation is often considered to provide a high level of accuracy. However, 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 sensitive to the impact of score normalization methods on the overall decision boundaries.

In the deep learning literature, Neverova *et al.* [159] proposed the multimodal dropout (ModDrop) for gesture recognition problems. First, the weights of each modality are pre-trained. Then a gradual fusion method is proposed by randomly dropping separate channels to learn cross-modal correlations while preserving uni-modality specific representation. When considering multi-modal fusion, one main challenge is eliminating the modal discrepancy and learning a joint feature space that can better fuse the features. Recent Generative Adversarial Networks (GANs) [49] have achieved significant success in mapping the data distribution into the desired one by adversarial training. Inspired by this, Mai *et al*. [160] built the encoder-decoder networks for different modal inputs to learn the latent feature embeddings. Adversarial learning was introduced on the encoder to learn the joint feature space for different modalities. Then a graph fusion network was applied to integrate encoded multimodal features. Zhou *et al*. [161] studied the multi-modal clustering problem. They developed the End-to-end Adversarial attention Multi-modal Clustering (EAMC) method, which consisted of the adversarial learning module and a modal attention module to align the feature distribution and quantify the important modal weights. A proposed clustering objective was added to guide the network training on the top of the network.

5.3 A siamese network for A-V fusion

This subchapter presents a new deep Siamese network for the fusion of face and voice modalities for accurate multi-modal 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 [162] CNN cascaded with an Long Short-Term Memory (LSTM) [163] network for the face modality. For the voice modality extracted from videos, we fine-tune a ResNet-50 pre-trained on VoxCeleb2 [47]. 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.

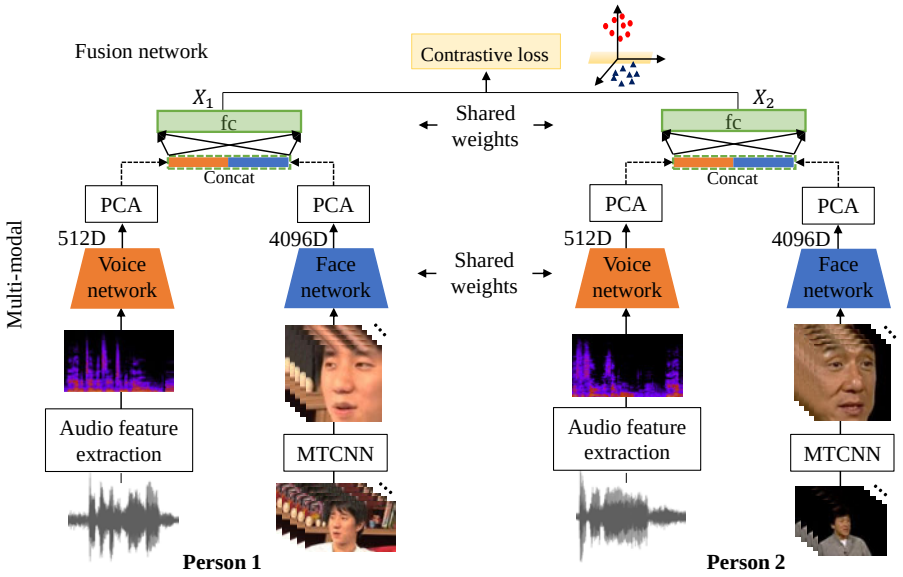


Fig. 12. Architecture of the proposed fusion method. Reprinted, with permission, from Paper V ©2019 IEEE.

Face network

We implement the VGG-face [162] CNN cascaded with an LSTM [163] network for facial representations. The 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 an RGB image with the size of $224 \times 224 \times 3$. As shown at the top of Figure 12, it is comprised of 13 convolution layers, each followed by a Rectified Linear Unit (ReLU). Some of them are also followed by a max pooling operator. The last three layers are FC layers. The first two FC layers have 4096 outputs and the last FC layer has N outputs as N-class predictions. We feed the facial frames one by one and collect the deep features from layer fc7. To integrate both spatial and temporal information, an LSTM layer with 4096 cells is stacked on top of it.

Voice network

In the previous research, the acoustic features are first extracted, and machine learning methods such as I-vector [164] and the Gaussian mixture model - universal background model (GMM-UBM) [165, 166] are used to analyze features. In this work (see the top of Figure 12), we use the ResNet-50 pre-trained with a large speaker verification dataset

called VoxCeleb2 [48], and then fine-tune 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 to 16 kHz to be the consistent with the VoxCeleb2 dataset. Then the audio samples are divided into 3-second segments. A Hamming window with 25ms width and 10ms step is applied to the audio. Following in the same manner as [48], 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.

Fusion network

We propose a deep Siamese network with contrastive loss [53] for kinship verification based on fusing videos and audio. The whole architecture is shown in Figure 12. For each voice and face network, we use contrastive loss to learn the intra-class similarity and inter-class dissimilarity among subjects. The contrastive loss is defined as:

$$L = \frac{1}{2N} \sum_{n=1}^{N} (y_n d^2 + (1 - y_n) \max(M - d, 0)^2), \tag{8}$$

where the 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, and y_n is the label of the sample pair. y_n equals 1 when the inputs have the kin relation; otherwise, y_n equals 0.

After training the face and voice networks, we can collect their features – 4096D features are extracted from the face network and 512D features are extracted from the voice network. Then, after performing PCA on them to reduce the dimension to 130, they are concatenated into a 260D feature and followed by an FC layer with 260 nodes. By adding contrastive loss during the 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 the 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* allows determining whether two inputs have a kin relation.

5.4 Unified adaptive adversarial multimodal learning approach

When fusing audio-visual features for the problem of FKV, based on our benchmarks and investigation, we find that inter-modal discrepancy and modal weighting are essential to exploit informative knowledge. Motivated by the adversarial learning [49] strategy and the self-attention mechanism [50], we propose the fusion method named Unified Adaptive Adversarial Multimodal Learning (UAAML) based on deep neural networks (DNN), which addresses the aforementioned challenges. The UAAML jointly considers multimodal feature learning and kinship attention weights with similarity learning. In particular, we introduce the L_2 norm layer [51] to generate the unified features before fusion and make the network training stable and efficient.

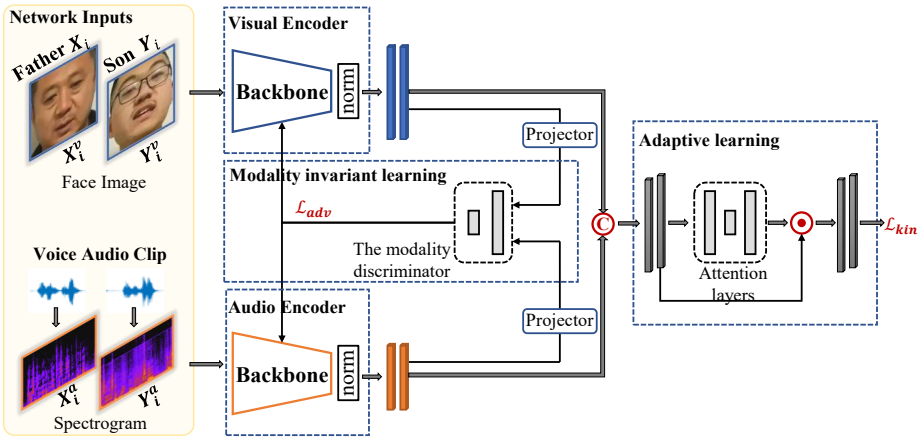


Fig. 13. The proposed unified adaptive adversarial multimodal learning method.

The overall framework of the proposed method is shown in Figure 13. It consists of modality-specific feature generators, modal fusion, and kinship assignment. The modality-specific networks are encouraged to exploit the distinct modal property. Then the modal fusion is trained to eliminate the cross-modal discrepancy to parse a better fusion of multiple feature vectors from different modalities. When obtaining the fused features, the contrastive loss is added to enforce the network to learn the compactness within kinship and separation between non-kinship.

5.4.1 Preliminaries

Let $\mathscr{D} = \{(\mathbf{X}_i, \mathbf{Y}_i, l_i) | i = 1, 2, \ldots N\}$ be the training set of N sample pairs, where $\mathbf{X}_i = \{\mathbf{X}_i^a, \mathbf{X}_i^v\}, \mathbf{Y}_i = \{\mathbf{Y}_i^a, \mathbf{Y}_i^v\}$. $\mathbf{X}_i$ and $\mathbf{Y}_i$ represent the ith sample pair that comes with both audio and visual modalities denoted by $\mathbf{X}_i^a, \mathbf{X}_i^v$ and $\mathbf{Y}_i^a, \mathbf{Y}_i^v$, respectively. The pairwise label l_i denotes whether the ith the sample pair has a kin relation, *i.e.,* $l_i = 1$ represents that $\mathbf{X}_i$ and $\mathbf{Y}_i$ have a kin relation, and $l_i = 0$ denotes that $\mathbf{X}_i$ and $\mathbf{Y}_i$ have a non-kin relation.

Our method has two feature encoders: the audio encoder $E_a(\cdot; \theta_a)$ and visual encoder $E_v(\cdot; \theta_v)$ that are parameterized by θ_a and θ_v. The audio and visual data are fed into the modal-specific encoder, and the feature representation is expected to be modal invariant. This is achieved by the adversarial learning associated with the discriminator $D(\cdot; \theta_d)$, where θ_d is the network parameter. Besides, to let the feature pay more attention to effective kinship traits and emphasize them, the attention mechanism is proposed to learn the weights for the feature-level fusion. The weight vector $\mathbf{w}$ is computed by the Multiple Layer Perceptron (MLP). The whole network is designed in the Siamese fashion that shares weights for two different inputs $\mathbf{X}_i$ and $\mathbf{Y}_i$. To preserve the kin discrimination of the network, we employ the contrastive loss L_{kin} to let the model learn the closeness of kinship pairs and the separation of non-kinship pairs.

5.4.2 Modality-specific networks

Different sources of data are difficult to combine at the raw data level. Therefore, we first adopt modality-specific networks to transform the face and voice data into the latent feature space. Following the work in [124], the network inputs are the facial image and a spectrogram computed from a particular speech. The residual network (ResNet) architecture [167] is adopted for both face and voice backbone networks, as described below. We take sample $\mathbf{X}_i$ as an example, which goes the same way to the input $\mathbf{Y}_i$.

