KINSHIP VERIFICATION ON FAMILYSHIP FACE VIDEOS

Anonymous VCIP Submission
Paper ID:

***Abstract*—In this paper, we investigate the problem of video-based parent-child relationship prediction via human face analysis. Most of the existing kinship verification methods are based on single images; these approaches cannot effectively utilize videos of the face for kinship verification. Recently, there have been many methods developed to determine parent-child relationships based on face video, but all of them only do pairwise comparisons between the father and the child or the mother and the child. Thus, they cannot effectively combine information about both the father's and the mother's faces when judging the blood relationship. In this paper, we propose our own dataset, Familyship Face Videos in the Wild (FFVW), which is captured both in wild conditions and standard reference, to deal with this issue. The inputs of FFVW are three separate videos of a family. To our knowledge, our paper is the first attempt at addressing this problem. In our preprocessing step, we extract four key frames from each video, before doing facial recognition and alignment. Finally, we use a convolutional neural network to make the prediction. Overall, the effectiveness of this approach is verified by experimental results, which show that our dataset outperforms previous approaches to kinship verification.**

***Index terms*– Kinship verification, Familyship Face Videos in the Wild (FFVW), Face recognition, CNN.**

I. INTRODUCTION

Kinship verification is especially different from other uses of face recognition, because it has important applications in the area of public security. In countries all over the world, there are many cases of lost children every day; however, it is often extremely difficult for the police to find the children's families. There are two main reasons: the amount of relevant images is limited because lost children are usually found in a different area or state and the cost of DNA recognition is very high but its return rate can be relatively low. As a result, many police stations are using short video clips from traffic supervision records to locate faces that might be victims of kidnappers. This can be widely used before resorting to DNA testing. Kinship verification is among one of the most popular topics in computer vision nowadays. Many algorithms, for example CNN[1] and Haar[2], have been widely applied to solve real life problems. Generally speaking, current face recognition approaches are divided into several categories: PCA, neural networks, and local face analyses. These methods have been proved to be efficient, reaching accuracies of $80\% - 90\%$, leading to their widespread application in facial recognition. However, these previous approaches have fallen short in one key aspect. They all rely on pairwise comparisons, which means that the face information of both the father and the mother cannot be effectively combined.

In this paper, we focus on a method to perform kinship verification between 3 members of a family based on short video clips instead of a single graph. For our neural network-based approach, we require a larger dataset of families. However, we failed to find sufficiently large, high-quality datasets currently in use by facial recognition researchers. Thus, we built such a dataset by ourselves. Then, we designed a pipeline to extract the most informative frames from video clips, to which we can then apply convolutional neural networks (CNNs) to classify the familial relationship.

II. RELATED WORKS

Face recognition has a long history of research. Galton published two articles on the use of human faces for identity recognition in Nature in 1888[3] and 1910[4] and analyzed human face recognition capabilities. However, it was not possible at that time to perform automatic recognition of human faces. The earliest research papers on AFR1, such as the technical report published by Chan and Bledsoe in 1965[5], have been in existence for 40 years. In recent years, well-known systems of methods in the field are designed to process face images. For instance, DeepFace[6] applied CNNs to minimize the distance between the distance between incongruous pairs. There are several approaches for face recognition, including Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Principal Components Analysis (PCA) and convolutional neural network. Among them, CNN provides better accuracy than searching for patterns or statistical approach. HOG generates result slow and performs sensitive to hot pixel despite of its simplicity in principle. LBP calculates in a relatively quick speed and is greyscale non-sensitive, however, the accuracy of it does not stand out and is not able to provide the corresponding robustness as CNN. PCA works well when the lightening variation is small but fails to show the same accuracy when training and testing datasets differ a lot in greyscale. As a result, though application of CNN requires robust datasets, its good behaviour stands out and becomes our final choice, which prompts the generation of our Familyship Face Videos in the Wild (FFVW). As well as algorithm, multiple public datasets such as LFW[7], CelebFaces[8], WDRefs[9] have been made available to researchers, speeding up training and allowing for common performance benchmarks. While most facial recognition datasets are composed of individual images without any correlation, video clips are more easily obtained in real life situations. This led us to develop our method to utilize video clips instead of single images.

