Like Father, Like Son: Facial Expression Dynamics for Kinship Verification

Hamdi Dibeklioğlu[1,2], Albert Ali Salah[3], and Theo Gevers[1]
[1]Intelligent Systems Lab Amsterdam, University of Amsterdam, Amsterdam, The Netherlands
[2]Pattern Recognition & Bioinformatics Group, Delft University of Technology, Delft, The Netherlands
[3]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
h.dibeklioglu@tudelft.nl, salah@boun.edu.tr, th.gevers@uva.nl

Abstract

Kinship verification from facial appearance is a difficult problem. This paper explores the possibility of employing facial expression dynamics in this problem. By using features that describe facial dynamics and spatio-temporal appearance over smile expressions, we show that it is possible to improve the state of the art in this problem, and verify that it is indeed possible to recognize kinship by resemblance of facial expressions. The proposed method is tested on different kin relationships. On the average, 72.89% *verification accuracy is achieved on spontaneous smiles.*

1. Introduction

Automatic detection of kinship from facial appearance is a difficult problem with several applications, including social media analysis [20, 21], finding missing children and children adoptions [9], and coaching for imitation and personification. Kinship is a genetic relationship between two family members, including parent-child, sibling-sibling, and grandparent-grandchild relations. Since a genetic test may not always be available for checking kinship, an unobtrusive and rapid computer vision solution is potentially very useful. This paper proposes such a novel approach for kinship detection.

Kinship may be verified between people that have different sex and different ages (e.g. father-daughter), which makes this problem especially challenging. Humans use an aggregate of different features to judge kinship from facial images [1]. Furthermore, depending on the age of the person assessed for kinship, humans use different sets of features consistent with the expected aging-related form changes in faces. For example, upper face cues are more prominently used for kids, as the lower face does not fully form until adulthood [13]. Automatic kinship detection methods also employ aggregate sets of features including color, geometry, and appearance. In Section 2 we summarize the recent related work in this area.

All the methods proposed so far to verify kinship work with images. In contrast to all published material, in this paper, we propose a method using facial dynamics to verify kinship from videos. Our approach intuitively makes sense: we all know people who do not look like their parents, until they smile. Furthermore, findings of [14] show that the appearance of spontaneous facial expressions of born-blind people and their sighted relatives are similar. However, the resemblance between facial expressions depends not only on the appearance of the expression but also on its dynamics, as each expression is created by a combination of voluntary and involuntary muscle movements. This is the key insight behind this paper. In this paper, we verify this insight empirically, and show that dynamic features obtained during facial expressions have discriminatory power for the kinship verification. This is the first work that uses dynamic features for kinship detection. By combining dynamic and spatio-temporal features, we approach the problem of automatic kinship verification. We use the recently collected UvA-NEMO Smile Database [3] in our experiments, compare our method with three recent approaches from the literature [8, 9, 21], and report state-of-the-art results.

2. Related work

In one of the first works on kinship verification, Fang *et al.* used the skin, hair and eye color, facial geometry measures, as well as holistic texture features computed on texture gradients of the whole face [8]. They have selected the most discriminative inherited features. Color based features performed better than the other features in general, since a good registration between individual face images was largely lacking in their approach. In the present study, we use their approach as a baseline under controlled registration conditions.

Different feature descriptors are evaluated for the kinship verification problem in the literature. In [9], eyes, mouth and nose parts are matched via DAISY descriptors. During matching, it is not expected to have good matches on all features, but on some features. Therefore, typically, the top few

