A Geometric and Fractional Entropy-based Method for Family Photo Classification

[1]Maryam Asadzadeh Kaljahi, [1]Palaiahnakote Shivakumara, [2]Tiang Ping, [1]Hamid A. Jalab, [1]Rabha W. Ibrahim, [3]Michael Blumenstein, [2]Tong Lu and [1]Mohamad Nizam Bin Ayub

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Email: asadzadeh@um.edu.my, shiva@um.edu.my, hamidjalab@um.edu.my, rabhaibrahim@um.edu.my, nizam_ayub@um.edu.my.

[2]National Key Lab for Novel Software Technology, Nanjing University, China. Email: htp@smail.nju.edu.cn, lutong@nju.edu.cn

[3]Faculty of Engineering and IT, University of Technology Sydney, Australia. Email: michael.blumenstein@uts.edu.au

Abstract

Due to the power and impact of social media, unsolved practical issues such as human trafficking, kinship recognition, and clustering family photos from large collections have recently received special attention from researchers. In this paper, we present a new idea for family and non-family photo classification. Unlike existing methods that explore face recognition and biometric features, the proposed method explores the strengths of facial geometric features and texture given by a new fractional-entropy approach for classification. The geometric features include spatial and angle information of facial key points, which give spatial and directional coherence. The texture features extract regular patterns in images. The proposed method then combines the above properties in a new way for classifying family and non-family photos with the help of Convolutional Neural Networks (CNNs). Experimental results on our own as well as benchmark datasets show that the proposed approach outperforms the state-of-the-art methods in terms of classification rate.

Keywords: Face recognition, Facial points, Facial geometric features, Fractional entropy, Convolutional neural networks, Family photo classification.

1. Introduction

The evolution of communication technologies, such as Facebook, Google+, Twitter, Instagram, Flicker and WhatsApp, help people to interconnect quickly (Zhen et al., 2018). One such example is photo-sharing services for social networking. By taking advantage of the advancements in mobile digital camera

technologies, people can easily take photos when they find something interesting and upload them to a social media platform to share exciting moments with their friends, families and colleagues (Cai et al., 2014). As a result, one can expect large collections, which is evident, as the uploaded photo count was "about 4.5 million daily" according to the report in (Cai et al., 2014). In addition, the development of multimedia technologies and cost effective CCTV cameras for surveillance applications produce diversified images or videos at a larger scale. This leads to a huge collection with a high degree of diversity and unstructured data (Shen et al., 2009). For instance, some sample images of family and non-family photos chosen from our dataset are shown in Fig. 1(a) and Fig. 1(b), respectively, where we can see each image has its own variety of foreground (face regions) and background information. In this context, face recognition alone may be insufficient to identify family or non-family photos. This is because the recognition methods developed might not work well for images which contain faces with multiple emotions, postures and actions. This makes the problem of finding photos that belong to the same family complex and challenging. As a result, family photo classification/identification can play a vital role in finding a solution to unsolved issues such as human trafficking, kinship recognition, and the problem of identifying/locating refugees (Robinson et al., 2018). Hence, there is an urgent need for developing an intelligent expert system for tackling the above-mentioned challenges.

(a) Examples of family photos

(b) Examples of non-family photos.

Fig. 1. Samples images of family and non-family photos chosen from our dataset

There are methods for identifying humans, facial expressions and emotions based on biometric features, which can be used for family and non-family image identification (Mehta et al., 2018; Haghighat et al., 2015). However, one major challenge of biometric systems is the variability in characteristics of the biometric of each individual. For example, the human face is complex, with features that change over time. In addition, facial features change due to variations in illumination, head pose, facial expression, cosmetics, aging, and occlusion because of beards or glasses (Haghighat et al., 2015). In addition, most of the methods require cropped face images for achieving better results (Mehta et al., 2018; Haghighat et al., 2015). Therefore, recognition-based systems may not be suitable for family and non-family photo classification because the images can have unconstrained backgrounds and multiple faces with numerous emotions or expressions (Wang et al., 2017). Hence, we can conclude that we need an expert and robust system that can cope with background complexities and issues of multiple faces with different emotions and expressions.

In this work, we propose to find a solution for family and non-family photo classification based on the characteristics defined below for family and non-family images in (Wang et al., 2015, 2017).

In the case of family photos, it is expected that

- Photos will have parents and their children either sitting or standing in a cascaded order. It should not contain persons of different families, namely, more than one family.
- The number of persons in an image should be more than 3, including parents and one child.
- Photos can be captured at both indoor and outdoor areas, such as houses, scenery, parks and tourist places with persons present. In other words, an image can have persons with any background.

In the case of non-family photos, it is expected that

- Photos must have persons with almost the same age, and it is expected persons of different families, for example, friends and colleagues might be present.
- The number of persons in an image should be 3 at a minimum.
- Images must have persons with different poses and any order with any background, which may include indoor and outdoor scenes.

2. Related Work

To overcome the limitations of recognition-based systems, methods which use unsupervised features such as clustering, grouping, and similarity between the parents and children's faces, as well as personal traits such as age, race and gender (Dandekar et al., 2014) have been developed.

Ng et al. (2011) proposed social relationship discovery and face annotations for personal photo collections. This method explores the combination of ensemble RBFNN with pairwise social relationships as context for recognizing people. However, the method requires face annotations and parameter tuning for social relationship identification. In addition, the focus of the method does not relate to family and non-family image classification; rather it explores general social relationships.

Dandekar et al. (2014) proposed verification of family relationships from parents' and children's' facial images. The method uses local binary pattern features and degree of similarity between the faces of children and parents. The method follows conventional feature extraction and classifiers for achieving results. However, the method is good for cropped face images but not those with multiple faces, emotions, expressions and complex backgrounds. In addition, the main target of the method is to match children's' faces with parents' faces but not finding group images.

Xia et al. (2014) proposed face clustering in a photo album, where the method explores spectral features, similarity features, minimum cost flow and clustering. The proposed features are extracted from cropped face images. The main objective of the method is to find images which share the same faces. This idea is good for grouping personal collections but not family and non-family image classification.

Qin et al. (2015) proposed tri-subject kinship verification for understanding the core of a family. The method proposes a degree of similarity between children and parents, resulting in a triangular relationship. To achieve this, the method uses a relative symmetric bilinear model for estimating similarity. To improve the results, the method takes spatial information into account. This method is good as long as the recognition approach provides successful results; however, recognition-based methods may not be robust for the images affected by severe illumination, postures and actions.

Dai et al. (2015) proposed family member identification from photo collections. The method explores an unsupervised EM joint inference algorithm with a probabilistic CRF. The proposed model identifies role assignments for all detected faces along with associated pairwise relationships between them. The performance of the proposed model depends on the success of face detection and recognition; however, the extracted biometric features used to find relationships may not be sufficiently robust when images are exposed to an open environment. In addition, the main target is to identify relationships between members of a family but the approach does not focus on family and non-family classification.

Robinson et al. (2018) proposed visual kinship recognition of families in the wild. This method explores deep learning for face verification, clustering and boosted baseline scores. The method involves multimodal labeling to optimize the annotation process. This includes information of faces and metadata collected from

family photos. It is noted that although the method explores recent powerful deep learning approaches for kinship identification, it is still limited to family photos but not non-family photos.

Wang et al. (2017, 2015) proposed leveraging geometry and appearance cues for recognizing family photos. The methods identify facial points for each face in an image. Based on facial points, the method constructs polygons to study geometric features of faces in the image. Due to the height difference of persons and the arrangement of faces in family and non-family images, the method gets different polygons to study geometric features. It estimates pairwise relationships like kinship recognition, and generates a codebook using k-means clustering. Furthermore, the degree of similarity of each group is extracted for classifying family and non-family photos with the help of an SVM classifier. However, classification may not be accurate when the heights of persons in an image do not follow a hierarchical arrangement. In addition, one might expect that non-family members could have the same arrangements and heights.

In light of the above discussions, we can assert that a few methods have addressed family and non-family photo classification or identification, but most of the methods focus on kinship recognition based on face detection and recognition. These methods may not work well for images where we can see faces with multiple emotions, postures and actions. The methods which addressed family and non-family classification explore only foreground information (facial information) for achieving their results. This is good for images with simple backgrounds but not images that have complex backgrounds, where we can expect open scenes and outdoor environments in the case of non-family photos. Therefore, we can conclude that there is a critical need for an accurate method to classify family and non-family photos.

Hence, we propose a novel method which explores the advantages of spatial and angle information of facial key points and fractional entropy features for classification of family and non-family images. As noted from related work, facial points and geometric features for faces play a vital role in identifying members of a family, including kinship/relationships (Wang et al., 2017; Wang et al., 2015). Motivated by this argument, we propose spatial and angle features in a new way to study geometric structures of faces, which captures the spatial and directional coherence of the face regions. Furthermore, to improve the discriminative power of the features, the propose method explores regular patterns in images. It is observed that in general, persons' standing or sitting arrangements in family photos follow regular patterns such as particular orders, while non-family photos may not follow these. To extract such observations, we propose a novel fractional entropy feature to study the texture of facial regions as well as the background (other than facial region) of images. The combination of spatial information, angles that extract the geometric structure of faces, and fractional entropy that extracts the texture of facial and background regions, produces a feature vector. Furthermore, the feature vector is passed to a Convolutional Neural Network (CNN) to overcome the above-mentioned challenges.