For the tri-subject kinship verification problem, we searched for a variety of methods for evaluating the relationship precisely and timely. One method is DNA profiling (also called DNA testing), which uses PCR to determine genetic family relationships. This well-developed technology can reach an accuracy rate of more than 99.99%. However, DNA profiling falls short on several fronts. It is often difficult to collect the required DNA samples and requires several days to obtain results. This is only acceptable for situations where high accuracy is required and high latency is tolerable. The images in these databases were all from the network and contained 4 parent-child relationships. Currently known parent-child databases are the KinFaceW-I[10], KinFaceW-II[10], CornellKin[14] and UB KinFace[5] databases with 533, 1000, 150 and 90 pairs of parent-child images, respectively. As a result, we need a method for kinship verification that runs on much smaller timescales.

III. Dataset Generation

Dataset collection is an important aspect of video-based face recognition in order to generate a precise machine-learning result. A large, high quality video dataset yields better results for machine learning than an image-based database, as it provides much more temporally correlated data. Most of the existed datasets for face recognition are image-based, as shown in Fig 1, which give us a great example and guides our routine in building our own video-based dataset : Familyship Face Videos in the Wild (FFVW). We collected 100 groups of videos from different families, each containing 3 separated video of mother, father and child. Nearly 80% of these videos are recorded from talk shows online, while the other 20% are filmed by people in real life. We are planning to make our dataset, FFVW, open for everyone. After generating FFVW, we transformed it to image-based dataset using matlab and filtered the result to image of blood relationship. We present our dataset by Introduction of FFVW, Generation of FFVW, Strategy for filtering and Methods to prove precision.

A. Introduction of FFVW

Familyship Face Videos in the Wild (FFVW) is a video-based dataset for blood relationship face recognition. FFVW is grouped by family. Inside each group, it is identified by label Father_1, Mother_1, Child_1, etc(each group has 3 labels and each label has at least 4 elements for face recognition). Each element is a short video from 5 second to 60 second length which contains face of the candidates from the central view with the same size and resolution rate. These videos form 100 groups build up the basic dataset of FFVW, which were captured both in wild conditions and standard reference, as shown in Fig 2.

B. Generation of FFVW

The importance of FFVWs generation is incredible due to the foundation of both video-based and blood relationship-based. Restrains of multi-point blood relationship makes the available data from huge dataset decrease in an exponential trend. As a result, FFVW contains high-quality videos from different resources include public videos from the web and private videos from our volunteers. Public videos centered on celebrities, such as Obamas family and other royal families which have been in public for an interview and also family that participated together in TV shows. Private videos are generated by our volunteers (most are students from BUPT and their families). Obviously, existing data sample are mainly

Fig. 1. Existing dataset samples from TSKinFace database[11].

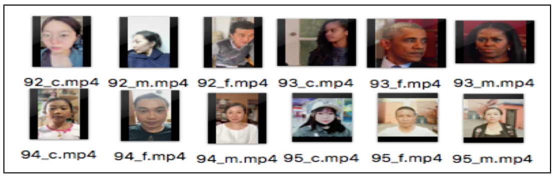

Fig. 2. FFWV samples.

faces with similar expressions and angles, while our FFVW contains short videos of different expressions, angles, colors, etc. FFVW not only provides material more readily available, but also illustrates the relationship between three members in a family rather than ordinary pairwise-based relationship.

IV. Data Preprocessing

As video-based dataset provides abundant data, data preprocessing is necessary in order to make the machine learning methods as precise and consistent as possible. We do the preprocesing in multiple steps.

A. General Manual Filtering of Videos

Before sorting the data, we first go through each video manually to remove potential errors in our dataset. Vague groups of videos were deleted, including those with uncertain familial relationships, poor video quality, or other mistakes that occurred during dataset generation. Videos where the face is visible for less than 10% of the total screen were also deleted or replaced. In our dataset, 100 groups (300 videos) remain after this process.

B. Extracting Key Frames from Videos

This step aims to reduce the variable length videos to a few key frames that are representative of the overall video. We applied k-means extraction algorithm based on video clustering[12]. First we get several frames of the video, as shown in Fig 3. Then algorithm is set to capture the steadiest 4 pictures, during which the person has the smallest change in facial expression.