Visual subnet

The visual backbone directly adopts the InsightFace with ResNet-34 architecture [168, 169]. Given an input facial image $\mathbf{X}_i^v \in \mathbb{R}^{D \times H \times W}$, we extract the corresponding feature embedding as $\mathbf{x}_i^v = E_v(\mathbf{X}_i^v)$. W and H indicate the spatial size, and D is the number of channels. As the facial image is cropped and resized to 112×112, the generated facial features fall into 512-D.

Audio subnet

The audio backbone employs the ResNet-50 pre-trained on Voxceleb2 [47, 170] to extract the vocal features from the spectrogram inputs. We extract a 3-second utterance clip and convert it into a single channel with a 16 kHz sampling rate. The spectrogram is generated by a sliding Hamming window of width 25ms and step 10 ms. Therefore, the audio network input $\mathbf{X}_i^a$ has the size of 512×300, and the corresponding output $\mathbf{x}_i^a = E_a(\mathbf{X}_i^a)$ is a 2048 dimensional feature embedding.

Similarly, we can have the audio and visual embedding for $\mathbf{Y}_i$ as $\mathbf{y}_i^a = E_a(\mathbf{Y}_i^a)$, $\mathbf{y}_i^v = E_v(\mathbf{Y}_i^v)$.

5.4.3 Model fusion

The modal fusion module fuses audio and visual features for comprehensive estimation. It consists of the unified feature operation, modal alignment, and feature fusion attention learning.

Multi-modal adversarial learning

When merging features generated from different modalities, they generally have different scales and norms. Directly combining these features leads to poor fusion performance since the larger features can overwhelm the smaller ones. Rather than carefully tuning the network parameters with efforts, Liu [51] found that normalizing the features before fusion improves the model stability. Therefore, before learning the modal invariant features, we add an L_2 normalization layer to transform the feature into a unified one. Formally, for the audio feature $\mathbf{x}_i^a$ and visual feature $\mathbf{x}_i^v$, we normalize them differently as $\hat{\mathbf{x}}_i^a$, $\hat{\mathbf{x}}_i^v$ using L_2-norm,

$$\begin{aligned} F(\mathbf{x}) = \hat{\mathbf{x}} &= \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \\ s.t.\ \mathbf{x} = \{x_1, x_2, \ldots, x_d\}&, \|\mathbf{x}\|_2 = (\sum_{i=1}^{d} |x_i|^2)^{\frac{1}{2}}. \end{aligned} \tag{9}$$

The audio and visual encoders learn multi-modal representations that may have a large gap between different modalities. Inspired by the recent generative adversarial networks [49], we introduce the discriminator $D(\cdot;\theta_d)$ to distinguish between the audio and visual features. Since the audio and visual features have different dimensions, we first feed them into one fully-connected layer that is $FC_a(\cdot)$ and $FC_v(\cdot)$ to map them

into a common length. Then the two-class classification is performed. The discriminator is optimized by the following objective function:

$$\begin{aligned}\min_{\theta_d} \mathscr{L}_d = -\mathbb{E}_{\mathbf{X},\mathbf{Y}\in\mathscr{D}} \sum_{i=1}^{N} & log(D(\hat{\mathbf{x}}_i^a)) + log(1-D(\hat{\mathbf{x}}_i^v)) \\ & + log(D(\hat{\mathbf{y}}_i^a)) + log(1-D(\hat{\mathbf{y}}_i^v)).\end{aligned} \tag{10}$$

One the other side, the modality-specific networks are trained to confuse the discriminator with the opposite modal label by minimizing the adversarial loss:

$$\begin{aligned}\min_{\theta_a,\theta_v} \mathscr{L}_{adv} = -\lambda_{adv}\mathbb{E}_{\mathbf{X},\mathbf{Y}\in\mathscr{D}} \sum_{i=1}^{N} & log(D(\hat{\mathbf{x}}_i^v)) + log(1-D(\hat{\mathbf{x}}_i^a)) \\ & + log(D(\hat{\mathbf{y}}_i^v)) + log(1-D(\hat{\mathbf{y}}_i^a)),\end{aligned} \tag{11}$$

where the λ_{adv} is the weight coefficient. The discriminator guides the modal encoders to learn the same distribution representations through min-max adversarial learning.

Feature fusion attention

After we obtain the modality-invariant representations, we concatenate the audio and visual features for two inputs as $\mathbf{x}_f = [\hat{\mathbf{x}}_i^a, \hat{\mathbf{x}}_i^v]$ and $\mathbf{y}_f = [\hat{\mathbf{y}}_i^a, \hat{\mathbf{y}}_i^v]$, the $[\cdot]$ denotes the concatenation operator. In particular, we design a fusion attention module to emphasize the efficient vector values. It consists of an MLP with the Sigmoid function, while the output is the weight vector $\mathbf{w}$ with the same dimension of $\mathbf{x}_f$ and $\mathbf{y}_f$, which can be calculated by:

$$\mathbf{w}_x = \sigma(FCs(\mathbf{x}_f)),\ \mathbf{w}_y = \sigma(FCs(\mathbf{y}_f)), \tag{12}$$

$$\mathbf{x} = \mathbf{x}_f \mathbf{w}_x,\ \mathbf{y} = \mathbf{y}_f \mathbf{w}_y, \tag{13}$$

where $\sigma(\cdot)$ is the Sigmoid function and $FCs(\cdot)$ is the stacked two fully-connected layers. $\mathbf{x}$ and $\mathbf{y}$ are the fused representations to get the kinship analysis. We denote the attention parameters as θ_{att}.

5.4.4 Learning kinship awareness embedding

To perceive the kinship traits, *i.e.,* similarity between kinship and difference between non-kinship, we adopt the contrastive learning scheme to train the network in a supervised

way. By integrating the kinship label l_i, the network objective can be expressed as:

$$\min_{\theta_a,\theta_v,\theta_{att}} \mathscr{L}_{kin} = \frac{1}{2N}\sum_{i=1}^{N}(l_i d^2 + (1-l_i)\max(M-d,0)^2), \tag{14}$$

where the threshold M is the margin, $d = \|\mathbf{x}-\mathbf{y}\|^2$.

During each training step, two multi-modal encoders are first trained alternatively in an adversarial way together with a discriminator without kin labels involved. Then the whole network is jointly trained using the kin labels. During the testing process, the cosine similarity $sim(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}$ is calculated to represent the distance between two inputs. A threshold applied to *sim* determines whether two inputs have a kin relation as has been done in [20].

5.5 Experimental settings

In this subchapter, we perform extensive kinship verification experiments on the TALKIN-Family dataset. We first compare the performance of our proposed method with other baseline methods. Then, an ablation study is conducted on our method. We also test the human ability to verify kinship from audio-visual videos. In the end, the benchmark experiments are also conducted in other environment settings.

5.5.1 Implementation detail

Our experiments evaluate audio-visual kinship verification without considering the video environment and what the subject has said. Our experiment is a frame-based architecture. It takes the frame-level inputs and then outputs the feature representations. During the test phase, we average the frame-level features extracted from one video as the final feature.

Data preparation for TALKIN

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 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.

Data preparation for TALKIN-Family

The TALKIN-Family is organized with a family structure. We first generate kinship pairs with 11 relationship types described in Chapter 3.3. After we obtain the kinship pairs (positive pairs), we split them into a maximum of five folds to conduct the K-fold validation [12, 20, 124]. Within each fold, we randomly generate non-kinship pairs as negative samples, where non-kinship subjects are from different families and biologically unrelated. The negative samples have the same size as the positive samples. Note that there is no family overlap between folds. The experimental data statistics distribution of audio-visual kinship verification in the wild is shown in Table 7. The stratified division in folds and relations is not possible for practical reasons, since certain relationships such as SS and BS, do not allow enough negative samples that belong to different families. We perform the data pre-processing on all videos for visual and audio data as introduced in Chapter 3.3. Since the video length varies from video to video and the neighbor video frames have a slight difference, we extract and align 60 facial frames and audio frames for each video. Due to extreme head pose variations and facial orientations that cause face detection failures, some frames of videos belonging to a few subjecs are discarded. Therefore, we could have about 60 pairs of facial frames and their corresponding utterance clips (3 seconds) as the multi-modal data sample for one kin sample pair.

Table 7. Data statistics for studying the audio-visual kinship verification in the wild on the TALKIN-Family dataset. # folds is the number of fold validations for each kin relation. # families and # subjects represent how many families and individuals are involved when studying the specific kin relation. # kin pairs is the number of kin pairs at the subject level. # videos is the total number of videos used, which is usually four times the number of subjects since each subject has four facial videos. # sample pairs is the number of frame-level sample pairs in each kin relation.

Relations	Siblings			Parent-Child				Grandparent-Grandchild			
	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD
# folds	5	4	4	5	5	5	5	2	2	3	3
# families	24	33	31	86	62	134	125	9	10	14	12
# subjects	50	70	73	196	136	308	285	19	21	31	27
# kin pairs	200	336	320	848	576	1296	1264	80	88	136	120
# videos	200	280	292	784	544	1232	1140	76	84	124	108
# sample pairs	11570	17530	15124	45444	30379	71740	69926	4328	4724	6181	6250

5.5.2 Compared methods

To verify the effectiveness of our proposed method on the TALKIN-Family dataset and compare the performance between the uni-modality and multi-modalities, we perform baseline methods on vocal and facial kinship verification and four fusion methods. We briefly introduce those methods below.

Voice features

We employ two traditional methods, GMM-UBM [165, 166] and I-vector [164], for audio analysis. We extract MFCCs with 12 cepstral coefficients from the audio samples. The UBM with 128 mixture components of GMM is trained with the training set.

- **GMM-UBM [165, 166].** The kin pair model is created from UBM using the Maximum A Posteriori (MAP) estimation. The verification likelihood is the log-likelihood ratio between speaker models and registered speakers' GMM.
- **I-vector [164].** UBM is trained using expectation-maximization (EM) with MFCCs. I-vector is obtained by MAP point estimate. Then the dimension of the I-vector is reduced by linear discriminant analysis (LDA). We compute the similarity between two speakers with the cosine similarity of I-vectors.