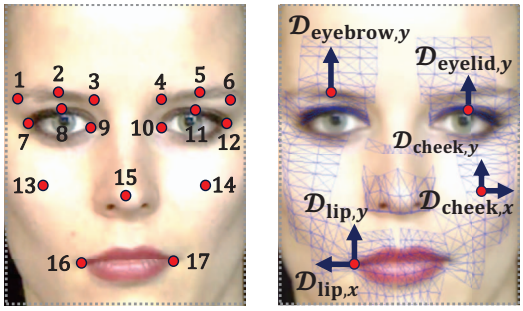
matching features are used for verification. In [21], Gabor-based Gradient Orientation Pyramid (GGOP) descriptors are proposed and used to model facial appearance for kinship verification. Support vector machines (SVM) with radial basis function kernels are used as the classifier. A mean accuracy of around 70% is reported on 800 image pairs. This is well within human kinship estimation range. In [11], the Self Similarity Representation of Weber face (SSRW) algorithm is proposed. Each face is represented by only its reflectance and difference of Gaussian filters are used to select keypoints to represent each face. SVM classifiers with different kernel functions are contrasted, and a linear kernel is found to be the most suitable. While SVM seems to be the classifier of choice for kinship verification, in [12], a metric learning approach is adopted. Samples that have the kinship relation are pulled close, and other samples are pushed apart. In this space, the transformation is complemented by defining a margin for kinship.

The evaluation protocols used for the kinship verification problem typically make use of pairs of photographs, where each pair is either a positive sample (i.e. kin) or a negative one. In [9], 100 face pairs with kinship and 100 pairs without are selected from family photos. There was no decomposition of results into specific kinship categories. In [8], [21], and [20] photos of celebrities have been downloaded from the Internet. In these studies, as well as in [12], four kinship relations (Father-Son, Father-Daughter, Mother-Son and Mother-Daughter) are analyzed separately. The largest database reported in the literature so far is the KinFaceW-II image database, with 250 pairs of kinship relations for each of these four categories.

In [14], Peleg *et al.* analyze the spontaneous facial expressions of born-blind people and their sighted relatives. They show that such expressions carry a unique family signature. Occurrences of a set of facial movements are used to classify families of blind subjects. Results show 64% correct classification on the average, with 60% in joy expressions. These results justify our motivation. Although [14] has focused on the facial movements for the task, they did not analyze the dynamics of expressions in terms of duration, intensity, speed, and acceleration, which is an empirical contribution of this paper.

3. Method

In this paper, we propose to combine spatio-temporal facial features and facial expression dynamics for the kinship verification. To this end, videos of enjoyment smiles are used. Our system analyzes the entire duration of a smile, starting from a moderately frontal and neutral face, the unfolding of the smile, and the return to the neutral face. Unlike other approaches proposed in the literature, our method works with videos of faces, rather than images. This is the first approach using videos for kinship verification.



(a) (b)

Figure 1. (a) The facial feature points used in this study with their indices, (b) the 3D mesh model and visualization of the amplitude signals, which are defined as the mean of left/right amplitude signals on the face. For simplicity, visualizations are shown on a single side of the face

We summarize the proposed method here. Our approach starts with face detection in the first frame and the localization of 17 facial landmarks, which are subsequently tracked during the rest of the video. Using the tracked landmarks, displacement signals of eyebrows, eyelids, cheeks, and lip corners are computed. Afterwards, the mean displacement signal of the lip corners is analyzed and the three main temporal phases (i.e. onset, apex, and offset, respectively) of the smile are estimated. Then, facial expression dynamics on eyebrows, eyelids, cheeks, and lip corners are extracted from each phase separately. To describe the change in appearance between the neutral and the expressive face (i.e. the apex of the expression), temporal Completed Local Binary Pattern (CLBP) descriptors are computed from the eye, cheek, and lip regions. After a feature selection step, the most informative dynamic features are identified and combined with temporal CLBP features. Finally, resulting features are classified using SVMs. In the rest of the section we provide more detailed information for each of these steps.

3.1. Landmark detection and tracking

Both the correct detection and accurate tracking of facial landmarks are crucial for normalizing and aligning faces, and for extracting consistent dynamic features. In the first frame of the input video, 17 facial landmarks (i.e. centers of eyebrows, eyebrow corners, eye corners, centers of upper eyelids, cheek centers, nose tip, and lip corners) are detected using a recent landmarking approach [4] (see Fig. 1(a)). This method models Gabor wavelet features of a neighborhood of the landmarks using incremental mixtures of factor analyzers and enables a shape prior to ensure the integrity of the landmark constellation. It follows a coarse-to-fine strategy; landmarks are initially detected on a coarse level and then fine-tuned for higher resolution. Then, these points are tracked by a piecewise Bézier volume deformation (PBVD) tracker [18] during the rest of the video.

