Explainable Kinship: The Importance of Facial Features in Kinship Recognition

Britt van Leeuwen[a], Arwin Gansekoele[b], Joris Pries[c], Etienne van de Bijl[d] and Jan Klein[e]

Centrum Wiskunde & Informatica, Stochastics group
Science Park 123, Amsterdam, the Netherlands
Email: [a]britt.van.leeuwen@cwi.nl, [b]arwin.gansekoele@cwi.nl,
[c]joris.pries@cwi.nl, [d]etienne.van.de.bijl@cwi.nl, [e]jan.klein@cwi.nl

***Abstract*—Kinship Recognition, the ability to distinguish between close genetic kin and non-kin, could be of great help in society and safety matters. Previous studies on *human* kinship recognition found an interesting insight when looking for the most important features. Results showed that analyzing only the top half of a face gives equal or even better performance compared to analyzing the whole face. In this paper, we aim to find the important features for *automated* kinship recognition based on the theory of *human* kinship recognition; this set of features was researched using features from pre-trained metrics from the StyleGAN2 model. We found that the most important facial features from the selection of 40 features are mostly focused on the facial hair traits. Furthermore, age-related features were found to be very important. This set of features does not entirely comply with the set of features important in *human* kinship recognition. Previous research has shown *human* kinship recognition performance does not decrease when removing the bottom half of the image of the face. In contrast, our results show that for *automated* kinship recognition, removing either the bottom or the top half of a face results in a decrease in the performance of our classifiers.**

***Keywords*—kinship recognition; StyleGAN2; Families-in-the-Wild; feature importance; transfer learning.**

I. INTRODUCTION

A. Kinship Recognition

Kinship Recognition (KR) is the ability to distinguish between close genetic kin and non-kin. The distinction involves people who are directly related and people who are not. One example of the usage of KR is on families who are spread throughout multiple refugee camps. One of these cases involved a father and his daughter being in one camp, while his wife and other children were in another camp. It took them over a year to get reunited by the Red Cross Restoring Family Links [1]. If a KR system is able to pick such family members out as a possible match for a kinship relation, a family could be reunited almost instantly. Issues with communication and limited manpower could be reduced with the discussed automation.

The main contribution of this paper is to make a first step towards understanding automated KR and the importance of facial features in it. In the field of KR, there is a lot of room for improvement, especially on the importance of facial features. This is what we tackled in our research by researching whether kinship is recognizable by using a set of extracted facial features with the use of machine learning. Specifically, we focus on what specific set of features is important for automated kinship recognition and if this set of features complies with the set of features important in human kinship recognition. First presented in this paper is a literature discussion on human as well as automated kinship recognition. Then, in Section II, the data is discussed. In Section III, an overview of the used models is presented. Next, the results of different experiments are discussed in Section IV. Lastly, a discussion and conclusion of the presented experiments is given in Sections V and VI.

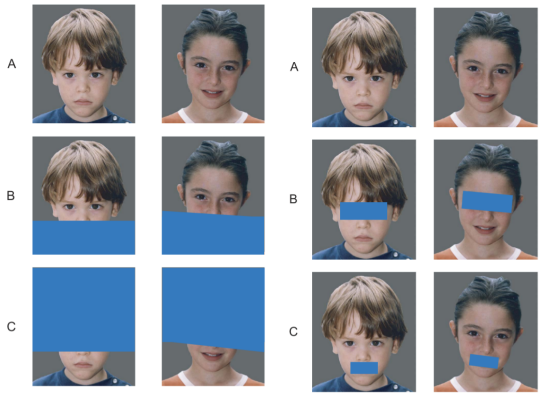
B. Related work

Studies on human KR contribute to our search for the set of important features in automated KR. Several studies [2]–[4] have been conducted on human KR, which showed that kinship is indeed recognizable by humans. Robinson et al. [5] used the Families-In-the-Wild (FIW) data set for their human performance measurement. This data set contains images of people's faces that are extracted from family pictures. Robinson et al. state that humans scored an overall average of 56.6% accuracy. Other research on KR [3], [6], [7] shows similar results. One of the interesting results is that the average accuracy of human KR is higher when face, hair color and background are taken into account compared to when the focus is purely on the face.