The contributions of this work are two-fold. (1) Exploring spatial and angle features for extracting the spatial and directional coherence through the geometric structure of face regions. (2) Introducing fractional entropy for extracting the texture of facial and background regions, which extracts regular patterns in the images.

3. Proposed Method

We noted from the Introduction and Related Work sections that facial features are important for discriminating between family and non-family persons. As a result, we propose to explore the same for finding facial key points (mouth, nose, left and right eyes and eyebrows) for the input of family and non-family images (Ren et al., 2014). The spatial relationship and angles between facial points provide unique cues for identifying a member of the same family or to distinguish between non-family members. Motivated by this observation, we propose to extract spatial and angle features for facial key points in a new way based on major and minor axes. It is stated in (Wang et al, 2017) that facial appearance in family images has a high degree of similarity with the unique pattern of spatial arrangement of persons (regular patterns), while in the case of non-family, one cannot expect such a high degree of similarity between faces and regular patterns in arranging persons (irregular patterns due to randomness in the ordering of persons). To extract such an observation, we propose to estimate the distance between facial key points with respect to major and minor axes of the respective face images, which gives spatial coherence. In the same way, we also estimate the angle between facial key points of the respective face images, which gives directional coherence. Spatial and directional coherence together extracts geometric properties of face images. However, the geometric features are limited to facial regions. In order to extract regular patterns from both the foreground and background (other than face regions), we further explore fractional entropy which extracts texture properties in the regions. In this way, the proposed method combines the strengths of geometric features and fractional entropy for classifying family and non-family images successfully.

The proposed method extracts 8 distances and 24 angle features using facial key points and two features from fractional entropy for face regions and background information (other than face regions). Therefore, for each input image, it gives a feature vector containing 26 features (8 + 16 + 2). Furthermore, the feature vector is fed to a Convolutional Neural Network (CNN) for classification (McAllister et al., 2016). The overall steps of the proposed method are shown in Fig. 2. In Fig. 2, P_1 to P_{68} are the points given by the face detection method (Ren et al., 2014), and based on those points, the same method detects five facial key points, namely, left Eyebrow (B_1), right Eyebrow (B_2), left Eye (E_1), right Eye (E_2), Nose (N), Mouth (M) and the centroid, using all the 68 points. The distances are estimated between facial key points (d) for each face and finally the proposed method computes the mean of all the 8 features of all the faces (f) in the image (D), which gives a vector of 8 features. Similarly, angles (θ) are estimated between the facial points, and

we compute the mean of all the angles of all the faces in image (γ), which gives a vector of 16 features. For faces and background regions, which are other than face regions, the proposed method extracts fractional entropy for each non-overlapping block (B). The mean (MT) and variance (VT) of the fractional entropy of all the blocks are considered as a feature vector containing features.

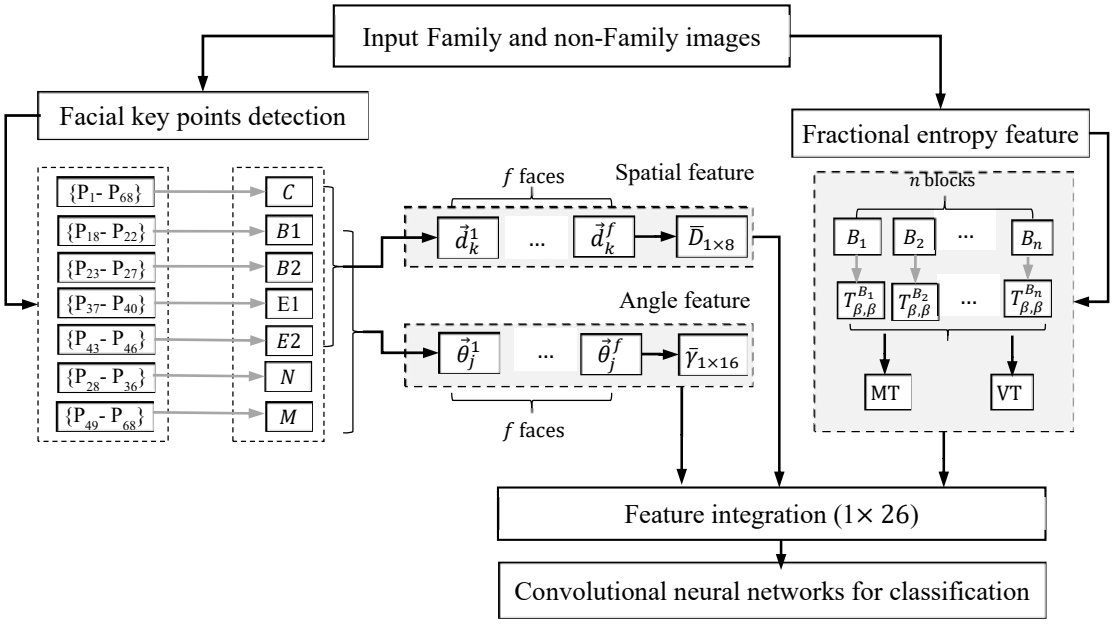
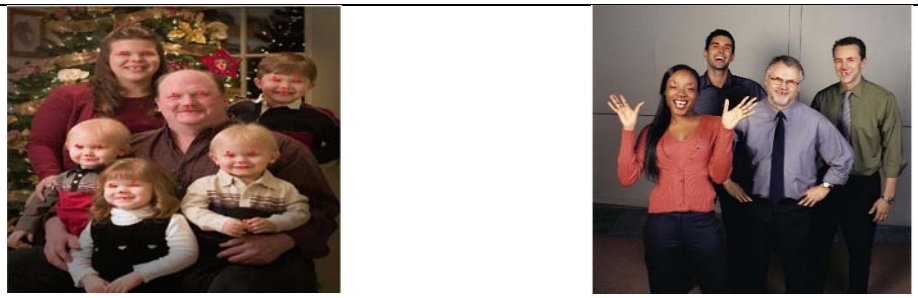
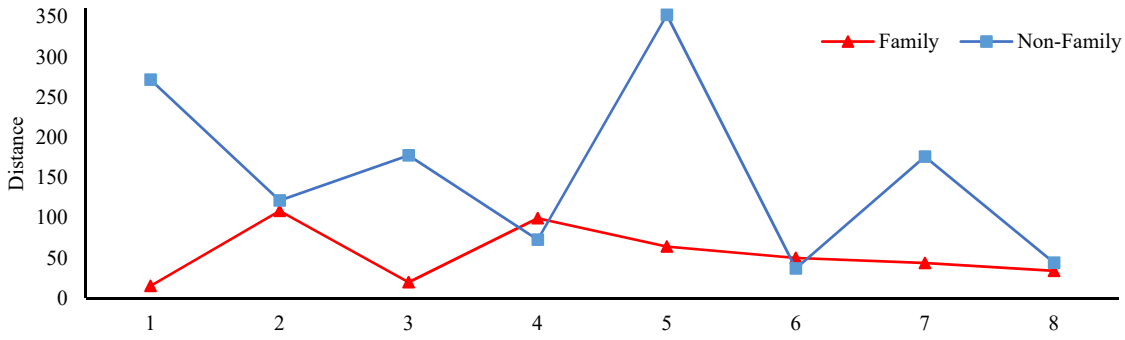


Fig. 2. Flow of the proposed method

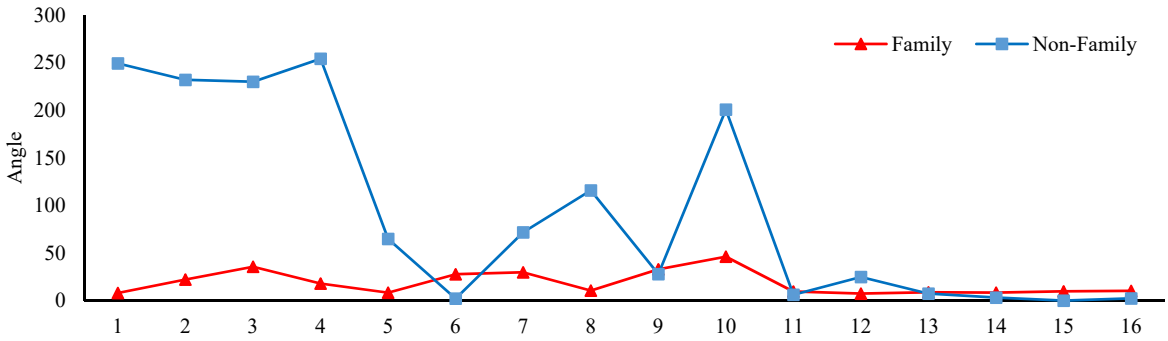
The above observations are illustrated in Fig. 3, where we draw line graphs for distance/angle features *vs* variances of distance/angle values for family and non-family images shown in Fig. 3(a). It is noticed from Fig. 3(b) and Fig. 3(c) that the line behavior which represents families is smoother than that representing non-family for both spatial and angle features. This confirms that the appearance of faces in a family image does not have many variations, while non-family have high variations. The same conclusion can be drawn from the illustration of the angle feature shown in Fig. 3(c). This motivates us to use spatial and angle features for family and non-family image classification. Detailed explanations for each step of the spatial and angle feature extraction process is discussed in subsequent sections.