Initially, the PBVD tracker warps a generic 3D mesh model of the face (see Fig. 1(b)) to fit the facial landmarks in the first frame of the image sequence. 16 surface patches form the generic face model. These patches are embedded in Bézier volumes to guarantee the continuity and smoothness of the model. Points in the Bézier volume, $x(u, v, w)$ can be defined as:

$$x(u,v,w) = \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l} b_{i,j,k} B_i^n(u) B_j^m(v) B_k^l(w), \quad (1)$$

where the control points denoted with $b_{i,j,k}$ and mesh variables $0 < \{u, v, w\} < 1$ control the shape of the volume. $B_i^n(u)$ denotes a Bernstein polynomial, and can be written as:

$$B_i^n(u) = \binom{n}{i} u^i (1-u)^{n-i}. \quad (2)$$

Once the face model is fitted, the 3D motion of the head, as well as individual motions of facial landmarks can be tracked based on the movements of mesh points. 2D movements on the face (estimated by template matching between frames, at different resolutions) are modeled as a projection of the 3D movement onto the image plane. Then, the 3D movement is calculated using projective motion of several points.

3.2. Registration

Faces in each frame need to be aligned before the feature extraction step. To this end, 3D pose of the faces are estimated and normalized using the tracked 3D landmarks ℓ_i (see Fig. 1(a)). Since a plane can be constructed by three non-collinear points, three stable landmarks (eye centers and nose tip) are used to define a normalizing plane $\mathcal{P}$. Eye centers $c_1 = \frac{\ell_7+\ell_9}{2}$ and $c_1 = \frac{\ell_{10}+\ell_{12}}{2}$ are the middle points between the inner and outer eye corners. Then, angles between the positive normal vector $\mathcal{P}$ and unit vectors on X (horizontal), Y (vertical), and Z (perpendicular) axes give the relative head pose. Computed angles (θ_z) and (θ_y) give the exact roll and yaw angles of the face with respect to the camera, respectively. Nevertheless, the estimated pitch (θ_x) angle is a subject-dependent measure, since it depends on the constellation of the eye corners and the nose tip. If the face in the first frame is assumed as approximately frontal, then the actual pitch angles (θ'_x) can be calculated by subtracting the initial value. After estimating the pose of the head, tracked landmarks are normalized with respect to rotation, scale, and translation. Aligned points ℓ'_i can be defined as follows:

$$\ell'_i = \left[\ell_i - \frac{c_1+c_2}{2}\right] R(-\theta'_x, -\theta_y, -\theta_z) \frac{100}{\rho(c1,c2)}, \quad (3)$$

$$R(\theta_x, \theta_y, \theta_z) = R_x(\theta_x) R_y(\theta_y) R_z(\theta_z), \quad (4)$$

and R_x, R_y, and R_z are the 3D rotation matrices for the given angles. ρ denotes the Euclidean distance between the given points. On the normalized face, the middle point between eye centers is located at the origin and the inter-ocular distance (distance between eye centers) is set to 100 pixels. Since the normalized face is approximately frontal with respect to the camera, we ignore the depth (Z) values of the normalized feature points ℓ'_i, and denote them as l_i.

3.3. Temporal segmentation

In the proposed method, dynamic and spatio-temporal features are extracted from videos of smiling persons. We choose to use the smile expression, since it is the most frequently performed facial expression, for showing several different meanings such as enjoyment, politeness, fear, embarrassment, etc. [5]. A smile can be defined as the upward movement of the lip corners, which corresponds to Action Unit 12 in the facial action coding system (FACS) [6]. Anatomically, the *zygomatic major* muscle contracts and raises the corners of the lips during a smile [7].