We take a look at the Feature Importance (FI) in some of these studies on human kinship detection. The reason behind this specific set of features for human KR might be of help in automated KR. One of the studies is by Martello and Maloney [2], [3], who raised the question which parts of a face are most important for human KR. In [2], they conducted a study in which humans were tested on their KR skills based on three separate conditions: (1) the right hemi-face masked, (2) the left hemi-face masked, and (3) the face fully visible. Most interestingly, the results showed that there is no significant difference in results for recognizing kinship by humans when the left or right part of the face is covered. On the contrary, a similar study [3] showed that the covering of the top or bottom part of a face does give a significant difference. The effect on kin recognition performance of masks that covered the upper half or the lower half of the face (experiment 1) and the eye region or the mouth region (experiment 2) were measured. An example of the covering up of facial parts for experiment 1 and 2 can be seen in Figure 1a and 1b below. In these experiments, it was found that masking the eye region led to

 ISBN: 978-1-68558-017-9

a 20% reduction in performance, whereas masking the mouth region led to a non-significant, although fascinating increase in performance. This leads us to consider the insight that the performance in KR is heavily dependent on only the upper half of a person's face.



(a) Experiment 1: Masking the bottom and top half of the face

(b) Experiment 2: Masking the eye area and lip area of the face

Figure 1. Illustration of the masking of faces in [3].

Overall, the theory that we researched is based on the possible change in performance when using a specific set of facial features compared to facial features from the whole face. This could lead to only requiring specific parts of faces to identify kinship relations, thus to more accessible data and a decrease of the computational costs of KR models.

For *automated* KR, several approaches have been proposed. Most approaches are not only focused on machine learning models, but also on feature selection. Feature-based methods aim to preserve facial, genetically determined characteristics in the feature descriptors used for the model. These methods identify local facial features such as inconsistencies of an individual's eyes, mouth, nose and skin from the individual's image. Feature-based methods can decrease computational cost and improve the model's performance. Most of the proposed models and algorithms were trained on only small data sets.

These data sets demonstrated to be insufficient for the task at hand. Most of the proposed classifiers are lower-level models and algorithms, which use handcrafted feature extraction (features using information presented in the image itself), Support Vector Machines (SVMs) or K-Nearest Neighbor classifiers.

Since 2016, a more extensive data set has been constructed in [4]: Families-in-the-Wild (FIW). This data set has been produced to verify kinship and classify relations [8]. The creators of this data set specify promising results in detecting kinship. Robinson et al. [5] state the best results were obtained when using the SphereFace model with an average accuracy of 69.18% and standard deviation of 3.68. All models performed well compared to previous work, although much improvement could still be made.

After publishing this FIW data set, more research in the field of KR models was done. Many models in KR include the use of FaceNet or other small feature selections for their models' input [9]. FaceNet is a neural network that extracts features of an image. The model provides a mapping from a picture of a face to the Euclidean space. The distances in this space correlate to the amplitude of face resemblance [10]. It produces an output vector to be used as input for a classification model. FaceNet creates embeddings by learning the mapping from images. A disadvantage of using FaceNet is that especially when looking at FI, information gets lost due to lack of feature interpretation [11].

FaceNet could help improve KR models, although we are interested in the similarities between faces by using facial features instead of the faces as a whole. Hence, we use a different approach than FaceNet. Fang et al. [12] proposed different feature extractions. One of the extractions is based on different colors of different parts of the face. Other extractions are based on image coordinates of certain parts of the face, facial distances and gradient histograms. Together, these feature extraction methods constructed 44 facial features. The top selected features are right eye RGB color, skin gray value, left eye RGB color, nose-to-mouth vertical distance, eye-to-nose horizontal distance and left eye gray value. The results show a high importance for eye related features. 10 out of the 14 top features include the eye area. While this study does include specific facial features like eye color, it only included 22 low-level features. It is indeed shown that most of the selected features are in the upper face area, which complies with the insight.

Most studies on the subject focus on either the overall similarities between faces, or on pre-determined facial feature sets. These studies treat KR tasks similar to the task of a standard facial recognition. Guo et al. [13] argue that kinship classification should be treated differently, since trait similarities are measured across age and gender. Additionally, kinship has a combination of traits and familial traits are special for each family pair.

Models proposed by researchers in this field are based on an input of just the images with little to no alterations. Although, some research focus on specific facial features by using for example a weighted graph embedding-based metric learning framework [14] or by using sparsity to model the genetic visible features of a face [15].