(a) Example of Family and Non-family images



(b) Variance of 8 distance features (spatial features) for family and non-family images in (a)



(c) Variance of 16 angle features (Angle information) for family and non-family images in (a)

Fig. 3. Cues for extracting geometrical features (Spatial + Angle) for family and non-family photo classification.

3.1. Geometric Features for Facial Key Points

For a given input image, the proposed method uses face alignment via regression of local binary features for detecting facial key points, namely, mouth, noise, left and right eyes, and eyebrows (Ren et al., 2014). The method basically proposes a better learning-based approach. It works based on learning with a "locality" principle. The principle is defined as: for locating a certain landmark at a given stage, the most discriminative texture information lies in a local region around the estimated landmark from the previous stage. Shape context, which gives locations of other landmarks and local textures of this landmark, provides sufficient information. With these observations, the method first learns intrinsic features to encode local textures for each landmark independently; it then performs joint regression to incorporate shape context. The method first learns a local feature mapping function to generate local binary features for each landmark. Here, it uses a standard regression random forest to learn each local mapping function. Then it concatenates all the local features to obtain the mapping functions. It learns linear projections by linear regression. This

learning process is repeated stage by stage in a cascaded fashion. After that, a global feature mapping function, and a global linear projection and objective function are used to incorporate shape context. This process can effectively enforce the global shape constraint to reduce local errors. In the case of the testing phase, shape increment is directly predicted and applied to update the current estimated shape. More details regarding implementation can be found in (Ren et al., 2014). The reason to choose this method is that it is said to be generic, efficient and accurate for finding facial key points. In addition, it can cope with issues of partial occlusion and distortion. This is justifiable because the proposed work considers family and non-family images with complex backgrounds and diversified content. The sample results of the above method are illustrated in Fig. 4, where (a) gives the results of candidate point detection for the input image, while Fig. 4(b) shows samples of facial key points for family and non-family images. It is noted from Fig. 4(b) that although the images are affected by distortion and poor quality, the method finds key points successfully.

(a) Candidate point detection (b) Facial key points of family and non-family images

Fig. 4. Facial key point detection for family and non-family images (Ren et al, 2014).

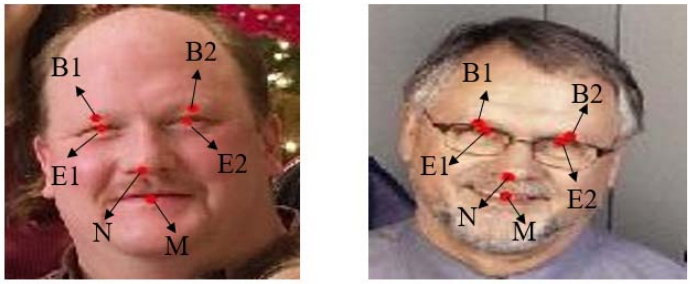


Fig. 5. Labelling six facial key points of the family and non-family images marked in Fig. 4.

Let B_1, B_2, E_1, E_2, N and M be center points given by the method (Ren et al., 2014), which denote left and right eyebrows, eyes, nose and mouth, respectively. These points are marked manually in Fig. 5 for the family and non-family faces chosen from the images shown in Fig. 4. To extract spatial features to study geometric characteristics, the proposed method finds the centroid using candidate points of the face region

as defined in Equation (1), where m is the number of candidate points given by the method (Ren et al., 2014).

$$(X_C, Y_C) = \left(\frac{\sum_{i=1}^{m} X_i}{m}, \frac{\sum_{i=1}^{m} Y_i}{m}\right) \quad (1)$$

With the help of the centroid (X_C, Y_C), the proposed method draws an ellipse to find the major and minor axis as shown in Fig. 6(a) and Fig. 6(b) for family and non-family faces, respectively. The proposed method moves in a perpendicular direction to each key facial point (B_1, B_2, E_1, E_2) of family and non-family images until it reaches pixels of the major axis as shown in the second illustration in Fig. 6(a) and Fig. 6(b), respectively. Similarly, the proposed method moves in a perpendicular direction to each key point of family and non-family images until it reaches pixels of the minor axis as shown in the third illustration in Fig. 6(a) and Fig. 6(b), respectively. Then the method finds the distance between facial key points r={B_1, B_2, E_1, E_2} and respective pixels of the major and minor axes in r'={major, minor} defined in Equation (2), which outputs 8 distances d_k , k={1,2,..,8} for each face i:

$$d_k^i = \sqrt{(X_r - X_{r'})^2 + (Y_r - Y_{r'})^2} \quad (2)$$

The distance features are extracted with respect to the major and minor axes to make the features robust to different rotations. In other words, if the input image is rotated in different directions, the feature still works well. For this step, we consider only four facial key points (that is, B_1, B_2, E_1 and E_2) for distance calculation because Mouth (M) and Nose (N) do not contribute much to classification because the M and N points lie on the minor axis. Note that the perpendicular distance is calculated by finding the smallest distance between facial key points and the pixels of major/minor axes. The step finds many distances by considering a few left and right pixels of major and minor axes to the key points. Then it chooses the pixel which produces the smallest distance between the pixels of the major/minor axis and key points. We believe that the smallest distance is the same as the perpendicular distance. Since the input image contains many faces and the number of faces is not predictable, the proposed method computes the mean of the 8 respective distances d of all faces in the input image as defined in Equation (3), resulting in an average distance vector $\overline{D}$ for each input image, where f is the number of faces:

$$\overline{D} = \frac{\sum_{i=1}^{f} d_k^i}{f} \quad (3)$$

To make the geometric features robust, we also propose to calculate the angles between facial key points to study the structure of the face region. This is because, as the face shape changes, the angle between facial key points also changes. To extract such observations, we construct a rectangle using B_1-B_2-E_1-E_2 as shown in the first image in Fig. 6(c), which gives four angles. In the same way, the proposed method forms triangles using B_1-B_2-N, B_1-B_2-M, E_1-E_2-N, E_1-E_2-M as shown respectively in Fig. 6(c), which gives twelve

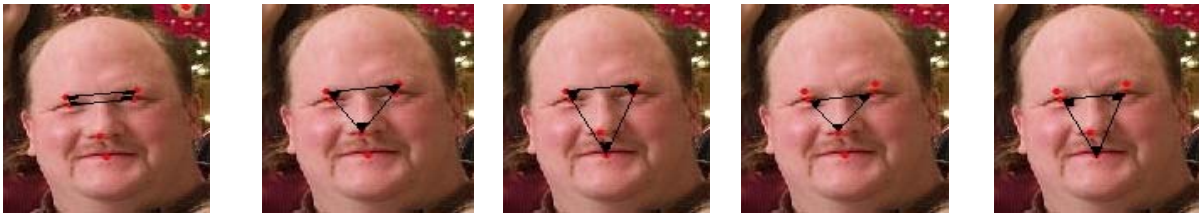
angles. In total, the proposed method obtains 8 spatial + 16 angles = 24 geometric features for family and non-family image classification.

Major and minor axis Distance to major axis Distances to minor axis Feature extraction

(a) Spatial feature extraction for sample family face.

Major and minor axis Distance to major axis Distances to minor axis Feature extraction

(b) Spatial features for sample non-family face.



4-angles for E1, E2, B2, B1 3-angles B1, B2, N 3-angles for B1, B2, M 3-angles for E1, E2, N 3-angles for E1, E2, M

(c) Angle-based features for sample family face.

4-angles for E1, E2, B2, B1 3-angles B1, B2, N 3-angles for B1, B2, M 3-angles for E1, E2, N 3-angles for E1, E2, M

(d) Angle-based features for sample non-family face.

Fig. 6. Geometric features using facial points for family and non-family image discrimination.