Most facial expressions are composed of three non-overlapping phases, namely: the onset, apex, and offset, respectively. Onset is the initial phase of a facial expression and it defines the duration from neutral to expressive state. Apex phase is the stable peak period (may also be very short) of the expression between onset and offset. Likewise, offset is the final phase from expressive to neutral state. Following the normalization step, we detect these three temporal phases of the smiles.

For this purpose, the amplitude signal of the smile $\mathcal{S}$ is estimated as the mean distance (Euclidean) of the lip corners to the lip center during the smile. Then, the computed amplitude signal is normalized by the length of the lip. Since the faces are normalized, center and length of the lip is calculated only once in the first frame. Afterwards, the longest continuous increase in $\mathcal{S}$ is defined as the onset phase. Similarly, the offset phase is detected as the longest continuous decrease in $\mathcal{S}$. The phase between the last frame of the onset and the first frame of the offset defines the apex.

3.4. Features

We extract two types of features from the faces. What we call *dynamic features* are based on the movement of landmark points in the registered faces over the expression duration. These do not contain appearance information. In contrast, what we call *spatio-temporal features* denotes appearance features obtained from multiple frames jointly, thus contain both spatial and temporal appearance information. These features are explained in detail next.

3.4.1 Extraction of dynamic features

To describe the smile dynamics, we use horizontal and vertical movements of tracked landmarks and extract a set of dynamic features separately from different face regions. Vertical and horizontal amplitude signals are computed from the movements of eyebrows, eyelids, cheeks, and lip corners. The (normalized) eye aperture $\mathcal{D}_{\text{eyelid}}$, and displacements of eyebrow $\mathcal{D}_{\text{eyebrow}}$, cheek $\mathcal{D}_{\text{cheek}}$ and lip corner $\mathcal{D}_{\text{lip}}$, are estimated as follows:

$$\mathcal{D}_{\text{eyelid}}(t) = \frac{\frac{l_7^t + l_9^t}{2} - l_8^t}{2\rho(l_7^t, l_9^t)} + \frac{\frac{l_{10}^t + l_{12}^t}{2} - l_{11}^t}{2\rho(l_{10}^t, l_{12}^t)}, \quad (5)$$

$$\mathcal{D}_{\text{eyebrow}}(t) = \frac{\frac{l_1^1 + l_2^1 + l_3^1}{3} - l_2^t}{2\rho(l_1^1, l_3^1)} + \frac{\frac{l_4^1 + l_5^1 + l_6^1}{3} - l_5^t}{2\rho(l_4^1, l_6^1)}, \quad (6)$$

$$\mathcal{D}_{\text{cheek}}(t) = \frac{\left|\frac{l_{13}^1 + l_{14}^1}{2} - l_{13}^t\right| + \left|\frac{l_{13}^1 + l_{14}^1}{2} - l_{14}^t\right|}{2\rho(l_{13}^1, l_{14}^1)}, \quad (7)$$

$$\mathcal{D}_{\text{lip}}(t) = \frac{\left|\frac{l_{16}^1 + l_{17}^1}{2} - l_{16}^t\right| + \left|\frac{l_{16}^1 + l_{17}^1}{2} - l_{17}^t\right|}{2\rho(l_{16}^1, l_{17}^1)}, \quad (8)$$

where l_i^t denotes the 2D location of the i^{th} point in frame t. Then, vertical (y) components of $\mathcal{D}_{\text{eyebrow}}$, $\mathcal{D}_{\text{eyelid}}$, $\mathcal{D}_{\text{cheek}}$, $\mathcal{D}_{\text{lip}}$, and horizontal (x) components of $\mathcal{D}_{\text{cheek}}$, $\mathcal{D}_{\text{lip}}$ are extracted (see Fig. 1(b)). Extracted sequences are smoothed by a 4253H-twice method [19]. These estimates are hereafter referred to as amplitude signals. Finally, amplitude signals are split into three phases as onset, apex, and offset, which have been previously defined using the smile amplitude $\mathcal{S}$.