Another group of researchers thought of combining the StyleGAN2 algorithm with KR [16]. In the task at hand, there is a restriction that family members should be recognized on the basis of physical facial features. However, several mentioned attempts neglect this constraint and do not employ any facial landmark before using a classification model. For this reason, Nguyen et al. [16] experimented with KR models using StyleGAN2 as an encoder to incorporate a facial landmark map. This method resulted in an average accuracy of 0.548 for recognizing kinship. Against expectations, no improvement was shown in the results from using StyleGAN2 in this manner, which is presumed to be due to the lack of a proper classification and thus it is argued to need more investigation. An algorithm proposed by Guo et al. [13] uses

 ISBN: 978-1-68558-017-9

familial traits extraction and kinship measurement based on a stochastic combination of the familial traits. The authors use a similarity score based on a Bayes decision for each pair of facial parts. However, facial features used by the algorithm are limited to the eyes, nose and mouth and, in line with the observations by Guo et al. [13], more parts of the face could be explored. Existing data sets use faces from the same family picture, so models learn about the background similarity. This causes the models to get a higher performance, but when tested on real life pictures, not taken from a family picture, the performance could be lower. When using pre-determined features, this does not present a problem.

II. DATA

We used the Families in The Wild data set. The data is split up into training and test data using hold-out cross validation. The data is split up in a 70/30 split, respectively. The training set consists of information on families, persons and relations between persons including images of the persons. The data is distributed as follows: an average of about 12 images per family, each with at least 3 and as many as 38 members. Each family is assigned a unique id, each person is assigned an id and each image collected is assigned a unique id. The data set includes good quality images of a person's face, but also blurry images of faces, as shown in Figures 2a and 2b, respectively.

(a) Image from training data

(b) Blurry image from training data

Figure 2. Example data from the Families-in-the-Wild data set

A file containing all matches in the training data set is available. However, this does not include data on combinations of persons that do not have a familial relationship. So, these pairs have been constructed by taking random pairs of images from the set of training images of the FIW data set. This is excluding existing related pairs and each pair is unique. This resulted in 205,285 related and 205,285 unrelated pairs of images.

StyleGAN2 metric: linear separability

This research is focused on FI in KR. To be able to understand the FI of a model, the features extracted from a model should be interpretable. To collect a bunch of features and to avoid having to do manual annotation, we decide to use a feature description method from the StyleGAN2 model. With this, it can be easily deducted which of the features of a face are seen as most important by a model for detecting kinship. The pictures in the data are of size 108x124, while the StyleGAN2 description method expects pictures of size 256x256 as input. Interpolation of the pictures in the data is used to overcome this problem. The StyleGAN2 model contains a certain metric called linear separability. StyleGAN2's linear separability metric can be used to steer a generated picture in a certain direction by specifying 40 facial features which are shown in Table I. For example, the models can be used to make the generated face have blond hair and high cheekbones. What we are most interested in for this research are the pre-trained models used in StyleGAN2 which produce probabilities of the 40 features to be true for an image of a person.

TABLE I. FACIAL FEATURES OF LINEAR SEPARABILITY METRIC

1) 5-o-clock-shadow,	15) double chin,	28) pointy nose,
2) arched eyebrows,	16) eyeglasses,	29) receding hairline,
3) attractive,	17) goatee,	30) rosy cheeks,
4) bags under eyes,	18) gray hair,	31) sideburns,
5) bald,	19) heavy make up,	32) smiling,
6) bangs,	20) high cheekbones,	33) straight hair,
7) big lips,	21) male,	34) wavy hair,
8) big nose,	22) mouth slightly open,	35) wearing earrings,
9) black hair,	23) mustache,	36) wearing hat,
10) blond hair,	24) narrow eyes,	37) wearing lipstick,
11) blurry,	25) no beard,	38) wearing necklace,
12) brown hair,	26) oval face,	39) wearing necktie,
13) bushy eyebrows,	27) pale skin,	40) young.
14) chubby,		

The metric was trained using the CelebA Data set (CelebFaces Attributes Data set). This is a face attributes data set with 202,599 celebrity images, each with five landmark locations and 40 attribute annotations. StyleGAN2's linear separability metric is meant to be used for the StyleGAN2 model and its corresponding data. We are interested in using the metric on the data from FIW. The information gathered from the linear separability metric (the facial features) is used as a starting point for the kinship classification models. Transfer learning does not only save time, but it also has the possibility of making a learning process more efficient [17].