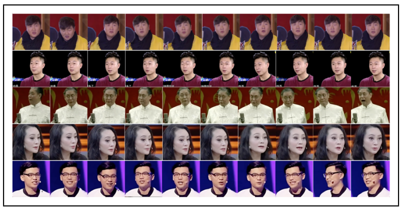

Fig. 3. Example of the steady frames in the videos, chosen by our algorithm.

V. FFVW Training

There are five steps in our FFVW training, as shown in the Fig 5. We discuss each step separately: video processing, face detection, face alignment, feature recognition and output results. For each set (videos of a family), we test 5 times overlying and our dataset is separated as shown in Fig 4 (60% training set, 20% development set and 20% test set):

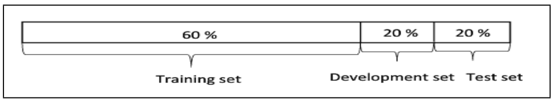

Fig. 4. Pipeline for kinship verification.

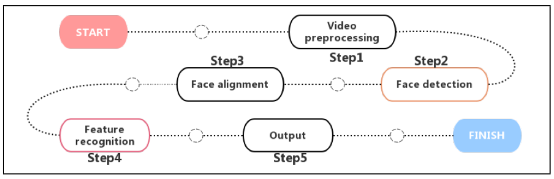

Fig. 5. Distribution of FFVW.

A. Video preprocessing

Fig. 6. After preprocessing, videos become a set of 4 representative images.

The first step can be concluded from dataset collection above. Preprocessing yields high quality inputs to the next steps of the pipeline. In Fig. 6, we were able to extract the best 4 frames from a video where the woman is making faces(sample of video 95_m.mp4).

B. Face detection

We use face detection to determine whether there is a face image in the captured image. Doing so allows us to extract more informative features downstream, and provide consistent inputs to our CNN. In Fig. 8, our face detection algorithm is even able to work in situations with different light intensity, unusual expressions, or abnormal inclination.

Fig. 7. Our face extraction process is robust to dim light, unusual expressions, and different face inclinations.

C. Face alignment

We apply ASM[13] to align the faces. ASM is an algorithm based on the Point Distribution Model (PDM). In PDM, objects with similar shapes — such as the face, hand, or heart — can be expressed in series by a number of key feature points (landmarks) in series. We locate 28 exact points on each image and align them.

Fig. 8. Example of facial landmark detection.

D. CNN Classification

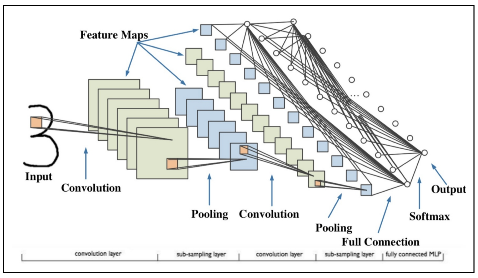

Fig. 9. Neural network architecture for classifier.

Based on their performance on other computer vision tasks, convolutional neural networks offer high accuracy and robustness when trained on sufficiently large datasets[citation]. We use the architecture outlined in Fig. 9. The convolution layer uses twelve 5×5 filters. We apply a maxpooling layer with a 2×2 kernel and a stride of 2. This downsampling procedure reduces the number of parameters in our neural network, while also introducing translational invariance into our classifier. We flatten the resulting output into a vector, apply a fully connected layer, and use a sigmoid nonlinearity to output the probability that the input faces correspond to a family.

VI. Experiments and Results

We did several experiments to not only compare FFVW to current methods, but also examine the effects of the data processing pipeline. To the best of our knowledge, not many people have yield results in the situation of FFVW and it is quite difficult to find an existing method which can directly compare to ours. In the following we present you with our tests in different datasets and some comparisons of necessary of the components in our structure. Fig. 10 shows the

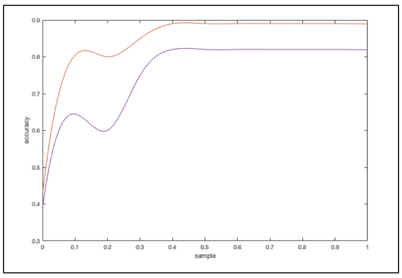

Fig. 10. ROC curve for FFVW with (blue) and without (red) normalizing the brightness of the input images.