Proposed dynamic features and their definitions are given in Table 1. It is important to note that the defined features are extracted separately from each phase of the smile. As a result, we obtain three feature sets for each of the six amplitude signals (see Fig. 1(b)). For a more detailed analysis, corresponding speed $\mathcal{V}(t) = \frac{d\mathcal{D}}{dt}$ and acceleration $\mathcal{A}(t) = \frac{d^2\mathcal{D}}{dt^2}$ signals are computed in addition to amplitudes.

In Table 1, signals marked with superindex ($^+$) and ($^-$) denote the increasing and decreasing segments of the related signal, respectively. For example, $\mathcal{D}^+$ pools the increasing segments in $\mathcal{D}$. η defines the length (number of frames) of a given signal, and ω is the frame rate of the video. For each phase of the amplitude signal, three 15-dimensional feature vectors are generated by concatenating these features. Combination of all the feature vectors forms the joint dynamic feature vector. In some cases, features cannot be calculated. For example, if we extract features from the amplitude signal of the lip corners $\mathcal{D}_{\text{lip}}$ using the onset phase, then decreasing segments will be an empty set ($\eta(\mathcal{D}^-) = 0$). For such exceptions, all the features describing the related segments are set to zero. This is done to have a generic feature vector format which has the same features for different phases of each amplitude signal.

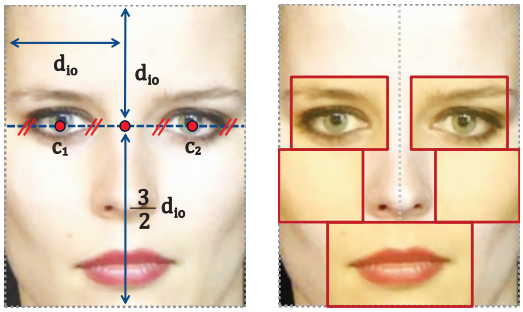
Table 1. Definitions of the extracted features.

Feature	Definition
Duration:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\omega}\ ,\ \frac{\eta(\mathcal{D}^-)}{\omega}\ ,\ \frac{\eta(\mathcal{D})}{\omega}\ \right]$
Duration Ratio:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\eta(\mathcal{D})}\ ,\ \frac{\eta(\mathcal{D}^-)}{\eta(\mathcal{D})}\ \right]$
Maximum Amplitude:	$\max(\mathcal{D})$
Mean Amplitude:	$\frac{\sum \mathcal{D}}{\eta(\mathcal{D})}$
Maximum Speed:	$\left[\ \max(\mathcal{V}^+)\ ,\ \max(\lvert\mathcal{V}^-\rvert)\ \right]$
Mean Speed:	$\left[\ \frac{\sum \mathcal{V}^+}{\eta(\mathcal{V}^+)}\ ,\ \frac{\sum \lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\ \right]$
Maximum Acceleration:	$\left[\ \max(\mathcal{A}^+)\ ,\ \max(\lvert\mathcal{A}^-\rvert)\ \right]$
Mean Acceleration:	$\left[\ \frac{\sum \mathcal{A}^+}{\eta(\mathcal{A}^+)}\ ,\ \frac{\sum \lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}\ \right]$

3.4.2 Extraction of spatio-temporal features

To describe the temporal changes in the appearance of faces, we employ a recently proposed spatio-temporal local texture descriptor, namely, the Completed Local Binary Patterns from Three Orthogonal Planes (CLBP-TOP) [16]. CLBP-TOP is a straightforward extension of Completed Local Binary Patterns (CLBP) operator [10] to describe dynamic textures (image sequences), which is calculated by extracting CLBP histograms from Three Orthogonal Planes XY, XT, and YT, individually, and by concatenating them as a single feature vector. Here, X and Y refer to the spatial extent of the image, and T denotes time. CLBP-TOP regards the face sequence as a volume, and the neighborhood of each pixel is defined in a three dimensional space, whereas CLBP uses only X and Y dimensions of a single image. Difference of the CLBP from the original LBP operator is that in addition to the sign of the local difference, it includes the center pixel of the local neighborhood and the magnitude of the difference.