Besides, we also evaluate the pre-trained deep models as feature encoders.

- **pyannote-S.** The pyannote.audio [171, 172, 166] is an End-to-End generic PyanNet that is trained on Voxceleb [48] and Voxceleb2 [47] datasets. The trained model takes the utterance and samples it with a sliding window to generate overlapping 512-D features. The pyannote-S means we evaluate the performance using only the single vocal feature.
- **pyannote-A.** For the utterance clip, we average all audio features for the sequence as its final feature representation.
- **VGG_M.** The model architecture is based on VGG_M [173], and takes the audio spectrogram as input. The spectrogram is computed with the same method described in Chapter 5.4. VGG_M is trained on the Voxceleb dataset [48] with the task of speaker verification. The final audio feature has a length of 1024 dimensions.
- **ResNet-50.** The model is trained on the Voxceleb2 dataset [47] and the audio embedding is collected from the FC layer with a length of 2048.

Facial features

We consider four traditional facial descriptors, the popular metric learning method [11] and five deep facial image encoders for visual-based evaluation. For traditional features, we employ color texture features [174, 175] that are evaluated effectively for facial kinship verification.

– **BSIF [149].** Each facial image is divided into non-overlapping 2×2 blocks in each color channel. Each block is represented using 256 features and the whole face with $256 \times 4 \times 3 = 3072$ features.
– **LPQ [176].** For each facial image, we divide it into non-overlapping 2×2 blocks in each color channel. Each block is represented using 256 features, leading to a 3072-dimensional ($256 \times 4 \times 3$) feature representation for the whole face.
– **LBP [80, 81].** We divide the image into non-overlapping 4×4 blocks in each color channel. The parameters of LBP are that the radius is set as 1, the sampling number is 8. Fifty-nine histogram values are used to represent each block. Thus, each facial image is represented using $59 \times 16 \times 3 = 2832$ features.
– **LBP-TOP [177].** The spatial-temporal descriptor, LBP-TOP, is also evaluated in our experiments. The frames are converted to grayscale. Then the face frames are divided into 4×4 non-overlapping blocks. We extract 59 histogram features for each block volume in XY, XT, and YT planes, respectively. Thus, one video can be represented as $59 \times 3 \times 16 = 2832$ features.

Metric learning is commonly used in the kinship verification problem, since it is able to represent kinship compactness and non-kinship separation. Among those studies, MNRML [11] method has attracted the most attention.

– **MNRML.** We implement MNRML by using multiple feature descriptors, LBP, LPQ, and BSIF features, to learn the multi-view data metric.

Furthermore, deep CNN models pre-trained on large-scale face datasets are also widely used in kinship verification to encode the facial image with output embedding.

– **SphereFace** [178, 179] is a CNN model trained with angular softmax (A-Softmax) to learn more discriminative features. The SphereFace is trained on the face dataset CASIA-WebFace [180]. Then the deep features can be collected from the FC1 layer with 512 dimensions.
– **VGG-Face** network [162] is trained on a large face dataset with 2.6 million images of over 2662 people. We feed the facial image into the network, and collect features from layer fc7.

– **FaceNet-C** [181, 182]. FaceNet is a deep CNN model trained with the Triple-let loss. FaceNet-C means the model trained on CASIA-WebFace [180]. The output feature is a 512-D embedding.
– **FaceNet-V** [181, 182] means the FaceNet trained on the VGGFace2 [183] dataset.
– **InsightFace.** As mentioned in Chapter 5.4, in our method, we apply the InsightFace, that is a ResNet-34 architecture-based model [168, 169]. Compared with SphereFace, InsightFace utilizes the AcFace loss that has fewer parameters but a better classification margin. The model is trained on the MS1MV2 dataset. The facial frames are fed into the pre-trained model, and we can get the final 512-D feature embedding.

Fusion methods

We perform both early fusion and two late fusion methods on audio-visual kinship verification. The Siamese fusion proposed in our ICB 2019 paper is also compared.

– **Early fusion.** The multi-view features are concatenated together as the fused feature for the later similarity comparison.
– **Late fusion (mean).** For the late fusion, the similarity scores are computed separately for each modality. Then the mean fusion average scores were obtained from multi-modalities as the final decision score.
– **Late fusion (max).** Rather than calculating the averaged score, max fusion takes the maximum score as the final decision score.

Implementation details

For the Siamese fusion network, after training the face and voice networks, we collect the 4096 features from the face network and 512 features from the voice network. To make the dimensional balance of both facial and vocal representations, we conduct PCA to reduce both the facial and vocal feature dimensions to 130. Then they are concatenated into a 260-dimensional feature, followed by an FC layer with 260 nodes. During the training procedure, our system is trained on TALKIN, using backpropagation and contrastive loss to learn the correlation between parent and child based on audio visual modalities, which has no family overlap between the training and testing subsets.

For UAAML, we implement our network on the PyTorch library. Since the released pretrained InsightFace net and ResNet-50 (voice) are implemented based on the MXNet and Matconvnet libraries, respectively. We first convert those models to PyTorch formats using open source code from Github [184] and [185]. We use the weights of ResNet-34 trained on MS1MV2 [168] for the visual network and the weights of

Table 8. Verification accuracy (%) for the face modality on the TALKIN dataset. Reprinted, with permission, from Paper V ©2019 IEEE.

Techniques	FS	FD	MS	MD	Average
BSIF-Average [149]	61.5	58.5	61.0	59.5	60.1
LPQ-Average [176]	62.5	58.0	60.5	59.0	60.0
LBP-Average [80, 81]	61.5	60.0	59.5	61.5	60.6
LBP-TOP [177]	64.5	60.0	67.0	59.5	62.8
VGG + LSTM	76.5	69.5	70.0	71.5	71.9

Table 9. Verification accuracy (%) for the voice modality on the TALKIN dataset. Reprinted, with permission, from Paper V ©2019 IEEE.

Techniques	FS	FD	MS	MD	Average
I-vector [164]	63.5	60.0	63.0	63.0	62.4
GMM-UBM [165, 166]	59.5	59.5	66.5	60.0	61.4
Resnet-50 [47]	73.0	60.0	63.5	66.5	65.8

ResNet-50 trained on VoxCeleb2 [47] for the audio network to initialize our network parameters. Parameters in other layers are initialized using random weights. For training the proposed method, the parameters of the network are optimized using the Adam optimizer with a typical learning rate of 1e-6, weight decay of 1e-4, and mini-batch size of 50. We train the whole network for 250 iterations. The program is run using two Nvidia V100 GPUs (32 GB). The hyper-parameter λ_{adv} determines the degree of multi-modal discriminative information used during the model training process. In the case of using small λ_{adv}, no sufficient modality discrimination could be applied. We set the λ_{adv} with 1 [186].

5.6 Experimental results and analysis

This subchapter presents experimental results of audio-visual kinship verification on the TALKIN and TALKIN-Family datasets from both single modality and multiple modalities. We first compare the verification performance from faces and voice respectively. Then, the fusion method is evaluated on the multi-modalities. The performance effectiveness of single-modality against the multi-modal fusion is also compared.

5.6.1 Single-modal kinship verification

Experimental results on the TALKIN dataset

Table 8 and Table 9 present the results of experiments for uni-modal kinship verification on the TALKIN dataset, respectively. VGG-Face with LSTM shows better performance compared to traditional hand-crafted features. For voice-based kinship verification, Resnet-50 has better performance, except for the mother-son relation, which has a small drop compared to the GMM-UBM method. Overall, the proposed methods for uni-modal based kinship verification show good efficiency, especially when compared with baseline methods.

Experimental results on the TALKIN-Family dataset

Tables 10 and 11 show the kinship verification accuracy of single-modality (based on one modality) on the TALKIN-Family dataset. For voice-based kinship verification, ResNet-50 has the best performance. The traditional methods of I-vector and GMM-UBM have comparatively low performance. Notice that the Grandparent-Grandchild results are not provided because the UBM is hard to converge due to limited data. A possible solution is to employ external data to train UBM. Regarding the pyannote model, the performance can be improved slightly by averaging all vocal features within one utterance.

For the kinship verification from faces (Table 11), deep models outperform the traditional descriptors (LBP, LPQ, BSIF) by a large margin. Compared with traditional descriptors, the metric learning method MRNML [11] has a better-averaged accuracy, and the spatial-temporal descriptor LBP-TOP also outperforms the averaged frame-level features. Among deep learning models, InsightFace surpasses others by a large margin except for GFGS, where VGG-Face achieves the best performance. The better face verification models boost the kinship verification performance due to the accurate feature representations [187]. Hence, we apply ResNet-50 (voice) and InsightFace (Face) as backbone networks for fusion.

5.6.2 Multi-modalities performance

Experimental results on the TALKIN dataset

Table 12 compares uni-modal based kinship verification with two baseline fusion methods and the proposed deep Siamese network method. Compared to uni-modals, the

Table 10. The average accuracies (%) for K-fold kinship verification with voice under the wild conditions in the TALKIN-Family dataset.

Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD	Average
I-vector [164]	63.11	65.99	60.99	63.07	63.23	61.22	62.17	-	-	-	-	-
GMM-UBM [165, 166]	70.67	64.69	62.63	65.85	66.87	70.15	70.68	-	-	-	-	-
VGG-M [47]	68.83	64.21	59.33	62.23	56.18	56.94	57.99	57.13	59.38	73.01	65.83	61.91
pyannote-S [172, 166]	71.61	61.40	65.08	61.22	55.63	59.21	58.67	64.85	57.50	62.54	63.61	61.94
pyannote-A [172]	72.11	65.70	63.95	63.58	58.44	57.93	59.27	67.13	57.92	71.60	58.61	63.30
ResNet-50 [48]	71.78	75.35	74.10	69.37	67.56	67.52	70.95	66.54	61.04	80.37	71.11	70.52

Table 11. The average accuracies (%) for K-fold kinship verification with faces under the wild conditions in the TALKIN-Family dataset.

Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD	Average
LBP [80, 81]	69.67	63.71	56.52	60.23	59.13	59.67	60.86	55.44	62.29	63.65	58.33	60.86
LPQ [176]	64.72	60.50	61.79	57.38	60.55	60.84	63.58	60.44	55.21	64.20	67.50	61.52
BSIF [149]	69.44	66.97	65.34	59.71	63.01	62.98	61.95	62.13	54.17	69.20	67.50	63.85
LBP-TOP [177]	69.28	62.31	64.63	63.07	60.70	63.71	64.28	67.94	59.17	60.23	64.17	63.59
MNRML [11]	66.89	64.41	64.38	59.45	62.49	63.20	64.78	64.41	58.54	69.53	67.50	64.14
SphereFace [178]	70.11	62.90	59.98	58.75	57.93	59.61	56.91	59.41	60.21	62.53	73.33	61.97
VGG-Face [162]	75.11	63.71	63.08	62.36	59.29	63.67	63.16	68.60	61.67	63.33	71.67	65.06
FaceNet-C [181, 180]	85.11	72.95	68.49	64.84	63.31	69.85	69.19	64.63	63.54	68.09	63.33	68.48
FaceNet-V [181, 183]	86.61	74.82	65.08	68.03	68.15	67.63	68.04	67.35	65.00	66.56	63.33	69.15
InsightFace [168]	87.22	83.31	76.45	78.00	73.21	75.65	76.12	65.66	70.83	75.26	72.50	75.84

Table 12. The comparison of verification accuracy (%) from uni-modal and multimodal techniques on the TALKIN dataset. Reprinted, with permission, from Paper V ©2019 IEEE.

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

feature fusion method and the score fusion method improve the accuracy in terms of average, while both have comparable accuracy as video modalities in father-daughter and mother-son relations. The proposed Siamese network shows a higher level of accuracy compared with the uni-modal and baseline fusion methods. The average accuracy is improved by about 3.8% from the uni-modal method and 1.0% from the baseline fusion method, while the feature fusion method has the best performance in the father-son relation, and both the feature fusion and score fusion methods have the highest accuracy in the mother-daughter relation.

Experimental results on the TALKIN-Family dataset

As presented at the end of Table 13, the proposed UAAML method shows improvement over single modalities for all 11 kin relations and the average level. Figure 14 visualizes the different methods' corresponding receiver operating characteristic (ROC) curves. It can be seen that by fusing the audio and visual features, the performance could be improved, demonstrating that the vocal and facial features complement each other. Besides, the proposed fusion method improves the single-modality verification accuracies and the baseline fusion methods to a certain extent. The average accuracy is improved by about 3.5%, 2.0% from the single modality and baseline fusion methods. Although the baseline fusion methods can not beat the UAAML method in terms of average, the score fusion methods show slightly higher accuracy in relations such as BS, NS, and GMGS. This is a motivation for future work that further explores multi-fusion strategies for audio-visual kinship verification.

5.6.3 Influence factors

The audio-visual kinship verification is affected by many factors. From the perspective of biological attributes, this includes the depth of the genealogical tree, age, and gender.

Table 13. The average accuracies (%) for K-fold kinship verification with faces and voices under the wild conditions in the TALKIN-Family dataset.

Relations	BB	SS	BS	FS	FD	MS	MD	GFGS	GFGD	GMGS	GMGD	Average
ResNet-50 (A) [48]	71.78	75.35	74.10	69.37	67.56	67.52	70.95	66.54	61.04	80.37	71.11	70.52
InsightFace (V) [168]	87.22	83.31	76.45	78.00	73.21	75.65	76.12	65.66	70.83	75.26	72.50	75.84
Late fusion (max)	81.56	81.52	79.53	77.05	71.90	73.35	75.81	68.02	68.33	85.78	71.94	75.89
Siamese fusion [124]	84.89	84.90	77.50	77.13	74.72	73.88	76.49	64.76	76.54	81.05	72.14	76.73
Early fusion	86.33	83.31	75.65	78.91	73.76	77.55	77.31	69.41	71.88	80.21	73.89	77.11
Late fusion (mean)	87.33	84.49	74.84	79.55	73.54	77.97	77.46	67.94	73.13	80.91	78.33	77.77
UAAML (Proposed)	90.02	86.95	78.30	80.74	77.17	76.97	76.65	70.48	78.67	83.18	73.47	79.33

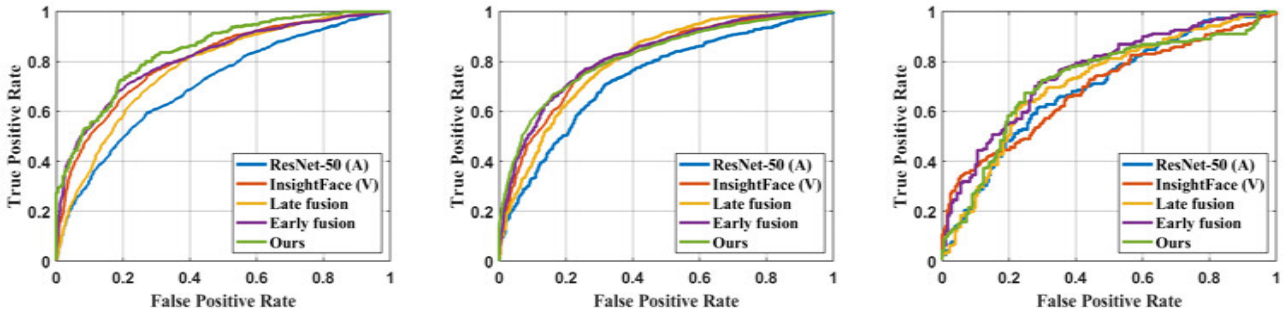


Fig. 14. The ROC curves of different methods on TALKIN-Family with the wild condition obtained on 14(a) siblings, 14(b) parent-child, and 14(c) grandparent-grandchild kin relations.

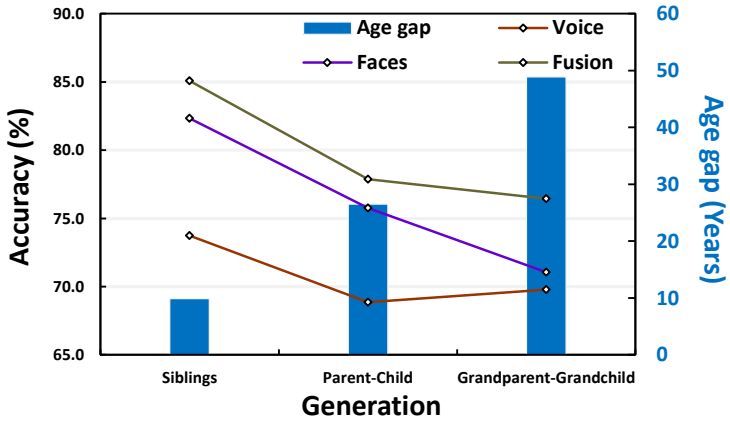


Fig. 15. The verification accuracy of different generations with the voice only, face only, and fusion of both. The performance drops sharply with deeper generation goes for faces. It also drops when fusing both modalities, but the trend slows down at grandparent-grandchild. Instead, the voice performance shows a slight decrease with generation goes and even an improvement from the first generation to the second generation. Overall, the fusion performance outperforms both face and voice modality.

From the data acquisition conditions, factors include the recording background and video speech content. We analyze how those factors influence the performance by providing the corresponding experimental results.

(1) Genealogical tree. Figure 15 shows the averaged verification accuracy for three generations of kinship with different inputs. It can be seen that as the genealogical tree becomes deeper, the performance on faces drops significantly. One reason for this is the age difference between kinship, as distributed in Figure 15. The siblings of the same generation have the smallest age difference of about ten years on average, of which parent-child has about a 26-year age difference. However, the second-generation subjects have an average age margin of about 50 years. As people age, the appearance of their faces varies in structure and texture. These differences affect the inner similarity of kin image pairs, consequently reducing the verification performance [57], whereas acoustic features compensate for the facial aging issues to some extent, especially for the Grandparent-Grandchild relationship.

(2) Gender factor. The experimental setting of relation-specific evaluation provides us the possibility to analyze the influence brought by the gender. From Table 13, we could observe that the gender influence is significant for the siblings, where the opposite gender (BS) has the comparatively lower accuracy than the cases with same gender

(BB, SS). While regarding to the parent-child and grandparent-grandchild relations, the influence of gender is limited, which the influence is mainly caused by the texture difference brought by the age gap.

(3) Recording conditions. The data collection conditions potentially influence system performance, such as the speech text in speaker verification [145], and the same photo issue exists in kinship verification [146] by providing latent clues. To control one variable factor at one time, we generate kinship pairs that 1) speak the fixed text but with different backgrounds (text-dependent) and 2) are recorded under the white background but with different speaking content (white background). Figure 16 shows the experimental results in text-dependent and white background conditions with different inputs. The background influence could be seen clearly from Figure 16 (a) that the visual-based performance has higher accuracy. Two reasons explain the phenomenon: 1) the noise effect is eased under the white background and 2) the white background videos within one family are possibly recorded in the same place, where the illumination provides the data bias [145]. This also explains why we asked the participants to take videos with two backgrounds, one of which is white, to easily distinguish the same or different backgrounds during training and test data selection. As illustrated in Figure 16 (b), the fixed text setting achieves comparable performance to the free-speaking setting due to the equal similarity within kin and non-kin pairs. Overall, the audio-visual fusion improves performance under all conditions, while under two semicontrolled environments, the improvement of fusion is comparably limited.

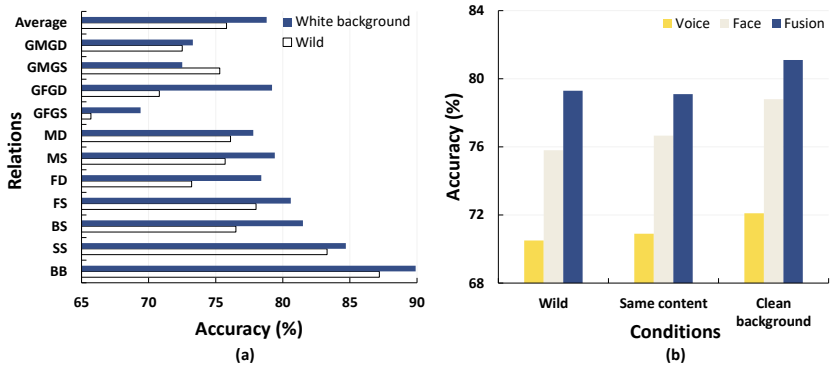


Fig. 16. The performance of kinship verification on TALKIN-Family under different conditions.