Let $A(x_A, y_A)$, $B(x_B, y_B)$, $C(x_C, y_C)$ be the coordinates of the ABC triangle. The inner angles $\hat{B}$ for the ABC triangle can be calculated as defined in Equation (4) and (5). Equation (4) computes a vector between B and A called $\overrightarrow{AB}$ and the vector between C and B similarly called $\overrightarrow{CB}$. Angle θ_B is driven by Equation (5) by computing the four-quadrant inverse tangent, where $\begin{vmatrix} x_{AB} & y_{AB} \\ x_{AB} & y_{AB} \end{vmatrix}$ is determinant, while $\overrightarrow{AB}.\overrightarrow{CB}$ is the scalar dot product of the two vectors. Similarly, the proposed method estimates angles for the rectangle and the other triangles in this work.

$$\overrightarrow{AB} = B - A \quad , \overrightarrow{CB} = C - B \tag{4}$$

$$\theta_B = Arctan2\left(\begin{vmatrix} x_{AB} & y_{AB} \\ x_{AB} & y_{AB} \end{vmatrix}, \overrightarrow{AB}.\overrightarrow{CB}\right) \tag{5}$$

Since we can expect many faces in a single input image, we propose to consider the average of the angles of the respective 16 angles. In order to average the respective angles of *f* faces, the circular mean is computed. First, since the angles $\{\theta_1, \theta_2, \ldots, \theta_j\}$, j=16 are defined on a circular coordinate system, the coordinate system should be changed to a rectangular one according to Equation (6), where θ_j^i is the j^{th} angle θ of the i^{th} face in the image. Afterwards, v_j the resultant vector and its direction are calculated as defined in Equation (7) and Equation (8), respectively. Finally, $\overline{\gamma_j}$ mean of the j^{th} angle for all the *f* faces is computed as defined in Equation (9).

$$\overline{x_j} = \frac{\sum_{i=1}^{f} \cos\theta_j^i}{f} \quad , \overline{y_j} = \frac{\sum_{i=1}^{f} \sin\theta_j^i}{f} \tag{6}$$

$$v_j = \sqrt{\overline{x_j}^2 + \overline{y_j}^2} \tag{7}$$

$$\cos\overline{\gamma_j} = \frac{\overline{x_j}}{v_j} \ , \ \ \sin\overline{\gamma_j} = \frac{\overline{y_j}}{v_j} \tag{8}$$

$$\overline{\gamma_j} = Arctan\left(\frac{\sin\overline{\gamma_j}}{\cos\overline{\gamma_j}}\right) \tag{9}$$

The proposed method computes the mean of distances to extract spatial features and the mean of angles for extracting angle features for each image. The reason for computing the average is to widen the difference between family and non-family images. As discussed in the Introduction Section, family images have persons with almost the same facial appearance, while non-family images have persons with different facial appearances. This is valid because one can expect a high degree of similarity between the appearances of faces from the same family. It may not be true for non-family images. In addition, family and non-family images can have any number of faces, which should be more than 3 persons in the images. In this situation, the average features for a family does not make much difference, while for non-family, the average makes a vast difference. Since the appearance of faces in a family have a high degree of similarity compared to those in non-family images, it is expected that the average gives almost the same values for family images while for non-family, we cannot predict the same values always. Besides, to make the spatial and angle features invariant to the number of faces, the proposed method considers the average for achieving better results.

3.2. Fractional Entropy Feature Extraction

As mentioned in the Introduction Section, it is found that the other than face region also provides cues for discriminating family and non-family images. However, the previous step does not explore other than face

region. Therefore, inspired by the method in (Ibrahim et al., 2015) where fractional calculus has been used for studying texture in splicing images, this section explores a new Tsallis fractional entropy-based texture (Tsallis et al., 2009) for studying variations in background as well as facial regions in family and non-family images. An overview of the Tsallis fractional entropy is presented in the following.

The Tsallis fractional entropy (Tsallis et al., 2009) measures the amount of uncertainty acting in the valuation of a random variable or the consequence of a random process. The general discrete form of this entropy is given in Equation (10).

$$T_{\beta}\ (\rho)(x) = \frac{1}{\beta-1}\left(1 - \sum_i {\rho_i}^{\beta}\ (x)\right), \tag{10}$$

where ρ is the q-Gaussian probability of pixel x, q $\neq$1 and $\beta\neq$1 are the fractional powers of the entropy, and the quantity 1/(β-1) is the capacity of the image. The q-Gaussian is a probability distribution ascending from the growth of the Tsallis entropy under suitable restrictions. It has the formal function as defined in Equation (11) to (13), where C_q is a normalization factor.

$$\rho\ (x) = \frac{\sqrt{\beta}}{\mathrm{C}_q}\ e_q\ (-\beta x^2), \tag{11}$$

$$e_q\ (-\beta x^2) = \ [1 + (1-q)(-\beta x^2)]^{\frac{1}{1-q}}, \qquad q \neq 1, \tag{12}$$

$$\mathrm{C}_q = \frac{\sqrt{\pi}}{\sqrt{q-1}}\ \frac{\Gamma(\frac{3-q}{2(q-1)})}{\Gamma(\frac{1}{q-1})}, \tag{13}$$

Since the variable is the pixel which has a positive value in the maximum entropy procedure, the q-exponential distribution is derived. Applying Equation (11-13) in (10), we have the following generalized formula of the fractional entropy:

$$T_{\beta,q}\ (x) = \frac{\sqrt{\beta}}{\mathrm{C}_q(\beta-1)}\left(1 - \sum_{i=1}^{Z}[1 + (1-q)(-\beta {x_i}^2)]^{\frac{\beta}{1-q}}\right) \tag{14}$$

In our discussion, let β=q, then we conclude

$$T_{\beta,\beta}\ (x) = \frac{\sqrt{\beta}}{\mathrm{C}_{\beta}(\beta-1)}\left(1 - \sum_{i=1}^{Z}[1 + (1-\beta)(-\beta {x_i}^2)]^{\frac{\beta}{1-\beta}}\right) \tag{15}$$

where Z is the total number of pixels in the image. The proposed method calculates the above Tsallis fractional entropy based on frequency details of the input image, which gives a texture property to study the structure of it. The advantage of Tsallis fractional entropy is that it is sensitive to non-textured regions (low frequency). In addition, it sharpens any changes in texture details in the regions, where pixel values are changing sharply (high frequency). The sample illustration for Tsallis fractional entropy for family and non-family images is shown in Fig. 7, where we can see all the dominant information represented by edges in the background and the facial regions are highlighted. Fig. 8 shows the clear discriminating power of Tsallis fractional entropy texture features for family and non-family images. Therefore, for the feature

matrix given by the Tsallis fractional entropy texture, we first split the input image into blocks with a size of $a \times a$ pixels, then the Tsallis fractional entropy for each block is computed. For all the blocks of the input image, the "mean" and the "variance" are computed and saved as the output texture features *MT* and *VT*, respectively. The pseudo-code for the proposed Tsallis fractional entropy algorithm is described as follows:

Algorithm: Fractional Entropy Feature Extraction

Initialization: I=Input image , a=3; β = 0.5
For each Input image I **do**
    $\{B_1, B_2, \ldots, B_n\} \leftarrow$split I into n blocks size of $a \times a$ pixels
        For i=1 to n **do**
        $T_{\beta,\beta}(B^i_{3\times3}) \leftarrow I$ // Fractional entropy is calculated as defined in Equation (15), where i denotes the i^{th} block of 3×3 dimension.
        End For
        $MT \leftarrow$ mean $(T^{B_i}_{\beta,\beta})$, i={1,2,…n}// Mean of Fractional entropy of all (n) blocks
        $VT \leftarrow$variance $(T^{B_i}_{\beta,\beta})$, i={1,2,…n}) //Variance of Fractional entropy of all (n) blocks
End For

Feature distributions of the spatial, angle and texture features for family and non-family images are shown in Fig. 9(a)-Fig. 9(c), respectively, where one can see that the feature distributions of geometric and Tsallis fraction entropy provides a clear distinction between family and non-family images in terms of histogram behavior.