We extract CLBP-TOP features from the previously detected smile onsets, since the onset phase shows the change from neutral to expressive face. On the selected frames, faces are normalized with respect to roll rotation using the eye centers c_1 and c_2. Then, each face is resized and cropped as shown in Fig. 2(a). For scaling and normalization, the inter-ocular distance d_{io} is set to 50 pixels. Resulting normalized face images have a resolution of 125×100 pixels. To provide more comparable onset du-



(a) (b)

Figure 2. (a) Scaling/cropping of a face image, and (b) the defined patches on eye, cheek, and mouth regions

rations for more reliable feature extraction, all smile onsets are temporally interpolated using bicubic interpolation on each YT plane. Then, six patches are cropped from each face (mouth, cheek, and eye regions) as shown in Fig. 2(b).

Each patch sequence is split into $X = 2{\times}Y = 2{\times}T = 3$ non-overlapping (equally-sized) blocks. Finally, CLBP-TOP features are extracted from these blocks using three neighborhood pixels (on a circle with a radius of a single pixel). All these features are concatenated to form the spatio-temporal feature vector.

3.5. Feature selection and classification

In kinship verification, the system is given pairs of samples, and the task is to verify whether the pair has the kinship relation or not. Essentially, a binary classification problem is solved. For this purpose, differences between feature vectors of the corresponding subjects are calculated. In our system, these differences are fed to individual support vector machine (SVM) classifiers trained with either dynamic features or spatio-temporal features. Afterwards, a weighted SUM rule is used to fuse the computed posterior probabilities for the target classes of these classifiers. To estimate these posterior probabilities, sigmoids of SVM output distances are used. Before classification, we employ the Min-Redundancy Max-Relevance (mRMR) algorithm [15] to select the discriminative dynamic features by eliminating feature redundancy. mRMR is an incremental method for minimizing the redundancy while selecting the most relevant information as follows:

$$\max_{f_j \in F - S_{m-1}} \left[I\left(f_j, c\right) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I\left(f_j, f_i\right) \right], \quad (9)$$

where I shows the mutual information function and c indicates the target class. F and S_{m-1} denote the feature set, and the set of $m-1$ features, respectively.

In our evaluation methodology, extreme care is taken to prevent overlearning of parameters. A separate validation set is used to determine the most discriminative dynamic features. Similarly, in order to optimize the SVM configuration, different kernels (linear, polynomial, and radial basis function (RBF)) with different parameters (size of RBF kernel, degree of polynomial kernel) are tested on the validation set and the configuration with the minimum validation error is selected. The test partition of the dataset is not used for parameter optimization.

4. Database

To analyze the role of smile dynamics in kinship verification, we have obtained the kinship annotations of the recently collected UvA-NEMO Smile Database [3]. By selecting the spontaneous and posed enjoyment smiles of the subjects who have kin relationships, we construct a kinship database which has 95 kin relations from 152 subjects. 15 of the subjects do not have spontaneous smile videos. And there is no posed video for six subjects. Each of the remaining subjects in the database has one or two posed/spontaneous enjoyment smiles. By using different video combinations of each kin relation, 228 pairs of spontaneous and 287 pairs of posed smile videos are included in the database. These pairs consist of Sister-Sister (S-S), Brother-Brother (B-B), Sister-Brother (S-B), Mother-Daughter (M-D), Mother-Son (M-S), Father-Daughter (F-D), and Father-Son (F-S) relationships. The relationship groups will be referred to as subsets in the remainder of the paper. Numbers of subjects and video pairs in each subset are given in Table. 2.

Table 2. Distribution of subject and video pairs in the dynamic kinship database.