5.6.4 Human performance

We test the human performance on kinship verification by using a subset of TALKIN-Family. Twelve volunteers from China participated in the experiments. Before the test, they had never seen or known any information about the dataset subjects and were asked to answer whether the given clips have a kin relation. In general, we set up three kinds of tasks, namely, kinship verification from *(1) facial videos without voice*, *(2) voice*, and *(3) facial videos with voice*. For each task, we select two kin pairs and two non-kin pairs from 11 kin relations, resulting in 22 positive pairs (kinship) and 22 negative pairs (non-kinship) in total. Note that there is no subject overlap between positive and negative pairs nor among the three sub-tasks to avoid miscellaneous information between tasks and comparison. Figure 17 illustrates the human performance results, in which subfigure (a) shows the overall accuracy and distribution of subject performance, and we compare the true positive and true negative accuracy in subfigure (b). Generally, humans tend to have a better ability to verify kinship from voice than from the face, while when given synchronous facial videos and voice, humans can make a better judgment. Figure 17 (a) indicates that face and voice information enables human observers to make a more stable assessment. Figure 17 (b) shows that humans have higher accuracy in verifying the negative samples, and multimodal information helps humans to recognize non-kinship, thus improving overall accuracy. It is worth noting that it takes about an hour for a person to complete the entire test, while machine learning methods spend much less time in the inference process. We conclude that machine learning methods can outperform human ability both efficiently and effectively.

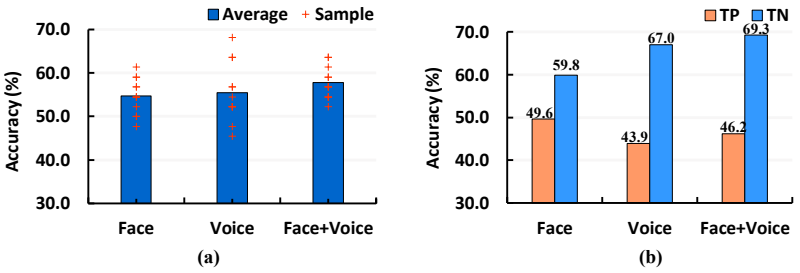


Fig. 17. The human performance on a subset of TALKIN-Family from the face, voice, and face&voice, respectively. (a) shows the overall verification performance with different modalities. (b) presents the true positive (TP) and true negative (TN) distributions of human performance under different settings.

5.7 Conclusion

Audio-visual kinship verification is a new and potential research topic. In this chapter, we systematically investigate the problem of audio-visual kinship verification. Moreover, the baseline experiments of single modal kinship verification are performed, in which vocal kinship verification is evaluated for the first time. Based on the single modal methods, a deep Siamese network for multimodal fusion is also proposed for metric learning of kinship verification. Experiments indicate that the proposed Siamese network improves accuracy over baseline uni-modal and multimodal fusion techniques for kinship verification. Additionally, audio (vocal) information is shown to be complementary and useful for the kinship verification problem. To improve the fusion performance, we further propose a deep learning framework named UAAML to jointly learn the modal invariant and adaptive fused features for kinship verification with contrastive loss. Our proposed fusion method outperforms both the baseline methods and human ability in kinship verification. The human performance experiment shows that by providing the face and voice of the subject, people could have higher kinship verification accuracy than using faces and voice only. We expect this work sets a milestone for audio-visual kinship verification. To stimulate future study, in the next chapter, we investigate the limitations of our datasets and the proposed approach and discuss future directions.

6 Discussion and summary

This thesis gave an overview of the audio-visual kinship verification problem and presented novel work in (1) establishing two comprehensive audiovisual kinship datasets, (2) studying the visual features for the FKV problem by introducing the color texture features and ELM classifier to solve the problem, and (3) extending the current FKV research from monomodal visual-based methods to acoustic methods, and therefore proposing the multimodal fusion method for audio-visual kinship verification. In this chapter, the findings and contributions of this thesis are summarized. In addition, we discuss the limitations of this thesis and present future research directions.

6.1 Contributions

In this thesis, we addressed the kinship verification from faces and voices. The objective is to capture the similarity of the kinship from biometric traits in an unobtrusive manner, *e.g.,* from faces or speaking voices as captured by audio-visual recording devices, and to verify whether or not two people have a kin relation from this audio-visual data. The main contribution of this thesis can be summarized in three aspects, *i.e.,* (1) developing novel audio-visual kinship datasets, (2) studying the effective visual features for the facial kinship verification problem, and (3) proposing methods for the new kinship recognition task of *audio-visual kinship verification*.

First, we considered the problem of kinship verification from both faces and voices, and raised a new subproblem such as *audio-visual kinship verification*. However, to our best knowledge, there is no research on kinship verification from audio-visual features or related datasets. Therefore, a new kinship dataset composed of facial videos and speaking voices was established. In particular, we selected 100 pairs of facial videos with speaking voices from YouTube for each of the four parent-child kin relations. However, the TALKIN dataset has limitations on size, kin relation, data variety, and application scenarios. Additionally, we extend the TALKIN dataset to a comprehensive dataset named TALKIN-Family, which consists of facial videos and synchronous speaking audio with properties that differ from the TALKIN dataset. In TALKIN-Family, there are 246 unique family trees and 1012 individuals with rich annotations of family relationships, age, gender, and scene conditions. The size of the family tree ranges from 2 to 14 subjects between 5 and 81 years of age. Each subject has multiple talking facial videos about 10 seconds in length under different conditions. Overall, there are 9.2 hours of videos in TALKIN-Family.

In our second contribution, we study kinship verification from visual features. Biologically, facial chromaticity is related to genetically expressed features, such as eye color or skin tone. Indeed, we studied three color spaces in our experiments: RGB, HSV and YCbCr. We found that the features extracted from these three color spaces can beat the features extracted from a gray image for different kin relations. Moreover, due to the orthogonality of the HSV color space, features extracted from it show a higher discrimination ability. In line with this finding, we proposed an ELM-based network for recognizing kinship and non-kinship. We compared the image similarity from the corresponding color spaces, and therefore a distance vector could be obtained as the input to the ELM for classification. In this way, the method could capture both the texture and color similarity of kinship.

Based on the audiovisual datasets and learned visual features, we investigate the problem of audio-visual kinship verification. Before that, we first provide the benchmarks on the single modality (*i.e.,* face and voice). The facial kinship verification problem has been widely explored, and based on that we establish architectures with various methods from traditional to novel deep learning approaches. Kin face research has attracted the attention of researchers. However, to the best of our knowledge, kinship verification from voices remains largely unstudied. Similar to the visual features, we extract various vocal features for kinship verification. For the first time, we proved that voices could be effective in verifying kinship. Based on our single modal benchmarks, we carried out the audio-visual kinship verification study. In particular, we proposed a deep Siamese fusion method that learns both kinship similarity and non-kinship dissimilarity and multimodal features at the same time. To compare the effectiveness of our proposed fusion method, we also conducted benchmark experiments on fusion methods such as early fusion and late fusion. Experimental results show that audio-visual fusion could outperform mono-modal methods. Kinship verification could benefit from a combination of discriminant information extracted from both video and audio signals. Furthermore, motivated by adversarial learning, we proposed a multimodal fusion network, UAAML, which can jointly learn modal invariant and attentive features with the unified multimodal features for kinship verification. Finally, extensive experiments show the effectiveness of the proposed method compared with baseline fusion methods. Evaluations of human performance also indicate that audiovisual information helps to improve judgment.

We believe the research is valuable and could provide insight for kinship recognition studies from faces and audio-visual data. We hope that the work will attract researchers from different disciplines to the field and promote the development of kinship analysis.

6.2 Limitations and future work

This thesis thoroughly summarizes the main research works in my Ph.D. studies on facial kinship verification. First, the audio-visual datasets are established, which until now did not exist. Second, kinship verification from visual features based on texture features is explored. Third, the problem of kinship verification from voices and the fusion of faces and voices is systematically explored.

After finalizing the work of this thesis, there are still several topics that could be further explored. First, the proposed TALKIN and TALKIN-Family datasets suffer from limited data. They could be improved in size and diversity (*e.g.,* kin relations). Second, although our findings in Articles III and IV indicate that the color features could help to improve the performance of FKV, these features are difficult to track when facing illumination changes when images taken are under different conditions. Finally, the performance of audio-visual fusion for kinship verification could be further improved in verification accuracy and computational efficiency.

Recently, many studies about FKV have been proposed that, focus on exploring discriminative kinship features. However, kinship verification study is still challenging due to the uncertainty and stochasticity in genetic heredity and environmental influences. The technology is still in its early stages and cannot satisfactorily address many of the challenges listed in Chapter 2.2 and many advanced applications. Still, the insufficient data sets pose challenges to efficiently and effectively study facial kinship verification, and deep learning has not yet reached its full potential as face verification.

In the following, we discuss the future research directions of FKV, which we hope will provide guidance and insight to interested researchers.

Large-scale dataset establishment for FKV. The availability of benchmark datasets has played a key role in advancing visual kinship recognition research. Clearly, there is a pressing need to build a large and well-annotated in-the-wild dataset that reflects the true data distribution of facial kinship worldwide and meets the requirements of data-hungry deep learning methods. However, despite the recent progress (*e.g.,* the FIW dataset [20]), such a goal remains unrealized. As we discussed in Chapter 2.2, there are still many problems in the current kinship datasets, such as the following. First, there is the issue of an unbalanced ethnicity distribution, which may lead to algorithmic biases such as demographic bias [188]. Secondly, the current datasets are not large and diverse enough to reflect real-world conditions. Last but not least, most of the current datasets focus on direct descendants without giving full consideration to the height or breadth of the family tree, which is also an important factor for FKV research. Due to privacy,

security, and labeling concerns, building a large, diverse, and comprehensive dataset for FKV is much more challenging than for face verification.