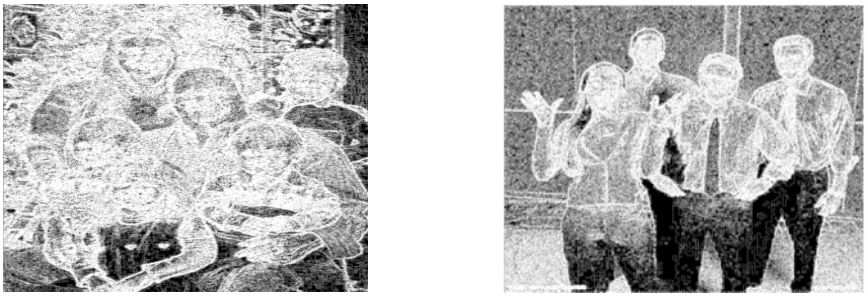


Fig. 7. Fractional entropy features for family and non-family images

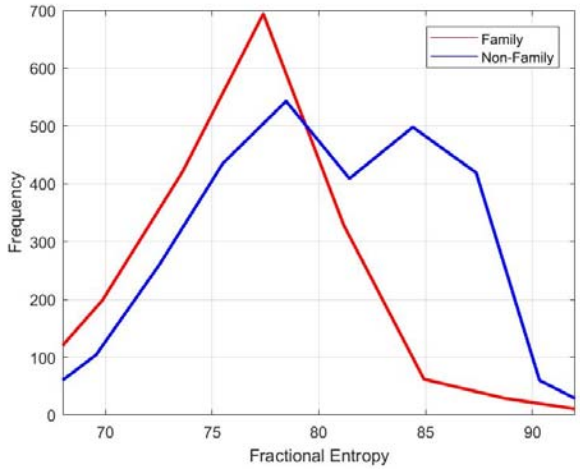
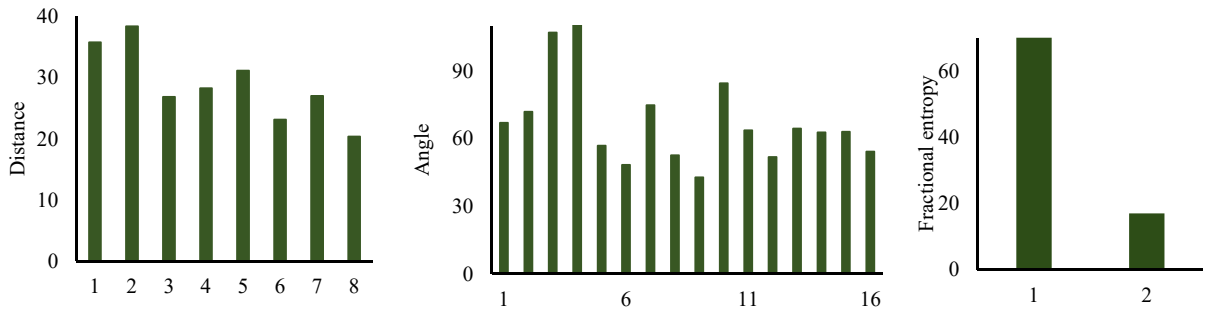
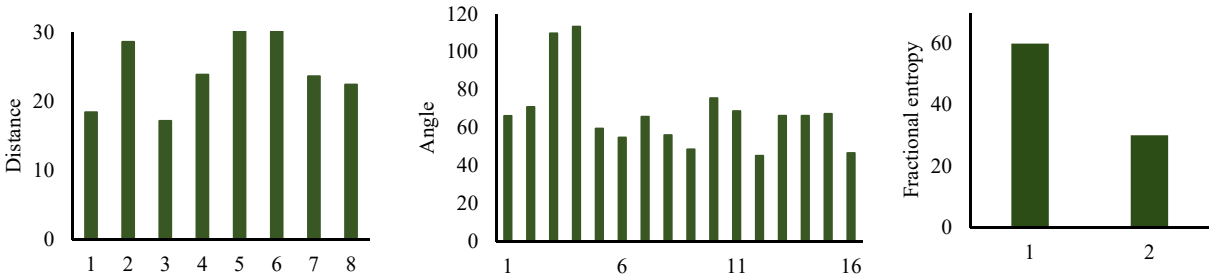


Fig. 8. Histogram of Fractional entropy features for family and non-family images.



(a) Mean of spatial, angle and fractional entropy feature distribution for family images.



(b) Mean of spatial, angle and fractional entropy feature distribution for non-family images.

Fig. 9. The proposed feature distribution for family and non-family images classification.

The concatenated features are then passed to a fully connected Convolutional Neural Network (CNN) for classifying family and non-family images (McAllister et al., 2016). Inspired by the method (Nanni et al. 2018), where it is mentioned that the combination of handcrafted features and the ensemble of CNNs give better results than deep learning tools such as GoogleNet, ResNet50 that use raw pixels of the input images for bioimage classification, we explore the same idea of combining the proposed features with the CNN for

family and non-family image classification in this work. Since the proposed work does not provide a large number of samples for training and labeling samples, we prefer to use the combination of the proposed features and the CNNs rather than raw pixels with the recent deep learning models. The main objective of the proposed work is to propose features that can classify the family and non-family photos. Thus, the proposed features are fed to a pre-defined CNN classifier which is available online (Arora & Suman, 2012) for classification in this work. For learning parameters of the classifier, we follow a 10-fold cross-validation procedure, which splits the dataset into training and testing components. The training samples are used for learning and adjusting the parameters of the classifier and the testing samples are used for evaluation. The complete algorithmic steps of the proposed method for classifying family and non-family images are presented below.

Algorithm: Feature Extraction for the Proposed Method

Initialization: I=Input image , m ={1,…,68}, set of points given by (Ren et al., 2014)
$f \leftarrow$ Number of faces
For i=1 to f **do**
    $(X_{P_m}, Y_{P_m}) \leftarrow$ facial points
    $\{B_1, B_2, E_1, E_2, N, M, C\} \leftarrow$ Key facial points, which includes eyebrows, eyes, nose, mouth and center of the all the key points, respectively.
    $(A_{major}\ ,\ A_{minor}) \leftarrow (x_c, y_c)$ as defined in Equation (1)
    $\vec{d}_k^i \leftarrow \{B_1, B_2, E_1, E_2\}$ &$\{A_{major}\ ,\ A_{minor}\}$:$\vec{d}_{(1\times 8)}^i$ as defined in Equation (2)
    $\vec{\theta}_j^i \leftarrow \{B_1, B_2, E_1, E_2, N, M\}$:$\vec{\theta}_{(1\times 16)}^i$ as defined in Equation (4) and Equation (5)
End For
$\bar{D}_k \leftarrow$mean $(\vec{d}_k)$: $\vec{\boldsymbol{D}}_{(\mathbf{1\times 8})}$ as defined in Equation (3)
$\bar{\gamma}_j \leftarrow$ circular mean$(\vec{\theta}_j)$: $\vec{\boldsymbol{\gamma}}_{(\mathbf{1\times 16})}$ as defined is Equation (6)-Equation (9)
$MT,\ VT \leftarrow$Fractional Entropy feature extraction as defined in the above algorithm
$\overrightarrow{feature}_{(1\times 26)} \leftarrow \bar{D} \| \bar{\gamma} \| MT \| VT$// Final feature vector having dimension, 1×26.
CNN classification //Classification of Family and Non-family photos.

4. Experimental Results

For experimentation, we created our own dataset by collecting images from social media, such as Facebook, Flickr, Instagram and from our own camera. This dataset includes indoor/outdoor scenes and images with 3-25 people. In addition, the dataset involves family and non-family photos of different cultures, such as Hindu and Chinese, and modern styles of family/non-family photos. This makes the dataset challenging for experimentation. For labeling the data as either family or non-family, we followed the instructions suggested in (Wang et al., 2017; Wang et al., 2015; Gallagher et al., 2009). Furthermore, the dataset includes one photo for one family. In other words, the dataset does not have multiple photos of the same family. In total, our dataset consists of 388 family images and 382 non-family images, which gives a total of 770 images.

To demonstrate that the proposed method is effective, we also considered the benchmark dataset collected from publicly available data in (Wang et al., 2017; Wang et al., 2015; Gallagher et al., 2009). This public

data provides a large number of images, which include many groups of photos and images containing both family and non-family categories. As a result, we chose the relevant family and non-family images and labelled these manually according to the instructions in (Wang et al., 2017; Wang et al., 2015; Gallagher et al., 2009). We consider this dataset as the benchmark dataset, which consists of 1790 family and 2753 non-family images. In total, there are 4543 images, which is larger than the dataset considered in (Mehta et al., 2018; Haghighat et al., 2015). Overall, we considered 5263 (770 from our dataset and 4543 from benchmark dataset) images for experimentation in this work. Sample images of family and non-family photos for ours and the benchmark dataset are shown in Fig. 10(a) and Fig. 10(b), respectively, where we can see intra- and inter-class variations. It is also observed from Fig. 10 that family and non-family images have both indoor and outdoor scenes as backgrounds. It is also true that height distribution of persons in a hierarchical order for family and a non-hierarchical order for non-family is not necessarily true as shown in Fig. 10. The detailed statistics of ours and the benchmark dataset are listed in Table 1, where we calculate the ratios (E_1 and E_2) as respectively defined in Equation (16) and Equation (17) using the count images with indoor and outdoor scenes, and hierarchical or non-hierarchical persons' height orders. The ratio in Table 1 indicates that our dataset is much more complex than the benchmark dataset because the ratio with respect to indoor backgrounds and the hierarchical order of our dataset are greater than those of the benchmark dataset. Note that in Equation (16) and Equation (17), *total* denotes the size of the dataset as given in Table 1 in the bracket.

$$E_1 = \frac{Outdoor(family)+Indoor(non-family)}{total} \tag{16}$$

$$E_2 = \frac{Non-Hierarchical(family)+Hierarchical(non-family)}{total} \tag{17}$$

Table 1. Statistics for ours and the benchmark dataset for family and non-family image classification

Dataset (Total)	Family				Non-Family				E_1	E_2
	Indoor	Outdoor	Hierarchical	Non-Hierarchical	Indoor	Outdoor	Hierarchical	Non-Hierarchical		
Our (770)	255	133	273	115	201	182	110	272	43.37	29.22
Benchmark (4543)	1172	618	1378	412	924	1829	513	2240	33.94	20.36

(a) Sample images for family and non-family photos from our dataset

(b) Sample images for family and non-family photos from the benchmark dataset

Fig. 10. Sample images of our dataset and the benchmark dataset (Gallagher et al, 2009)

To show that the proposed method is superior in comparison to existing methods, we implemented two state-of-the-art methods, namely, (Wang et al., 2015), which explores facial geometric features and facial appearance model-based features. The features are passed to an SVM classifier for family and non-family image classification. Please note, the same idea is extended and the results are improved in (Wang et al., 2017) for the purpose of family and non-family image classification. However, both the ideas focused only on facial regions for achieving results; these also ignored background clues.