Consequently, some adjustments to the data were necessary to apply the metric. This resulted in an output of 40 features for all images in the data set, which then could be used to train the chosen automated KR models. As data points for the models, we chose a list of length 40 and a list of length 80, composed of the metric values for the features per two pictures. Two input types were experimented with: (1) a list of 80 features, consisting of 40 features per image, and (2) a list of 40 features, taking the absolute difference of the feature values between the images per feature.

III. MODEL DESCRIPTION

We implemented and tested several models to see how well the models work on our data and to find a recurring pattern in FI. For all models, the FI is investigated. The results of this are then used to understand whether the theory of human KR will hold for automated KR as well. Various machine learning models were selected for this task. For each model, the accomplished accuracy is obtained by K-fold cross validation. The number of folds is set to 10 and the data is shuffled before splitting into batches.

 ISBN: 978-1-68558-017-9

Machine learning methods

Using StyleGAN2's linear separability metric on our data results in an output of 40 features for all images in the data set, which then are used to train the models. As data points for the models, we chose a list of length either 40 or 80, composed of the metric values for the features per two pictures. The models we decided to experiment with are the following:

First, we have the decision tree algorithm with maximum depth set to 10, where we obtain the FI by using the Gini importance. Second, we have the random forest consisting of 100 trees, where the FI is obtained by using the impurity importance. Then, we have the Gaussian Naive Bayes, which obtains FI by using the permutation importance. Next is the linear SVM, where the weights of the model are used to determine FI. Lastly, we have logistic regression, where the FI is determined by using the coefficients of the decision function.

IV. Results

Two different approaches have been researched, the original StyleGAN2 description method and the bottom and top masked method. The results of these approaches are discussed and an overview of the results is provided.

A. Original StyleGAN2 descriptor experiment

The initial approach is taking the results of the StyleGAN2 model and using them as input for the different algorithms. Over all images, we calculated the probabilities of the image complying with the given 40 features. Extracting 40 features per picture resulted in 80 different values since we were working with two images per data point. The FI was determined per model. For the 80 feature input, we took the sum of each feature per picture. An overview of all the results from the StyleGAN2 descriptor experiment can be found in Table II and Table III.

Decision Tree: The accuracy of the decision tree with 40 features as input has mean 0.61 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.66 with a standard deviation of 0.005. The model is more leaning towards giving a positive (related) classification. For the decision tree model with input of 40 features, *arched eyebrows*, *no beard* and *heavy makeup* are the most important features. For the input of 80 features, the top three of important features is *young*, *no beard* and *wearing necklace*.

Random Forest: The accuracy of the random forest with 40 features as input has mean 0.74 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.80 with a standard deviation of 0.004. The model does not have a clear preference for either a positive or a negative classification. With the model giving 51.39% and 50.63% positive classifications for 40 and 80 features respectively, the even distribution of the data in half positive and half negative data points is represented well with a slight deviation towards positive classifications. For the random forest model with input of 40 features, *arched eyebrows*, *mustache* and *heavy make up* are the most important features. For the input of 80 features, the top three of important features is *young*, *no beard* and *mustache*.

Gaussian Naive Bayes: The accuracy of the Gaussian naive Bayes with 40 features as input has mean 0.60 with a standard deviation of 0.004. The 80 features input gives a mean accuracy of 0.59 with a standard deviation of 0.005. The model has a preference for a positive classification. With the model giving 59.56% and 64.21% positive classifications for 40 and 80 features respectively, most errors are false positives. For the Gaussian naive Bayes model with input of 40 features, *eyeglasses*, *mustache* and *arched eyebrows* are the most important features. For the input of 80 features, the top three of important features is *eyeglasses*, *rosy cheeks* and *no beard*.

Linear Support Vector Machine: The accuracy of the linear SVM with 40 features as input has mean 0.59 with a standard deviation of 0.004. The 80 features input gives a mean accuracy of 0.63 with a standard deviation of 0.005. The model does not have a clear preference for a positive or negative classification. With the model giving 47.08% and 52.92% positive classifications for 40 and 80 features respectively, we see a slight effect of the different input values. The 40 values input gives the model a bit more lenience towards negative classification and the 80 values input gives the model slightly more lenience towards positive classification. For the LSVM model with input of 40 features, *arched eyebrows*, *no beard* and *heavy make up* are the most important features. *no beard* and *arched eyebrows* are also among the most important features for the input of 80 features. Here the top three of features is *arched eyebrows*, *narrow eyes* and *no beard*.