Bias and Fairness. Recently, the AI research community has realized the importance of developing fair and unbiased AI systems [189, 190]. For instance, facial recognition, Is facial recognition too biased to be let loose [188], has been shown to have serious demographic bias [191, 192]. Kinship recognition relies on people-centric data and also faces such issues. This is especially concerning since kinship recognition systems are intended to be used in critical security applications such as crime scene investigation, border control, and searching for missing children. As we discussed above, datasets play a critical role in FKV. If the training datasets reflect unwanted demographic bias and imbalance, the learned model is unlikely to perform well in the wild. Therefore, taking both algorithm and data biases into consideration in kinship understanding is an important future research direction.

Accurate features suitable for FKV. Accurate feature representations, suitable for the verification of kinship, are critical to obtaining good FKV performance. However, it still remains a challenging open problem. As we discussed in detail in Chapter 2.2, in contrast to face verification, kin faces are not identical. FKV has large interpersonal variations. The facial similarities of kin faces are often not obvious and vary considerably between different families. All these factors pose great challenges for accurate feature representation. Furthermore, there is another important question: *what features are suitable for kinship verification*? Maybe it is helpful for researchers to be aware of relevant findings in psychology, neuroscience, and anthropology. Finally, facial attributes like gender, age, and skin color may be helpful for FKV. This is the key to learning how to fuse multiple features effectively. Besides, fusing complementary features is also promising in boosting performance.

Transformer models for kinship verification. The transformer architectures have been recently applied to a variety of vision problems, and have been shown to be a potential alternative to traditional CNNs [193]. As the key component of the transformer, the self-attention mechanism has the ability to learn the long-range dependencies in the network and extract the intrinsic features [194]. Nevertheless, the potential of the transformer for the FKV problem has barely been explored. Therefore, it is valuable to explore the transformer potential of FKV for image/video representation learning, and multimodal tasks [195].

Multimodality. Most current research focuses on kinship verification from facial images, while only a few works consider facial videos. It has been shown that visual kinship verification performance can be enhanced by incorporating multimodal signals

such as expressions [22], voice [124], gait [135], infrared images [196], hyperspectral images [197], 3D facial images [198], and facial sketches [199].

Interdisciplinary research. FKV is an important yet challenging problem, with many open issues. It has been studied in several fields, including psychology, anthropology, neuroscience, computer vision, and machine learning. Towards ultimately solving the problem of FKV, we argue that interdisciplinary research should be advocated. For instance, in genetics, automatic computational kinship verification can be applied in exploring facial traits' computation [200] and genetic problems such as evolutionary patterns of DNA methylation sites [201].

References

[1] L. T. Maloney and M. F. Dal Martello, "Kin recognition and the perceived facial similarity of children," *Journal of Vision*, vol. 6, no. 10, pp. 4–4, 2006.

[2] R. H. Porter, "Mutual mother-infant recognition in humans," *Kin recognition*, pp. 413–432, 1991.

[3] M. F. Dal Martello and L. T. Maloney, "Lateralization of kin recognition signals in the human face," *Journal of Vision*, vol. 10, no. 8, pp. 9–9, 2010.

[4] L. M. DeBruine, F. G. Smith, B. C. Jones, S. C. Roberts, M. Petrie, and T. D. Spector, "Kin recognition signals in adult faces," *Vision research*, vol. 49, no. 1, pp. 38–43, 2009.

[5] M. F. Dal Martello and L. T. Maloney, "Where are kin recognition signals in the human face?" *Journal of Vision*, vol. 6, no. 12, pp. 2–2, 2006.

[6] A. Alvergne, F. Perreau, A. Mazur, U. Mueller, and M. Raymond, "Identification of visual paternity cues in humans," *Biology letters*, vol. 10, no. 4, p. 20140063, 2014.

[7] F. Hansen, L. M. DeBruine, I. J. Holzleitner, A. J. Lee, K. J. O'Shea, and V. Fasolt, "Kin recognition and perceived facial similarity," *Journal of Vision*, vol. 20, no. 6, pp. 18–18, 2020.

[8] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *2010 IEEE International conference on image processing*. IEEE, 2010, pp. 1577–1580.

[9] A. M'charek, "Tentacular faces: Race and the return of the phenotype in forensic identification," *American Anthropologist*, vol. 122, no. 2, pp. 369–380, 2020.

[10] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Supervised mixed norm autoencoder for kinship verification in unconstrained videos," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1329–1341, 2019.

[11] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 331–345, 2014.

[12] J. Lu, J. Hu, X. Zhou, J. Zhou, M. Castrillón-Santana, J. Lorenzo-Navarro, L. Kou, Y. Shang, A. Bottino, and T. F. Vieira, "Kinship verification in the wild: The first kinship verification competition," in *IEEE International Joint Conference on Biometrics*. IEEE, 2014, pp. 1–6.

[13] W. Jang, A. Chhabra, and A. Prasad, "Enabling multi-user controls in smart home devices," in *Proceedings of the 2017 Workshop on Internet of Things Security and Privacy*, 2017, pp. 49–54.

[14] E. Dahan and Y. Keller, "A unified approach to kinship verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[15] W. Li, S. Wang, J. Lu, J. Feng, and J. Zhou, "Meta-mining discriminative samples for kinship verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021.

[16] S. Huang, J. Lin, L. Huangfu, Y. Xing, J. Hu, and D. D. Zeng, "Adaptively weighted k-tuple metric network for kinship verification," *IEEE Transactions on Cybernetics*, 2022.

[17] X. Qin, D. Liu, and D. Wang, "A literature survey on kinship verification through facial images," *Neurocomputing*, vol. 377, pp. 213–224, 2020.

[18] M. Bordallo Lopez, A. Hadid, E. Boutellaa, J. Goncalves, V. Kostakos, and S. Hosio, "Kinship verification from facial images and videos: human versus machine," *Machine Vision and Applications*, vol. 29, no. 5, pp. 873–890, 2018.

[19] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: Understanding the core of a family," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1855–1867, 2015.
[20] 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*, vol. 40, no. 11, pp. 2624–2637, 2018.
[21] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *European Conference on Computer Vision*. Springer, 2012, pp. 525–538.
[22] H. Dibeklioglu, A. Ali Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1497–1504.
[23] H. Yan and J. Hu, "Video-based kinship verification using distance metric learning," *Pattern Recognition*, vol. 75, pp. 15–24, 2018.
[24] R. T. Sataloff, "Genetics of the voice," *Journal of Voice*, vol. 9, no. 1, pp. 16–19, 1995.
[25] W. G. Van, J. Vercammen, and F. Debruyne, "Voice similarity in identical twins." *Acta oto-rhino-laryngologica Belgica*, vol. 55, no. 1, pp. 49–55, 2001.
[26] S. P. Whiteside and E. Rixon, "Speech tempo and fundamental frequency patterns: a case study of male monozygotic twins and an age-and sex-matched sibling," *Logopedics Phoniatrics Vocology*, vol. 38, no. 4, pp. 173–181, 2013.
[27] M. Weirich and L. Lancia, "Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers." in *ICPhS*, 2011, pp. 2118–2121.
[28] F. Debruyne, W. Decoster, A. Van Gijsel, and J. Vercammen, "Speaking fundamental frequency in monozygotic and dizygotic twins," *Journal of Voice*, vol. 16, no. 4, pp. 466–471, 2002.
[29] F. Nolan, K. McDougall, and T. Hudson, "Some acoustic correlates of perceived (dis) similarity between same-accent voices." in *ICPhS*, 2011, pp. 1506–1509.
[30] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.
[31] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
[32] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.
[33] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.
[34] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[36] N. Kohli, R. Singh, and M. Vatsa, "Self-similarity representation of weber faces for kinship classification," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2012, pp. 245–250.
[37] G. Guo and X. Wang, "Kinship measurement on salient facial features," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2322–2325, 2012.
[38] A. Puthenputhussery, Q. Liu, and C. Liu, "Sift flow based genetic fisher vector feature for kinship verification," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 2921–2925.

[39] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid, "A weighted exponential discriminant analysis through side-information for face and kinship verification using statistical binarized image features," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 1, pp. 171–185, 2021.

[40] X. Zhou, J. Hu, J. Lu, Y. Shang, and Y. Guan, "Kinship verification from facial images under uncontrolled conditions," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 953–956.

[41] X. Zhou, J. Lu, J. Hu, and Y. Shang, "Gabor-based gradient orientation pyramid for kinship verification under uncontrolled environments," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 725–728.

[42] B. Patel, R. Maheshwari, and B. Raman, "Evaluation of periocular features for kinship verification in the wild," *Computer Vision and Image Understanding*, vol. 160, pp. 24–35, 2017.

[43] A. Moujahid and F. Dornaika, "A pyramid multi-level face descriptor: application to kinship verification," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 9335–9354, 2019.

[44] H. Yan, "Learning discriminative compact binary face descriptor for kinship verification," *Pattern Recognition Letters*, vol. 117, pp. 146–152, 2019.

[45] A. Goyal and T. Meenpal, "Patch-based dual-tree complex wavelet transform for kinship recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 191–206, 2020.

[46] 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*, vol. 7, no. 5, pp. 525–532, 1999.

[47] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[48] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[50] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[51] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[52] X. Wu, X. Feng, L. Li, E. Boutellaa, and A. Hadid, "Kinship verification based on deep learning," in *Deep Learning in Object Detection and Recognition*. Springer, 2019, pp. 113–132.

[53] 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 on Image Analysis and Recognition*. Springer, 2016, pp. 539–548.

[54] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[55] D. B. Krupp, L. M. Debruine, and P. Barclay, "A cue of kinship promotes cooperation for the public good," *Evolution and Human Behavior*, vol. 29, no. 1, pp. 49–55, 2008.

[56] N. Kohli, *Automatic kinship verification in unconstrained faces using deep learning*. West Virginia University, 2019.

[57] M. Georgopoulos, Y. Panagakis, and M. Pantic, "Modeling of facial aging and kinship: A survey," *Image and Vision Computing*, vol. 80, pp. 58–79, 2018.

[58] C. Xu, Q. Liu, and M. Ye, "Age invariant face recognition and retrieval by coupled auto-encoder networks," *Neurocomputing*, vol. 222, pp. 62–71, 2017.