Relation	Spontaneous		Posed	
	Subject	Video	Subject	Video
S-S	7	22	9	32
B-B	7	15	6	13
S-B	12	32	10	34
M-D	16	57	20	76
M-S	12	36	14	46
F-D	9	28	9	30
F-S	12	38	19	56
All	75	228	87	287

The kinship annotations in the database are obtained from the forms filled by the database subjects, and not verified with DNA analysis. This may have an insignificant effect on the annotation veracity. Ages of subjects vary from 8 to 74 years. Videos have a resolution of 1920×1080 pixels at a rate of 50 frames per second. These videos are publicly available[1] for research purposes. We will make the

[1] http://www.uva-nemo.org

collected kinship annotations available to the research community.

5. Experimental results

To evaluate our system, and to assess the reliability of facial expression dynamics and spatio-temporal facial information for kinship verification problem, we employ the above described database. Kinship relations in the database are used as positive samples of the classification problem. Negative samples are prepared using randomly selected samples with no kinship relation. These negative pairs are specifically constructed for each subset. For example, in the negative sample for a father-son relationship we retain the father, but replace the son with another male child.

In our experiments, the optimum number of selected dynamic features, fusion weights, kernel and parameters of SVM classifiers are determined on a separate validation partition. For this purpose, a two level leave-one-out cross-validation scheme is used. Each time videos of a test pair (subjects) are separated, the system is trained and parameters are optimized using leave-one-out cross-validation on the remaining pairs (without using the test partition). Similar to the results reported in [11] linear SVM is found to perform better than polynomial and RBF alternatives in our experiments. The tracking is initialized using the automatically annotated facial landmarks [4].

5.1. Dynamics versus spatio-temporal appearance

We train different systems using individual feature sets to compare the discriminative power of facial expression dynamics and spatio-temporal appearance. Then the outputs of these systems are fused with weighted SUM rule for assessing combined usage of dynamic and spatio-temporal features. Spontaneous smiles are used in this experiment. Correct verification rates for different features on the subsets and the whole set are given in Table 3.

Table 3. Correct verification rates for different features using spontaneous smiles.

Feature	Correct Ver. Rate (%)	
	Subsets (Mean)	Whole Set
Dynamics	60.84	54.61
Spatio-temporal	64.51	60.31
Combined	72.89	67.11

Results show that dynamic features do not perform as well as spatio-temporal appearance, but still provide discriminative information for kinship verification. The mean accuracy of the dynamic features on the subsets is only 60.84%, whereas the use of spatio-temporal appearance provides 64.51% correct verification rate. However, the output-level fusion of these two individual feature sets provide a statistically significant improvement on the verification accuracy, as determined by t-test (with $p < 0.01$). The higher accuracy of spatio-temporal features is expected, since they describe both facial appearance and the change in time, whereas the dynamic features can only define the duration and the dynamics of change in facial geometry.

We observe that the correct verification rates on the whole set are lower than the mean accuracies on the individual subsets. Dynamic features cause a higher decrease (6.23%) in accuracy on the whole set in comparison to that of spatio-temporal features (4.20%). This can be explained by the effect of age and gender on facial dynamics, since group specific training leads to dynamic features with better accuracy. We refer the reader to [3] and [2] for more detailed analysis of facial dynamics.

5.2. Role of face regions

To assess the influence of different regions on kinship verification accuracy, we train and test our system with region-specific features. Since both facial dynamics and spatio-temporal appearance features are extracted from regions of eye (eyelid and eyebrow), cheek, and mouth, these three regions and their combination are used in our tests. Spontaneous smiles are used in this experiment.

Table 4. Correct verification rates for different regions using spontaneous smiles.

Region	Correct Ver. Rate (%)	
	Subsets (Mean)	Whole Set
Eye	68.65	61.40
Cheek	64.48	55.70
Mouth	66.39	58.55
All	72.89	67.11

As shown in Table 4, features extracted from the eye region provide the highest verification accuracy on both group specific data (68.65%) and the whole set (61.40%). Mouth follows the eye region, and the cheek comes last. When we use all regions, a higher accuracy is obtained on both dataset settings. When we analyze the correct verification rates on group specific subsets and the whole set, it is seen that the decrease in the accuracy of using cheek region for the whole set is 13.61% (relative), whereas the average decrease for eye and mouth regions is 11.18% (relative). This result suggests that the cheek region is more sensitive to changes between different kin relations. This can be explained by the fact that spatio-temporal features describe the skin texture and the skin on the cheek surface can be discriminative for changes in age.