Logistic Regression: The accuracy of the logistic regression with 40 features as input has mean 0.60 with a standard deviation of 0.003. The 80 features input gives a mean accuracy of 0.63 with a standard deviation of 0.005. The model does not have a clear preference for a positive or negative classification. The model gives 50.40% and 51.61% positive classifications for 40 and 80 features respectively, which shows the balance of the data with a slight deviation towards positive classification. For the logistic regression model with input of 40 features, *arched eyebrows*, *no beard* and *eyeglasses* are the most important features. These are also among the important features for the input of 80 features. Here the top three of important features is *no beard*, *arched eyebrows* and *pale skin*.

B. Masked StyleGAN2 descriptor experiment

To support the theory we found, all of StyleGAN2's linear separability features were taken of not the original image, but over an image with the bottom part of the face masked black like shown in Figure 3. The same was done with the top part of the face masked black, comparable to the experiments performed by Martello et al. [2], [3]. All the models are exactly the same as for the original StyleGAN2 description method. Only the input changed.

Bottom half masked: This experiment was done with all models previously used in the original StyleGAN2 descriptor

 ISBN: 978-1-68558-017-9

(a) (b)

Figure 3. Example data from the Families in the Wild data set with bottom masked (a) and top masked (b)

experiment. The accuracy and FI were obtained for the decision tree, random forest, Gaussian naive Bayes, LSVM and logistic regression models. An overview of the accuracy and important features for all the models from the bottom masked StyleGAN2 descriptor experiment can be found in Table II and Table III. Again, the results show that the 80 value input gives and overall better performance than the 40 value input and the best performing model is random forest for both inputs. Some of the most important features for the bottom masked approach are related to the nose (*pointy nose* and *big nose*) and the hair (*grey hair*, *blond hair* and *waivy hair*).

Top half masked: This experiment was done with all models previously used in the original StyleGAN2 descriptor experiment. The accuracy and FI were obtained for the decision tree, random forest, Gaussian naive Bayes, SVM and logistic regression models. An overview of the accuracy and important features for all the models from the bottom masked StyleGAN2 descriptor experiment can be found in Table II and Table III. Again, the results show the 80 value input to give overall better performance than the 40 value input and the best performing model is random forest for both inputs.

TABLE II. ACCURACY FOR THE 40 AND 80 VALUE INPUT PER EXPERIMENT: COMPLETE, BOTTOM MASKED AND TOP MASKED

	40 Compl.	**40 Bottom**	**40 Top**	**80 Compl.**	**80 Bottom**	**80 Top**
Decision Tree	0.61 ± 0.003	0.57 ± 0.004	0.57 ± 0.003	0.66 ± 0.005	0.64 ± 0.004	0.65 ± 0.003
Random Forest	0.74 ± 0.003	0.62 ± 0.003	0.63 ± 0.002	**0.83 ± 0.004**	0.81 ± 0.001	0.82 ± 0.001
Gaussian Naive Bayes	0.60 ± 0.004	0.53 ± 0.003	0.55 ± 0.002	0.59 ± 0.005	0.55 ± 0.003	0.57 ± 0.002
Support Vector Machine	0.59 ± 0.004	0.55 ± 0.002	0.57 ± 0.002	0.63 ± 0.005	0.60 ± 0.002	0.61 ± 0.002
Logistic Regression	0.60 ± 0.003	0.55 ± 0.003	0.57 ± 0.002	0.63 ± 0.005	0.59 ± 0.003	0.61 ± 0.002

V. DISCUSSION

Multiple models have been tested on FI. Some approaches were based on the human KR experiments from [2], [3]. These experiments showed a certain area of the face to contain the important facial traits needed for KR. We researched the set of features that is most important for automated KR. Pre-trained metrics from the StyleGAN2 model that are meant to be used for synthesizing artificial examples of faces were used. The pre-trained models give 40 values for specific facial features. These 40 values can also be taken from pictures using the pre-trained models. These values were used as input for our machine learning models: decision tree, random forest, Gaussian naive Bayes, support vector machine and logistic regression. These models were trained and evaluated to show which of the features were seen as most important by the models. More experiments were conducted with the top and bottom parts of a face masked black to also test the theory of human KR.