[59] S. Wang, Z. Ding, and Y. Fu, "Cross-generation kinship verification with sparse discriminative metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2783–2790, 2018.

[60] S. Xia, M. Shao, and Y. Fu, "Kinship verification through transfer learning," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[61] S. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. Phillips, A. Sachs, and E. Schadt, "Genetic inheritance of gene expression in human cell lines," *The American Journal of Human Genetics*, vol. 75, no. 6, pp. 1094–1105, 2004.

[62] A. Alvergne, C. Faurie, and M. Raymond, "Differential facial resemblance of young children to their parents: who do children look like more?" *Evolution and Human behavior*, vol. 28, no. 2, pp. 135–144, 2007.

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

[64] I. Ö. Ertugrul and H. Dibeklioglu, "What will your future child look like? modeling and synthesis of hereditary patterns of facial dynamics," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 33–40.

[65] P. Gao, S. Xia, J. Robinson, J. Zhang, C. Xia, M. Shao, and Y. Fu, "What will your child look like? dna-net: Age and gender aware kin face synthesizer," *arXiv preprint arXiv:1911.07014*, 2019.

[66] C. Xia, S. Xia, Y. Zhou, L. Zhang, and M. Shao, "Graph based family relationship recognition from a single image," in *Pacific Rim International Conference on Artificial Intelligence*, 2018, pp. 310–320.

[67] W. Wang, S. You, S. Karaoglu, and T. Gevers, "Kinship identification through joint learning using kinship verification ensembles." in *European Conference on Computer Vision*, 2020, pp. 613–628.

[68] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[69] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[70] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[71] H. Yan, J. Lu, and X. Zhou, "Prototype-based discriminative feature learning for kinship verification," *IEEE Transactions on cybernetics*, vol. 45, no. 11, pp. 2535–2545, 2014.

[72] H. Yan, J. Lu, W. Deng, and X. Zhou, "Discriminative multimetric learning for kinship verification," *IEEE Transactions on Information forensics and security*, vol. 9, no. 7, pp. 1169–1178, 2014.

[73] S. Xia, M. Shao, and Y. Fu, "Toward kinship verification using visual attributes," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 549–552.

[74] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1046–1056, 2012.

[75] X. Wang and C. Kambhamettu, "Leveraging appearance and geometry for kinship verification," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5017–5021.

[76] A. Goyal and T. Meenpal, "Detection of facial parts in kinship verification based on edge information," in *2018 Conference on Information and Communication Technology (CICT)*. IEEE, 2018, pp. 1–6.

[77] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 815–830, 2009.

[78] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3444–3451.

[79] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[80] 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, 2006.

[81] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[82] P. Alirezazadeh, A. Fathi, and F. Abdali-Mohammadi, "A genetic algorithm-based feature selection for kinship verification," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2459–2463, 2015.

[83] A. Bottinok, I. U. Islam, and T. F. Vieira, "A multi-perspective holistic approach to kinship verification in the wild," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 2. IEEE, 2015, pp. 1–6.

[84] L. Cui and B. Ma, "Adaptive feature selection for kinship verification," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 751–756.

[85] X. Chen, L. An, S. Yang, and W. Wu, "Kinship verification in multi-linear coherent spaces," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4105–4122, 2017.

[86] E. Xing, M. Jordan, S. J. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, pp. 521–528, 2002.

[87] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and trends in machine learning*, vol. 5, no. 4, pp. 287–364, 2012.

[88] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Asian conference on computer vision*. Springer, 2014, pp. 252–267.

[89] J. Hu, J. Lu, Y.-P. Tan, J. Yuan, and J. Zhou, "Local large-margin multi-metric learning for face and kinship verification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1875–1891, 2017.

[90] L. Kou, X. Zhou, M. Xu, and Y. Shang, "Learning a genetic measure for kinship verification using facial images," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[91] Y. Zhang, B. Ma, L. Huang, and H. Hu, "Transfer metric learning for kinship verification with locality-constrained sparse features," in *International Conference on Neural Information Processing*. Springer, 2015, pp. 234–243.

[92] Z. Wei, M. Xu, L. Geng, H. Liu, and H. Yin, "Adversarial similarity metric learning for kinship verification," *IEEE Access*, vol. 7, pp. 100 029–100 035, 2019.

[93] X. Zhou, Y. Shang, H. Yan, and G. Guo, "Ensemble similarity learning for kinship verification from facial images in the wild," *Information Fusion*, vol. 32, pp. 40–48, 2016.

[94] M. Xu and Y. Shang, "Kinship measurement on face images by structured similarity fusion," *IEEE Access*, vol. 4, pp. 10 280–10 287, 2016.

[95] X. Zhou, H. Yan, and Y. Shang, "Kinship verification from facial images by scalable similarity fusion," *Neurocomputing*, vol. 197, pp. 136–142, 2016.

[96] M. Xu and Y. Shang, "Kinship verification using facial images by robust similarity learning," *Mathematical Problems in Engineering*, vol. 2016, 2016.

[97] X. Qin, X. Tan, and S. Chen, "Mixed bi-subject kinship verification via multi-view multi-task learning," *Neurocomputing*, vol. 214, pp. 350–357, 2016.

[98] Y. Fang, Y. Y. S. Chen, H. Wang, and C. Shu, "Sparse similarity metric learning for kinship verification," in *2016 Visual Communications and Image Processing (VCIP)*. IEEE, 2016, pp. 1–4.

[99] Y. Guo, H. Dibeklioglu, and L. Van der Maaten, "Graph-based kinship recognition," in *2014 22nd international conference on pattern recognition*. IEEE, 2014, pp. 4287–4292.

[100] J. Liang, Q. Hu, C. Dang, and W. Zuo, "Weighted graph embedding-based metric learning for kinship verification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1149–1162, 2018.

[101] H. Yan, X. Zhou, and Y. Ge, "Neighborhood repulsed correlation metric learning for kinship verification," in *2015 Visual Communications and Image Processing (VCIP)*. IEEE, 2015, pp. 1–4.

[102] H. Yan, "Kinship verification using neighborhood repulsed correlation metric learning," *Image and Vision Computing*, vol. 60, pp. 91–97, 2017.

[103] X. Lei, B. Li, and J. Xie, "Locality discriminative canonical correlation analysis for kinship verification," in *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2017, pp. 1870–1874.

[104] J. Zhang, S. Xia, H. Pan, and A. K. Qin, "A genetics-motivated unsupervised model for tri-subject kinship verification," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2916–2920.

[105] H. Liu, J. Cheng, and F. Wang, "Kinship verification based on status-aware projection learning," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1072–1076.

[106] H. Liu and C. Zhu, "Status-aware projection metric learning for kinship verification," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 319–324.

[107] Y. Wu, Z. Ding, H. Liu, J. Robinson, and Y. Fu, "Kinship classification through latent adaptive subspace," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 143–149.

[108] Y.-G. Zhao, Z. Song, F. Zheng, and L. Shao, "Learning a multiple kernel similarity metric for kinship verification," *Information Sciences*, vol. 430, pp. 247–260, 2018.

[109] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *CVPR 2011*. IEEE, 2011, pp. 785–792.

[110] X. Gao, S. C. H. Hoi, Y. Zhang, J. Wan, and J. Li, "Soml: sparse online metric learning with application to image retrieval," in *AAAI'14 Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1206–1212.

[111] M. Shao, S. Xia, and Y. Fu, "Genealogical face recognition based on ub kinface database," in *CVPR 2011 WORKSHOPS*. IEEE, 2011, pp. 60–65.

[112] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[113] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2020.

[114] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 8590–8605, 2020.

[115] M. Wang, Zechao Li, Xiangbo Shu, Jingdong, and J. Tang, "Deep kinship verification," in *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, 2015, pp. 1–6.

[116] H. Yan and S. Wang, "Learning part-aware attention networks for kinship verification," *Pattern Recognition Letters*, vol. 128, pp. 169–175, 2019.

[117] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang, "Kinship verification with deep convolutional neural networks," in *British Machine Vision Conference 2015*, 2015.

[118] Q. Duan and L. Zhang, "Advnet: Adversarial contrastive residual net for 1 million kinship recognition," in *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, 2017, pp. 21–29.

[119] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.

[120] Y. Li, J. Zeng, J. Zhang, A. Dai, M. Kan, S. Shan, and X. Chen, "Kinnet: Fine-to-coarse deep metric learning for kinship verification," in *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, 2017, pp. 13–20.

[121] X. Zhou, K. Jin, M. Xu, and G. Guo, "Learning deep compact similarity metric for kinship verification from face images," *Information Fusion*, vol. 48, pp. 84–94, 2019.

[122] L. Zhang, Q. Duan, D. Zhang, W. Jia, and X. Wang, "Advkin: Adversarial convolutional network for kinship verification," *IEEE transactions on cybernetics*, vol. 51, no. 12, pp. 5883–5896, 2020.

[123] H. Dibeklioglu, "Visual transformation aided contrastive learning for video-based kinship verification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2459–2468.

[124] X. Wu, E. Granger, T. H. Kinnunen, X. Feng, and A. Hadid, "Audio-visual kinship verification in the wild," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[125] Y. Suh, B. Han, W. Kim, and K. M. Lee, "Stochastic class-based hard example mining for deep metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7251–7259.

[126] S. Wang and H. Yan, "Discriminative sampling via deep reinforcement learning for kinship verification," *Pattern Recognition Letters*, vol. 138, pp. 38–43, 2020.

[127] S. Wang, Z. Ding, and Y. Fu, "Coupled marginalized auto-encoders for cross-domain multi-view learning." in *IJCAI*, 2016, pp. 2125–2131.

[128] J. Liang, J. Guo, S. Lao, and J. Li, "Using deep relational features to verify kinship," in *CCF Chinese Conference on Computer Vision*. Springer, 2017, pp. 563–573.

[129] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah, "Who do i look like? determining parent-offspring resemblance via gated autoencoders," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1757–1764.

[130] S. Wang, J. P. Robinson, and Y. Fu, "Kinship verification on families in the wild with marginalized denoising metric learning," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 216–221.