To measure the performance of the proposed and existing methods, we generate confusion matrices for family and non-family classification and the classification rate. The Classification Rate (CR) is defined as the number of images classified correctly by the proposed method (R) divided by the total number of images in the class (MG) as defined in Equation (18). The Average Classification Rate (ACR) is calculated for diagonal elements of the confusion matrices to evaluate the overall performance of the proposed and existing methods.

In this work, we undertake 10-fold cross-validation for choosing the number of training and testing samples. The criteria divides the whole dataset into 10 equal-sized sub-folds. For each iteration, images from each sub-fold are considered as testing samples, while images from the other sub-folds are considered as training samples for classification, which results in a confusion matrix for one sub-testing fold out of 10 sub-folds. This process indicates that the chosen training samples are used for training the classifier and the testing samples are used for evaluation. In this way, the process considers every sub-fold as testing samples at each iteration, which results in 10 confusion matrices i.e. 10-fold. The average of all the 10 confusion matrices are considered as the final confusion matrix for evaluation in this work.

$$CR = \frac{R}{MG} \tag{18}$$

4.1. Evaluating the Proposed Classification

The proposed method consists of three key steps, namely, extracting spatial/angle-based geometric features and fractional entropy-based texture features for classifying family and non-family images. In order to assess the contribution of each key step, we conducted experiments on both our dataset and the benchmark dataset individually to calculate average classification rates. The results reported in Table 2 show that the combined Spatial + Angle achieves the best results compared to the individual features for both our dataset and the benchmark dataset. It is also noted from Table 2 that the ACR of angle-based features is better than Spatial, but lower than Spatial + Angle for both the two datasets. This shows that angle-based features are better than spatial-based features, and the combination is better than both individual features. This is understandable because the spatial structure alone is not sufficient for handling the problem of complex backgrounds as it only extracts 8 features. However, the improvement is marginally different. Therefore, we can conclude that spatial and angle-based features contribute equally for achieving the best results.

Table 2. Confusion matrices of spatial, angle and spatial + angle on ours and the benchmark dataset in (%)

Classes	Spatial				Angle				Spatial + Angle			
	Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	66.5	33.50	32.62	67.37	74.22	25.77	39.40	60.59	76.80	23.2	43.92	56.07
non-Family	35.07	64.92	39.88	60.11	21.46	78.53	33.92	66.12	20.94	79.05	34.96	65.08
ACR	65.70		46.36		76.37		52.76		77.92		54.49	

In this work, we extracted fractional entropy-based texture features for the whole image, which includes facial regions and background information. We conducted experiments for calculating classification rates only for the Facial region (FEF), Background (FEB), and the whole image (FEW) to identify the effectiveness of the facial region and background information, individually. Note: the facial regions detected by facial point detection are considered as foreground, and the rest of the region is considered as the background for experimentation. The results of FEF, FEB and FEW are reported for both our dataset and the benchmark dataset in Table 3. It is observed from Table 3 that the FEF is the best at ACR compared to FEB for both the datasets. This shows that facial regions contribute more compared to the background. This is justifiable because sometimes, family and non-family photos may share the properties of the background. It is evident from the results of FEB for the benchmark dataset in Table 3, where most family images are misclassified as non-family. This shows that the features of the background of family images overlap with the features of the background of non-family images. However, facial regions alone are not

sufficient to achieve the best ACR compared to FEW. Therefore, we can conclude that the features of the foreground and background are important to achieve the best results for classification.

Table 3. Confusion matrices of FEF, FEB and FEF+FEB on ours and the benchmark dataset in (%)

Classes	FEF				FEB				FEW (FEF + FEB)			
	Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	75.51	24.48	50.09	49.01	64.43	35.56	4.99	95.33	80.92	19.07	54.65	45.35
non-Family	28.53	71.46	14.67	85.52	44.24	55.75	5.43	94.61	12.82	87.17	12.71	87.29
ACR	73.48		67.80		60.09		49.63		84.04		70.97	

It is noted from Table 2 and Table 3 that Spatial + Angle and FEW are better compared to individual features on both our dataset and the benchmark datasets for family and non-family image classification. In order to decide the best combination, we conducted experiments for the following combinations: Spatial + FEF, Spatial + FEB, Spatial + FEW, as reported in Table 4 and Angle + FEF, Angle + FEB and Angle + FEW, as reported in Table 5. When we look at the ACR of all the combinations in Table 4 and Table 5, Spatial + FEW and Angle + FEW are the best compared to the other combinations for both our dataset and the benchmark dataset. It is justifiable because Spatial + FEW and Angle + FEW include features of facial regions and background information. Therefore, to achieve the best results, we propose the combination of Spatial + FEW and Angle + FEW, which is the proposed method and the results are reported in Table 6 for our dataset and the benchmark dataset.

When we compare ACR of Spatial + FEW and Angle + FEW with the results of the proposed method (Spatial + Angle + FEW), ACRs for the respective three experiments on our dataset are almost the same. This is because the proposed method has been developed based on our dataset. However, when we compare the ACR of Spatial + FEW, Angle + FEW, and the proposed method on the benchmark dataset, there is a significant improvement for the proposed method compared to Spatial + FEW and Angle + FEW. Hence, we can conclude that the proposed method is capable of handling complex datasets. It is observed from the ACR of the proposed method on our dataset and the benchmark dataset reported in Table 6 that the proposed method scores highly on the benchmark dataset compared to our dataset. The reason is that our dataset includes diverse images such as those of different culture, modern families and non-family photos. At the same time, the benchmark dataset provides a large number of images for training, i.e., 1790 for family and 2753 for non-family compared to 388 for family and 382 for non-family images of our dataset. This is the advantage of the benchmark dataset for achieving the best results compared to ours. This is because when we feed a large number of training samples to the classifier, it covers more possible variations in images.

Therefore, a large number of training samples and more variations led to achieving the best results for the benchmark dataset by the proposed method compared to our dataset.

Table 4. Confusion matrices of Spatial + FEF, Spatial + FEB and Spatial + FEW on ours and the benchmark dataset in (%).

Classes	Spatial + FEF				Spatial + FEB				Spatial + FEW			
	Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	73.71	26.28	33.68	66.31	62.62	37.38	37.35	62.57	85.56	14.44	71.04	28.88
non-Family	27.74	72.25	16.48	83.51	44.50	55.49	29.12	70.92	14.39	85.60	15.94	84.10
ACR	72.98		58.59		59.42		54.13		85.58		77.57	

Table 5. Confusion matrices of Angle + FEF, Angle + FEB and Angle + FEW on ours and the benchmark dataset in (%).

Classes	Angle + FEF				Angle + FEB				Angle + FEW			
	Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	84.02	15.97	43.64	56.35	74.48	25.52	45.83	54.16	83.24	16.75	67.86	32.13
non-Family	19.10	80.89	14.26	85.73	26.17	73.82	14.22	85.52	11.78	88.21	19.80	80.20
ACR	82.46		64.68		74.15		65.67		85.72		74.03	

In the case of spatial and angle feature extraction discussed in the Proposed Methodology Section, the proposed method computes the mean for distances and angles of all the faces in the respective images separately. To assess the influence of averaging (the mean), we conduct experiments for calculating the classification rate using the proposed method without averaging. In other words, the proposed method considers all the distance and angle features of faces in images as distance and angle feature vectors respectively for classification. The results are reported in Table 6, where one can see the proposed method without averaging the scores providing very poor results compared to the proposed method with averaging for both the datasets. This shows that the operation, namely, averaging, plays a vital role in achieving better results for family and non-family classification.