TABLE III. MOST IMPORTANT FEATURES PER EXPERIMENT

	Complete	**Bottom Masked**	**Top Masked**
Decision Tree	young, no beard, arched eyebrows, eyeglasses	attractive, blond hair, pointy nose, grey hair	young, no beard, arched eyebrows, eyeglasses
Gaussian Naive Bayes	eyeglasses, no beard, young, arched eyebrows	wavy hair, blond hair, pale skin, heavy makeup	eyeglasses, no beard, young, arched eyebrows
Support Vector Machine	young, no beard, pointy nose, arched eyebrows	grey hair, pale skin, wavy hair, big nose	young, no beard, pointy nose, arched eyebrows
Logistic Regression	blurry, no beard, wearing necklace, pointy nose	wavy hair, young, grey hair, big nose	blurry, no beard, wearing necklace, pointy nose
Random Forest	young, no beard, mustache, arched eyebrows	pointy nose, grey hair, smiling, attractive	young, no beard, mustache, arched eyebrows

Major findings: Interesting results were found when comparing the different models using the original StyleGAN2 description method. Four out of five models had a higher accuracy score when all features for both pictures were kept separate. The models are able to learn about combinations of different features between the two pictures, which has a positive influence on the accuracy score of the models.

The best performing model seems to be the random forest. Since this model has a very high accuracy compared to the other models, we are specifically interested in its corresponding FI scores. Accordingly, we mainly focus on the results of the random forest model. This model gives high importance values to the features *young*, *no beard*, *mustache* and *arched eyebrows*. It is also noticeable that in two of the five models, the feature *young* is found to be very important and in the other three models, the FI increases when using 80 features instead of 40 features as input. On top of that, in all models, the features *arched eyebrows* and *no beard* are in the top four of the most important features for the model. There is a clear pattern in the importance of facial hair. Beards, mustaches and arched eyebrows are found to be important features for most of the models. Another pattern is the age difference. This gives us reason to believe that the combination of facial hair and the age of a person is strongly correlated to the classification.

 ISBN: 978-1-68558-017-9

While the correlation scores do not show a correlation between the two features, the combination of the features do matter when comparing two pictures. A reason for these features to be found important is that most of the kinship relations (75%) in the data set are zero generation and first generation relations. Young people are not able to grow facial hair, if they have the genes, it comes with age. This would explain why both facial hair and age are found to be more important.

This set of features that are found to be the most important in our research do not comply with the selection of features proposed by Fang et al. [12]. The set of features used in their research is different, although it is clear that the eye area was found to be the most important by them. Contrasting, the set of important features we found is not particularly focused on the eye area.

For the masked experiment, all five models had a higher accuracy score when all features for both pictures were kept separate. When looking at the bottom masked method results, a clear decrease in the performance is found compared to the original StyleGAN2 description method. Remarkable is that the feature *young* and the features on facial hair are not found in the top features of almost all models. The original StyleGAN2 approach showed these features to be very important. This leads to the believe that the bottom part of a face is essential for extracting the feature *age*. This would also explain why the feature *grey hair* is found to be important in three out of five models. Grey hair is usually a sign of a higher age. When looking at the top masked method results, a decrease in the accuracy is found, although this decrease is not as excessive as with the bottom masked method. Above is mentioned that the feature young is likely to be extracted from mostly the bottom of a face. However, this is not shown in the results of the top masked method. It is curious that the feature *young* is still not found to be one of the most important. Like the original approach, the top masked method shows the feature *arched eyebrows* to be important. Although a pattern is difficult to find in the top masked method results.

For the bottom masked approaches the difference with the original approach is clear. Where humans showed equal or even better performance when masking the top half of a picture, the algorithms showed the opposite effect.

Limitations: The data set might not be very compatible with the StyleGAN2 metrics, which is an uncertainty. However, as for now, there are no other data sets that contain enough images which are of adequate quality. So for now, we have to accept this limitation. An issue was also encountered when using the linear separability metric for a different purpose than StyleGAN2. The results for the top masked method showed one very noteworthy important feature, namely the *arched eyebrows* feature. This feature should be focused on the top part of a face. However, it is found to be important when the top part of a face is masked. More features which show unusual behavior are *smiling* and *pointy nose*, since these are found to be important when masking the bottom half of the face. This is one of the problems that is encountered when combining StyleGAN2 metrics with other models. The models that are trained for the linear separability metric behave different than intuitively expected. Using the metric in tasks for which it is not initially intended can cause limitations to the models.