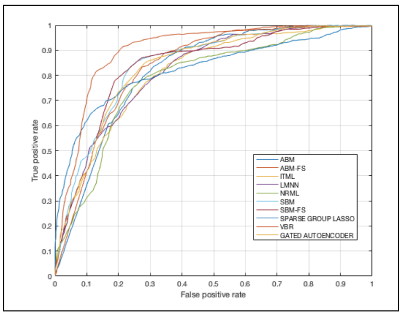

Fig. 11. ROC curve comparison of our approach to current methods.

increased accuracy when normalizing the brightness of input images.According to well-known experience, the intensity of light is a very important factor in face recognition. Pictures in dim light always yield unsatisfying results compared to those in regular light. We tested our database using existing learning and training method of CNN, and evaluate the accuracy of prediction. Our results are shown in the figure, where dim light still proves to have some influence of losing accuracy. Furthermore, we found that the face alignment process increases the classifier accuracy from 83.06% to 89.42%. Both of these results demonstrate large improvements from ensuring the consistency of inputs. We noticed that traditional CNN doesn't contain the process of face alignment, which cannot be left out when using our FFVW dataset. The reason is that we have to deal with video jitter. Screenshots from the video may provide different angles of faces and this makes face alignment indispensable. We compare the accuracy of prediction results with and without face alignment. Fig. 11 compares the ROC curves corresponding to our method and current methods. Table from [11] is combined to summarize the results. It can be seen, vividly, our video-based blood relationship (VBR) based on FFVW improves the performance by 4-48.6% compared to other methods like SBM and ABM. It can be concluded that CNN can reveal an amazing performance on our dataset, not to mention the performance of other advanced algorithms. Finally, we observe that our method is able to obtain an accuracy for pairwise kinship verification. This implies that the FFVW can be successful with fewer inputs.

VII. Conclusions

In this work, we made two main contributions. First, we use videos instead of graphs to work on tri-subject kinship verification. We built a video-based dataset of families, allowing us to provide more training data and learn a more robust classifier. The second contribution is that we apply convolutional neural networks, along with improvements like face alignment, leading to encouraging results. Furthermore, we prove that our method is feasible for the traditional pairwise kinship problem. In the future, we would like to expand our dataset and construct deeper models, which may allow for even better accuracy.

References

[1] Martin Thelwall. Introduction to Webometrics:Quantitative Web Research for the Social Sciences. 2009.

[2] Radomir S. Stankovic; Claudio Moraga; Jaakko Astola. Fourier Analysis on Finite Groups with Applications in Signal Processing and System Design. 2005.

[3] Francis Galton, Personal identification and description. Nature, 1888, pp.173-7, 201-2.

[4] Francis Galton, Numeralised profiles for classification and recognition. Nature, 1910, pp.127.

[5] Shao Ming, Kit D, Fu Yun. Generalized transfer subspace learning through low- rank constraint[J]. International Jour- nal of Computer Vision, 2014, 109(1/2): 74-93.

[6] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, British Machine Vision Conference, 2015.

[7] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]//Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition. 2008.

[8] Y.Sun,X.Wang,and X.Tang.Deeplearning face representation from predicting 10,000 classes. In Proc. CVPR, 2014.

[9] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In Proc. ECCV, pages 566579, 2012.

[10] Lu Jiwen, Zhou Xiuzhuang, Tan Y-P, et al. Neighborhood repulsed metric learning for kinship verification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(2): 331-345.

[11] Xiaoqian Qin, Xiaoyang Tan,Songcan Chen, Tri-Subject Kinship Verification: Understanding the Core of A Family, IEEE Transactions on Multimedia, Accepted, 2015.

[12] N. Petrovic, N. Jojic and T. H. Huang, Hierarchical video clustering, IEEE 6th Workshop on Multimedia Signal Processing, 2004., 2004, pp. 423-426.

[13] Milborrow S., Nicolls F. (2008) Locating Facial Features with an Extended Active Shape Model. In: Forsyth D., Torr P., Zisserman A. (eds) Computer Vision ECCV 2008. ECCV 2008. Lecture Notes in Computer Science, vol 5305. Springer, Berlin, Heidelberg.

[14] Fang Rugou, Tang K D, Snavely N, et al. Towards compu- tational models of kinship verification[C]//Proceedings of the 17th IEEE International Conference on Image Processing, Hong Kong, China, Sep 26-29, 2010. Piscataway, USA: IEEE, 2010: 1577-1580.