Since we use the CNN for classification, to show its effectiveness compared to the SVM and the use of raw pixels with the CNN, the proposed method is used for experimentation of the proposed features with an SVM as well as feeding raw pixels to a CNN for ours and the benchmark dataset. For experiments using raw pixels of the images, the proposed method considers each pixel value as a feature and it converts a two-

dimensional image matrix to single-dimensional feature vector in a row-wise fashion. The converted single-dimensional feature vector is passed to a CNN for classification. This experiment does not involve the proposed distance, angle and fractional entropy-based features for calculating the measures. The results are reported in Table 6. It is noted from Table 6 that the results of feeding raw pixels directly to a CNN performs poorly in terms of classification rate compared to the proposed features with a CNN. The main reason for the poor results is that since the number of samples for the training set is small, it may not cover all the possible variations of images when we feed raw pixels to the CNN directly. For experimentation with an SVM and for a fair comparative study with the CNN, the proposed method uses a polynomial kernel as it is non-linear like the CNN classifier. When we compare the results of the proposed features with an SVM and the proposed features with a CNN, the proposed features with an SVM achieve poorer results compared to the proposed features with a CNN. It is justifiable because the SVM does not have a generalization ability as is the case with a CNN. In addition, the performance of the SVM depends on the kernel type and size. On the other hand, the CNN can learn complex non-linear input and output relationships. Therefore, for the proposed problem, which is complex in terms of foreground and background variations according to the statistics reported in Table 1, the proposed features with a CNN perform better than the proposed features with an SVM.

Table 6. Confusion matrix and classification rate of the proposed method (spatial + angle + FEW) without an averaging operation, with CNNs, SVMs and CNN on raw pixels in the images on ours and the benchmark dataset (in %).

Classes	Proposed Method (with averaging)				Proposed without averaging				Proposed method CNN on raw pixels				Proposed method with SVM			
	Our		Benchmark		Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	88.40	11.59	96.36	3.63	77.31	22.68	48.92	51.08	62.62	37.37	91.58	8.42	80.87	19.13	91.67	8.32
non-Family	15.70	84.29	1.16	98.83	80.36	19.63	19.22	80.78	56.70	43.29	88.84	11.16	12.56	87.43	5.70	94.29
ACR	86.34		97.59		49.86		64.85		52.59		53.53		84.13		93.26	

We also report the results of two existing methods on our dataset and the benchmark datasets in Table 7. Since (Wang et al., 2017) is the improved version of (Wang et al., 2015), whereby Wang et al. (2017) gives better results in terms of ACR. When we compare the ACR of the proposed method with two existing methods, the proposed method gives better results than (Wang et al., 2015) and (Want et al., 2017). This is understandable as both the existing methods use only facial regions for classification, while the proposed method uses both facial and background regions. In addition, the proposed method extracts geometric features based on spatial and angle information, and the new fractional entropy feature are an enhancement on existing methods and hence it makes a positive difference.

Table 7. Confusion matrix of the proposed (spatial + angle + FEW) and existing methods on our dataset and the benchmark dataset in (%).

Classes	Proposed				Wang et al. (2015)				Wang et al. (2017)			
	Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	88.40	11.59	96.36	3.63	76.28	23.71	88.54	11.45	82.73	17.27	94.52	5.47
non-Family	15.70	84.29	1.16	98.83	37.43	62.56	24.95	75.04	22.51	77.49	14.38	85.61
ACR	86.34		97.59		69.42		81.79		80.11		90.06	

Sample qualitative results of the proposed method on our dataset and the benchmark dataset are shown in Fig. 11. Fig. 11 also includes the results of misclassifications by the proposed method on our dataset and the benchmark dataset. The reason for misclassification is that when the images of family and non-family images share geometric structures of the faces and the properties of backgrounds, the proposed method fails to perform correct classification. Therefore, there is scope for improvement in the future.

Sample family images from our dataset and the benchmark dataset classified successfully

Sample family images from our dataset and the benchmark dataset classified incorrectly

Sample non-family images from our dataset and the benchmark dataset classified successfully

Sample non-family images from our dataset and the benchmark dataset classified incorrectly

Fig. 11. Qualitative results of successful and unsuccessful classification employing the proposed method on our dataset and the benchmark dataset

The existing methods (Wang et al., 2015, 2017) that work based on the fact that the height distributions of persons in images should satisfy a hierarchical order for family, while it does not for non-family. In the same way, according to the statistics in Table 1, the benchmark dataset contains more images with indoor scenes for the family class, and more images with outdoor scenes for the non-family class. However, the proposed method does not consider these two constraints for the classification of family and non-family images. It is evident from the statistics reported in Table 1 for our dataset, where it can be seen that the ratio of hierarchical to the total number of family images and non-hierarchical to the total number of non-family images is greater compared to that from the benchmark dataset. The same thing is true for images with indoor scenes for family and outdoor scenes for non-family images. To validate the statement, we conducted

experiments for Family-Hierarchical *vs.* Non-family-Non-hierarchical and Family-Non-hierarchical *vs.* Non-family-Hierarchical on both ours and the benchmark dataset, and the results are reported in Table 8. It is observed from Table 8 that the classification rate for the expected order is higher than that of the other order. Therefore, we can conclude that there is not much influence on the overall performance of the proposed method. Similarly, images with indoor and outdoor scenes do not have much of an effect on the overall performance of the proposed method. It is evident from the results reported in Table 8 for Family-Indoor *vs.* Non-family-Outdoor and vice versa. In summary, for all the experiments listed in Table 8 for both ours and the benchmark dataset, the proposed method achieves almost consistent average classification rates. This demonstrates that the proposed method works well irrespective of the background complexities and hierarchical distribution of heights. However, when we compare the results of the proposed method on the whole dataset (Table 7) without separation and the results in Table 8, the results of the proposed method in Table 7 are higher than those in Table 8 due to fewer training samples which represent the variations in the case of individual experiments listed in Table 8.

Table 8. Confusion matrices and classification rate for family and non-family images with different foreground and background patterns on ours and the bechmark dataset (in %).

Classes	Family-Hierarchical vs Non-family Non-Hierarchical				Family-Non-Hierarchical vs Family Hierarchical				Family-Indoor vs Non-family-Outdoor				Family-Outdoor vs Non-family-Indoor			
	Our		Benchmark		Our		Benchmark		Our		Benchmark		Our		Benchmark	
	Family	non-Family	Family	non-Family	Family	Non-family	Family	Non-Family	Family	Non-Family	Family	Non-Family	Family	Non-Family	Family	Non-Family
Family	72.0	27.0	77.0	23.0	66.0	34.0	73.0	27.0	68.0	31.0	76.0	24.0	57.0	42.0	66.0	34.0
non-Family	31.0	69.0	33.0	67.0	35.0	64.0	43.0	57.0	29.0	71.0	23.0	77.0	26.0	74.0	27.0	73.0
ACR	70.5		72.0		65.0		65.0		69.5		76.5		65.5		69.5	

5. Conclusions and Future Work

In this paper, we have proposed a new idea for classifying family and non-family photos by combining facial structure and background texture. The proposed method explores distances between facial key points for extracting spatial features. In addition, angles between facial key points are also explored for studying the structures of faces, which are called geometric features. To make use of the background information and textural properties of facial regions, we have proposed novel Tsallis fractional entropy-based features. Furthermore, the proposed method combines spatial, angle and fractional entropy features to obtain the feature vector. The feature vector is applied to a conventional convolutional neural network for classification. Experimental results on our own dataset and the benchmark datasets show that the proposed method is better than two state-of-the-art methods in terms of average classification rate.

The main contributions are the following. It is inherent that facial regions are the key factor for family and non-family photo image classification. Based on this observation, we explore distance features for facial

key points as spatial features to study the structure of facial regions. We have used angle information for facial key points to make spatial features robust to extract the detailed structure of facial regions. The way we combine spatial and angle-based features as geometric ones is novel and an interesting approach to tackle the issues of family and non-family photo classification. To extract regular patterns in facial and background regions (other than facial region), we propose a novel idea of introducing Tsallis fractional entropy for extracting texture properties of facial regions and other background regions. Furthermore, the proposed method combines geometric and fractional entropy features in a different way for achieving the best results.

Despite having proposed a new idea for family and non-family images classification, there are some limitations to the proposed approach. Sometimes, when family and non-family image share the same properties of facial regions with the background, the proposed method fails to yield good results. This is understandable because one can expect similar patterns of foreground and background for both family and non-family images. In this case, we need a method, which can work irrespective of background and facial regions. One way is to introduce context features using foreground and background information to find a solution regarding context, which can be independent of facial regions and the background.

When photos contain both family and non-family members, the proposed method may not work well. It is beyond the scope of the proposed work as it is hard to separate family or non-family members in the same image. To find a solution, one possible way is to bring multimodal concepts, such as face, skin, dress, and structure of the body. This is due to the potential of sharing personal traits and habits with members belonging to the same family. If individuals do not belong to a particular family, we can expect different habits, structures (apart from the face), skin, etc.

In summary, this paper presents a new idea for finding a solution to family and non-family photo classification. The proposed work demonstrates a promising direction for solving a number of issues, including human trafficking. There are several potential concepts, which can be considered as new research directions for future study. In order to support reproducible research, the dataset and code will made available to readers upon request.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant 61672273 and Grant 61832008, and the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021.