Unexpected findings: A surprising matter is the difference in performance between the top masked and bottom masked StyleGAN2 description method. Masking the bottom half of the face decreased the performance. As masking the top half of the face decreased the performance as well, it still performed better than the bottom masked method. This is against expectations and raises the question whether the bottom part of a face contains more information than the top part of a face does for KR.

VI. Conclusion

We researched the set of features that is most important for automated KR. For this, multiple models have been tested on FI. The results showed that the most important facial features from the selection of 40 features are mostly focused on the facial hair traits and age related features.

One of the issues we ran into is on transfer learning. The question rises whether StyleGAN2 is compatible enough for transfer learning combined with our data set. It could be more effective to write a new metric that focuses on more solid facial features. Despite that, the StyleGAN2 metrics are the most elaborate method in finding pre-determined facial features. Other models include not as many facial features or need manual annotation. It would be contributory to find a way to annotate all parts of the face for many more features as to train the models on.

In conclusion, this paper is an important first step towards understanding automated KR, but there are many challenges to be faced before it can be used in real-world applications. As it is now, a large set of clear pictures of complete faces are needed for a model to perform decently. Learning more about the most important parts of our face for automated KR is the next step to take to improve the field of KR.

VII. Acknowledgment

We thank Rob van der Mei and Sandjai Bhulai for their useful comments on drafts of this paper.

References

[1] E. Seselja, "How the Red Cross and a radio reconnected a family torn apart by conflict - ABC news", https://www.abc.net.au/news/2021-08-29/red-cross-reconnect-family-separated-by-conflict-after-16-years-/100413214, August 2021, (Accessed on 23/06/2022).

[2] M. F. Dal Martello and L. T. Maloney, "Lateralization of kin recognition signals in the human face", The Association for Research in Vision and Ophthalmology - Journal of vision, vol. 10(8), 2010.

[3] M. F. Dal Martello and L. T. Maloney, "Where are kin recognition signals in the human face?", The Association for Research in Vision and Ophthalmology - Journal of vision, vol. 6(12), 2006.

[4] J. P. Robinson, M. Shao, Y. Wu and Y. Fu, "Family in the Wild (FIW): A Large-scale Kinship Recognition Database", CoRR, abs/1604.02182, 2016.

 ISBN: 978-1-68558-017-9

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

[6] 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 - Elsevier, vol. 49(1), pp. 38–43, 2009.

[7] G. Kaminski, S. Dridi, C. Graff, and E. Gentaz, "Human ability to detect kinship in strangers' faces: effects of the degree of relatedness," Proceedings of the Royal Society B: Biological Sciences, vol. 276(1670), pp. 3193-3200, 2009.

[8] 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(11), pp. 2624–2637, 2018.

[9] R. F. Rachmadi, I. K. E. Purnama, S. M. S. Nugroho and Y. K. Suprapto, "Family-aware convolutional neural network for image-based kinship verification", International Jour- nal of Intelligent Engineering and Systems, vol 13(6), pp. 20–30, 2020.

[10] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[11] L. Dulčić, "Face Recognition with FaceNet and MTCNN - Ars Futura", https://arsfutura.com/magazine/face-recognition-with-facenet-and-mtcnn/, (Accessed on 23/06/2022).

[12] R. Fang, K. Tang, N. Snavely, and T Chen, " Towards computational models of kinship verification", IEEE International Conference on Image Processing, pp. 1577-1580, 2010.

[13] G. Guo and X. Wang, "Kinship measurement on salient facial features", IEEE Transactions on Instrumentation and Measurement, vol. 61(8), 2012.

[14] 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(3) pp. 1149–1162, 2019.

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

[16] T. H. Nguyen, H. H. Nguyen and H. Dao, "Recognizing families through images with pretrained encoder", arXiv, 2020.

[17] Seldon, "Transfer learning for machine learning", https://www.seldon.io/transfer-learning/, June 2021, (Accessed on 23/06/2022).

 ISBN: 978-1-68558-017-9