[131] S. Ozkan and A. Ozkan, "Kinshipgan: Synthesizing of kinship faces from family photos by regularizing a deep face network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2142–2146.

[132] S. Zhang, D. Chen, J. Yang, and B. Schiele, "Guided attention in cnns for occluded pedestrian detection and re-identification," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1875–1892, 2021.

[133] W. Shen, W. Bao, G. Zhai, L. Chen, X. Min, and Z. Gao, "Blurry video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5114–5123.

[134] M. Jin, Z. Hu, and P. Favaro, "Learning to extract flawless slow motion from blurry videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8112–8121.

[135] S. E. Bekhouche, A. Chergui, A. Hadid, and Y. Ruichek, "Kinship verification from gait," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2950–2954.

[136] E. Boutellaa, M. Bordallo López, S. Ait-Aoudia, X. Feng, and A. Hadid, "Kinship verification from videos using spatio-temporal texture features and deep learning," *arXiv preprint arXiv:1708.04069*, 2017.

[137] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Supervised mixed norm autoencoder for kinship verification in unconstrained videos," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1329–1341, 2018.

[138] http://iab-rubric.org/resources/KIVI.html.

[139] X. Wu, X. Feng, X. Cao, X. Xu, D. Hu, M. Bordallo López, and L. Liu, "Facial kinship verification: A comprehensive review and outlook," *International Journal of Computer Vision*, pp. 1–32, 2022.

[140] Y. Sun, J. Li, Y. Wei, and H. Yan, "Video-based parent-child relationship prediction," in *2018 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2018, pp. 1–4.

[141] R. T. Sataloff, "Genetics of the voice," *Journal of Voice*, vol. 9, no. 1, pp. 16–19, 1993.

[142] N. Kohli, M. Vatsa, R. Singh, A. Noore, and A. Majumdar, "Hierarchical representation learning for kinship verification," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 289–302, 2016.

[143] X. Wu, X. Feng, X. Zhang, M. Bordallo López, and L. Liu, "Audio-visual kinship verification: a new dataset and a unified adaptive adversarial multimodal learning approach," *submitted to IEEE Transactions on Cybernetics, (Major revision)*, 2022, https://doi.org/10.36227/techrxiv.19776007.v1.

[144] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[145] M. Bordallo López, E. Boutellaa, and A. Hadid, "Comments on the "kinship face in the wild" data sets," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2342–2344, 2016.

[146] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[147] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[148] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 356–370, 2012.

[149] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE, 2012, pp. 1363–1366.

[150] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[151] Q. He, X. Jin, C. Du, F. Zhuang, and Z. Shi, "Clustering in extreme learning machine feature space," *Neurocomputing*, vol. 128, pp. 88–95, 2014.

[152] X. Wu, E. Boutellaa, X. Feng, and A. Hadid, "Kinship verification from faces: Methods, databases and challenges," in *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, 2016, pp. 1–6.

[153] S. M. Taylor, *Acoustic Correlates of Aging and Familial Relationship*. Brigham Young University, 2018.

[154] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[155] 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*, 2008.

[156] A. Rosenberg, "Listener performance in speaker verification tasks," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 221–225, Jun. 1973.

[157] 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*, vol. 48, no. 4, pp. 182–186, Dec. 2008.

[158] H. J. Künzel, "Automatic speaker recognition of identical twins." *International Journal of Speech, Language & the Law*, vol. 17, no. 2, 2010.

[159] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.

[160] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 164–172.

[161] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 619–14 628.

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

[163] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[164] 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*, vol. 19, no. 4, pp. 788–798, 2010.

[165] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[166] H. Bredin, "TristouNet: Triplet Loss for Speaker Turn Embedding," in *42nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*, 2017, uRL = http://arxiv.org/abs/1609.04301.

[167] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[168] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[169] J. Guo and J. Deng, "Insightface," 2021, url=https://github.com/deepinsight/insightface.

[170] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2," 2017, url=https://github.com/a-nagrani/VGGVox.

[171] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio," 2017, url=https://github.com/clcarwin/sphereface_pytorch.

[172] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[173] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[174] X. Wu, E. Boutellaa, M. Bordallo López, X. Feng, and A. Hadid, "On the usefulness of color for kinship verification from face images," in *2016 IEEE International workshop on information forensics and security (WIFS)*. IEEE, 2016, pp. 1–6.

[175] X. Wu, X. Feng, E. Boutellaa, and A. Hadid, "Kinship verification using color features and extreme learning machine," in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2018, pp. 187–191.

[176] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[177] 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*, vol. 29, no. 6, pp. 915–928, 2007.

[178] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[179] carwin, "Sphereface," 2020, url=https://github.com/pyannote/pyannote-audio-hub.

[180] D. Yi, Z. Lei, S. Liao, and S. Li, "Casiawebface: Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[181] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[182] T. Esler, "Facenet," 2021, url=https://github.com/timesler/facenet-pytorch.

[183] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[184] E. Nizhibitsky, "pytorch-insightface," 2019, url=https://github.com/nizhib/pytorch-insightface.

[185] S. Albanie, "pytorch-mcn," 2018, url=https://github.com/albanie/pytorch-mcn.

[186] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowledge-Based Systems*, vol. 180, pp. 38–50, 2019.

[187] A. Shadrikov, "Achieving better kinship recognition through better baseline," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 872–876.
[188] D. Castelvecchi, "Is facial recognition too biased to be let loose?" *Nature*, vol. 587, no. 7834, pp. 347–349, 2020.
[189] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.
[190] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
[191] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295.
[192] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.
[193] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2022.
[194] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[195] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[196] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon, "Refining geometry from depth sensors using ir shading images," *International Journal of Computer Vision*, vol. 122, no. 1, pp. 1–16, 2017.
[197] S. Arya, N. Pratap, and K. Bhatia, "Future of face recognition: a review," *Procedia Computer Science*, vol. 58, pp. 578–585, 2015.
[198] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.
[199] S. Nagpal, M. Vatsa, and R. Singh, "Sketch recognition: What lies ahead?" *Image and Vision Computing*, vol. 55, pp. 9–13, 2016.
[200] S. Richmond, L. J. Howe, S. Lewis, E. Stergiakouli, and A. Zhurov, "Facial genetics: a brief overview," *Frontiers in genetics*, vol. 9, p. 462, 2018.
[201] D. Gokhman, M. Nissim-Rafinia, L. Agranat-Tamir, G. Housman, R. García-Pérez, E. Lizano, O. Cheronet, S. Mallick, M. A. Nieves-Colón, H. Li *et al.*, "Differential dna methylation of vocal and facial anatomy genes in modern humans," *Nature communications*, vol. 11, no. 1, pp. 1–21, 2020.

Original publications

I Wu X, Feng X, Cao X, Xu X, Hu D, Bordallo López M & Liu L (2022) Facial kinship verification: A comprehensive review and outlook. International Journal of Computer Vision (IJCV), vol. 130, no. 6, pp. 1494–1525.

II Wu X, Boutellaa E, Feng X & Hadid A (2016) Kinship verification from faces: Methods, databases and challenges. Proc. IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2016), Hong Kong, China, pp. 1–6.

III Wu X, Boutellaa E, Bordallo López M, Feng X & Hadid A (2016) On the usefulness of color for kinship verification from face images. IEEE International Workshop on Information Forensics and Security (WIFS 2016), Abu Dhabi, UAE, pp. 1–6.

IV Wu X, Feng X, Boutellaa E & Hadid A (2018) Kinship verification using color features and extreme learning machine. Proc. 2018 IEEE International Conference on Signal and Image Processing (ICSIP 2018), Shenzhen, China, pp. 187–191.

V Wu X, Granger E, Kinnunen T H, Feng X & Hadid A (2019) Audio-visual kinship verification in the wild. Proc. IEEE International Conference on Biometrics (ICB 2019), Crete, Greece, pp. 1-8.

VI Wu X, Feng X, Zhang X, Bordallo López M & Liu L (2022) Audio-visual kinship verification: a new dataset and a unified adaptive adversarial multimodal learning approach. Submitted to IEEE Transactions on Cybernetics, (Major revision). `https://doi.org/10.36227/techrxiv.19776007.v1`.

Original publications are not included in the electronic version of the dissertation.

827. Li, Yante (2022) Machine learning for perceiving facial micro-expression

828. Behzad, Muzammil (2022) Deep learning methods for analyzing vision-based emotion recognition from 3D/4D facial point clouds

829. Leppänen, Tero (2022) From industrial side streams to the circular economy business : value chain, business ecosystem and productisation approach

830. Tuomela, Anne (2022) Enhancing the safety and surveillance of tailings storage facilities in cold climates

831. Kumar, Dileep (2022) Latency and reliability aware radio resource allocation for multi-antenna systems

832. Niu, He (2022) Valorization of mining wastes in alkali-activated materials

833. Jayasinghe, Laddu Praneeth Roshan (2022) Coordinated multiantenna interference mitigation techniques for flexible TDD systems

834. Zhu, Ruixue (2022) Interaction peculiarities of red blood cells and hemorheological alterations induced by laser radiation

835. Kuosmanen, Elina (2022) Technological support for Parkinson's disease patients' self-care

836. Li, Jing (2022) Advanced high and low field ^{1}H and ^{129}Xe NMR methods for studying polymerization, curing and pore structures of geopolymers

837. Carneiro de Melo, Wheidima (2022) Deep representation learning for automatic depression detection from facial expressions

838. Hannula, Jaakko (2022) Effect of niobium, molybdenum and boron on the mechanical properties and microstructures of direct quenched ultra-high-strength steels

839. Rajaniemi, Kyösti (2022) Electrocoagulation in water treatment: continuous versus batch processes and sludge utilization

840. Ramezanipour, Iran (2022) Hybrid spectrum mechanism for energy vertical

841. Saukko, Laura (2022) Managing integration capabilities in collaborative inter-organizational projects

842. Tiensuu, Henna (2022) Modelling the quality of the steel products under challenging measurement conditions

843. Avsievich, Tatiana (2022) Red blood cells and novel nanomaterials: towards nanosafety and nanomedicine

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-3423-6 (Paperback)
ISBN 978-952-62-3424-3 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)

UNIVERSITY OF OULU