References

Arora, R and Suman, (2012), Comparative analysis of classification algorithms on different datasets using WEKA, International Journal of Computer Applications, 54, pp 21-25. (https://github.com/amten/NeuralNetwork. Accessed 15 Sep 2016)

Cai, G., Hio, C., Bermingham, L., Lee, K. and Lee, I., (2014), Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method, Expert Systems with Applications, 41, pp 3514-3526.

Dai, Q., Carr, P., Sigal, L. and Hoiem, D., (2015), Family member identification from photo collections, In Proc. WCACV, pp 982-989.

Dandekar, A. R. and Nimbarte, M. S., (2014), A survey: Verification of family relationship from parents and child facial images, In Proc. SCEECS.

Dandekar, A. R. and Nimbarte, M. S., (2014), Verification of family relation from parents and child facial images, In Proc. INPAC, pp 157-162.

Gallagher, A. C. and Chen, T., (2009), Understanding images of groups of people. In Proc. CVPR, pp 256-263.

Haghighat, M., Zonouz, S. and Mottaleb, M. A., (2015), CloudID: Trustworthy cloud-based and cross-enterprise biometric identification, Expert Systems with Applications, 42, pp 7905-7916.

Ibrahim, R. W., Moghaddasi, Z., Hamid. A. J., Noor. R. M., (2015), Fractional differential texture descriptors based on the machado entropy for image splicing detection. Entropy17(7), pp 4775-4785.

McAllister, P., Zheng, H., Bond, R. and Moorhead, A., (2016). Towards Personalized Training of Machine Learning Algorithms for Food Image Classification Using a Smartphone Camera. In Proc. ICUCAI, pp 178-190.

Mehta, J., Ramnani, E. and Singh, S., (2018), Face detection and tagging using deep learning, In Proc. ICCCCSP.

Nanni, L., Chidoni, S and Brahnam, S., (2018), Ensemble of convolutional neural networks for bioimage classification, Applied Computing and Informatics.

Ng, W. W. Y., Zheng, T. M., Chan, P. P. K. and Yeung, D. S., (2011), Social relationship discovery and face annotations in personal photo collection, In Proc. ICMLC, pp 631-637.

Qin, X. ,Tan, X. and Chen, S., (2015), Tri-subject kinship verification: Understanding the core of a family, IEEE Trans, Multimedia, 17, pp 1855-1867.

Ren, S., Cao, X., Wei, Y. and Sun. J., (2014), Face alignment at 3000 fps via regressing local binary features. In Proc. CVPR, pp 1685-1692.

Robinson, J. P., Shao, M., Wu, Y., Gillis, T. and Fu, Y., (2018), Visual kinship recognition of families in the wild, IEEE Trans. PAMI, 40, pp 2624-2837.

Shen, C. T., Liu, J. C., Shih, S. W. and Hong, J. S., (2009), Towards intelligent photo composition-automatic detection of unintentional dissection lines in environmental portrait photos, Expert Systems with Applications, 36, pp 9024-9030.

Tsallis, C., (2009), Introduction to nonextensive statistical mechanics. Springer, Science + Business Media.

Wang, X., Guo, G., Merler, M., Codella. N. C., Rohith, M., Smith, J. R. and Kambhamettu, C., (2017), Leveraging multiple cues for recognizing family photos. Image and Vision Computing, 58, pp 61-75.

Wang, X., Guo, G., Rohith, M. and Kambhamettu. C., (2015), Leveraging geometry and appearance cues for recognizing family photos. In Proc. ICWAFG, pp 1-8.

Xia, S., Pan, H. and Qin, A. K., (2014), Face clustering in photo album, In Proc. ICPR, pp 2844-2848.

Zhen, L., Caiming, Z. and Caixian, C., (2018), MMDF-LDA: An improved multi-modal latent dirichlet allocation model for social image annotations, Expert Systems with Applications, 104, pp 168-184.

Detailed Responses to Reviewers' Comments

Title: A Geometric and Fractional Entropy-based Method for Family Photo Classification (ESWA-D-18-05162R1)

Reviewer#1

Comment 1.1: Thank you to the Authors for the work and changes. The manuscript has certainly gained in quality, many things are now presented more clearly and accurately.

Response 1.1: Thank you very much for appreciating the work in terms of quality and clarity.

Comment 1.2: However, I am not convinced if I can recommend this work for publication. The most important thing that I miss is an accurate description of the classification model. The mere mention that this is a convolutional network is not enough. In the previous review I pointed to a strange description of the architecture of this network. As a result, in the current version, this description simply disappeared and has not been improved. Indication that a traditional, pre-trained network is used (how was it pre-trained?) is not enough. This is particularly important in this case, where (as I understand it) the network uses as inputs both raw pixels as well as the vector of features calculated by the method proposed by the Authors. How does the network use inputs of different types (2D pixel data and 1D feature vector)?

Response 1.2: We apologize for misunderstanding the comments. In the proposed work, for classification, the extracted feature vector of size, 1×28 is fed to a pre-defined convolutional network classifier, which is available online (https://github.com/amten/NeuralNetwork. Accessed 15 Sep 2016). For training and testing, we use 10-fold cross validation, which automatically splits the data into training and testing

components. The number of training samples are used for learning the parameters while the testing samples are used for calculating measures.

Regarding the experiments on raw pixels reported in Table 6, this experiment is independent, which does not involve any feature vectors extracted by the proposed method. Since there was a suggestion by the reviewer previously, we have considered each pixel value in the input image as a feature and converted the whole 2D image matrix into a single-dimensional feature vector (column wise) before feeding it to the CNN for classification. This experiment is performed to in order to compare with the proposed features. In order to avoid confusion, we have added more details about the experiments on the raw pixels in Section 4.1 in the revised manuscript.

Comment 1.3: In addition, it is not clear to me why the Authors use the SVM linear classifier? What's more, SVM is presented as a strictly linear model. I have the impression that the Authors do not know that there are also SVM models that are non-linear models. Example, it is justifiable because the SVM is a linear classifier, which is good for simple classification problems". Therefore, using only the linear SVM model for comparison and indicating that it is worse than a multi-layer neural network is unconvincing.

Response 1.3: Thank you very much for noticing the mistake. As per your earlier suggestion for comparing the proposed features-SVM with the proposed features-CNN, we had used a linear SVM model for experiments. As you pointed out, this is not a fair comparative study because CNNs have to be compared with non-linear SVM models for experimentation. Therefore, we have run again the proposed method with an SVM by considering a non-linear polynomial kernel. The new results are reported in Table 6. It is observed from Table 6 that when we compare the results of the proposed method with a non-linear SVM, the SVM method obtains lower results compared to the proposed method using the CNN. The main reason is that the CNN has the ability to solve complex classification problems because it involves generalization ability through weighted calculations. On the other hand, the SVM does not have the same level of generalization ability compared to the CNN as it depends on the kernel type and size. Therefore, we can conclude that for the proposed classification, the proposed feature with a CNN is better than the proposed features with an SVM. We have included the new results for the SVM with a suitable discussion in Section 4.1 in the revised manuscript.

Comment 1.4: Therefore, we use the predefined trained CNN with default parameters and architecture available to the public in WEKA" - more description is needed; how do you include your feature vector into the pre-trained model? How was it pre-trained? On what data?

Response 1.4: Thanks for your comment. As mentioned in the Response to Comment 1.2, in this work, since we have used a pre-defined CNN, which is available online for classification, we follow 10-fold cross-

validation for calculating the measures. The pre-defined CNN network considers training samples given by 10-fold cross validation for learning the parameters, and testing samples for classification. Therefore, we have passed the extracted feature vector given in the proposed method to the CNN for classification. We have included more details to avoid confusion in Section 4.1 in the revised manuscript.

Comment 1.5: "The feature vector is applied to a conventional convolutional neural network for classification." - How? What are the inputs organized with pixels?

Response 1.5: Thanks for your comment. We request that you refer to the Response to Comment 1.2 and 1.3.

Comment 1.6: In the description of the cross validation, the Authors name the individual parts of data, subclasses. According to the accepted terminology, they should be referred to as folds. Also, the response 1.5 is not really an answer for Comment 1.5.

Response 1.6: We apologize for the typo error. We have corrected the mistakes in the revised manuscript. For the comment 1.5 (originally), as per our knowledge and dataset creation, we have collected one photo for one family. As a result, the dataset does not include multiple photos of the same family. We believe this is true for the benchmark dataset also. In the case that the dataset contains a few photos of the same family for both training and testing by mistake, it may not affect the overall performance of the proposed method much because it is very rare to have multiple photos of the same family. We have added a statement in Section 4 in the revised manuscript for the purpose of improving the clarity of the dataset creation.

Comment 1.7: Before eq. 4: "where … is determined" - determinant? Section 4: "This dataset includes indoor/outdoor scenes and images with 2-25 people" - in the introduction you defined family/non-family images as containing at least 3 persons. "The criteria divides the whole dataset into 10 equal-sized sub-classes." - … 10 equal-sized folds

Response 1.7: We appreciate your observation and effort. We have taken care of such mistakes in the revised manuscript.