Table 5. Correct verification rates (%) for different methods.

Method	Spontaneous Smiles									Posed Smiles	
	S-S	B-B	S-B	M-D	M-S	F-D	F-S	Subsets (Mean)	Whole Set	Subsets (Mean)	Whole Set
Proposed: Dynamics	63.64	70.00	57.81	58.77	61.11	55.36	59.21	60.84	54.61	58.16	54.01
Proposed: Spatio-temporal	63.64	73.33	65.63	65.79	61.11	58.93	63.16	64.51	60.31	62.37	57.84
Proposed: Combined	75.00	70.00	68.75	67.54	75.00	75.00	78.95	72.89	67.11	70.02	64.98
CLBP-TOP: Smile Onset	63.64	66.67	60.94	60.53	58.33	67.86	65.79	63.39	57.02	60.22	55.23
CLBP: Neutral face	72.73	56.67	62.50	58.77	63.89	60.71	59.21	62.07	56.80	58.98	53.66
CLBP: Expressive face	56.82	60.00	57.81	55.26	58.33	57.14	55.26	57.23	53.07	56.79	54.88
Fang *et al.* (2010) [8]	61.36	56.67	56.25	56.14	55.56	57.14	55.26	56.91	53.51	53.11	52.79
Guo & Wang (2012) [9]	65.91	56.67	60.94	58.77	62.50	67.86	55.26	61.13	56.14	58.32	54.18
Zhou *et al.* (2012) [21]	63.64	70.00	60.94	57.02	56.94	66.07	60.53	62.16	58.55	57.42	54.18

5.3. Comparisons with other methods

The proposed system is compared with three recent approaches from the literature [8, 9, 21]. In [8], Fang *et al.* propose a system for kinship verification which uses shape and texture based features such as colors of eyes and skin, and distances between different facial locations, etc. The difference of these features for kin pairs is calculated and fed to a K-nearest neighbor classifier, with Euclidean distance. In [9], DAISY descriptors (fast local descriptor for dense matching) are extracted from eye, nose, and mouth patches for kinship analysis. The extracted features for each patch are matched with a scheme similar to the modified Hausdorff distance. Best three of the computed matching scores are used to verify the kinship between pairs using a Bayesian based voting. In [21], GGOP features are extracted from image pairs. Then, cosines of the difference between pairs are modeled by SVMs using radial basis function kernels. We have implemented both methods, and report results with the same experimental protocol (i.e. same training, validation, and test partitions).

We have also implemented a spatio-temporal baseline using CLBP-TOP features [16]. Frames in the smile onset portion of the videos (from neutral to expressive face) are split into $X = 8 \times Y = 8 \times T = 3$ non-overlapping blocks, and CLBP-TOP features are extracted from these blocks using three neighborhood pixels. Finally, we implemented two more baselines using CLBP features [10] on neutral and expressive faces. CLBP features are extracted from 8×8 non-overlapping blocks on the faces. All these baselines employ SVM classifiers using differences between the extracted features. CLBP-TOP and CLBP were not used for kinship verification before, but were successfully employed for analysis of face dynamics, which makes them suitable baseline approaches.

Both spontaneous and posed smiles are used for the comparisons. Table 5 shows the correct verification rates of baseline methods in addition to the proposed approaches. Because of the space constraints, individual accuracies of subsets are given only for spontaneous smiles. The proposed system that combines facial expression dynamics and spatio-temporal appearance, outperforms all baselines for each data setting. Moreover, even using only spatio-temporal features provides more accurate verification than the baseline methods. Proposed dynamic features cannot perform as well as spatio-temporal appearance, however, their combination with spatio-temporal features provides higher accuracy for both spontaneous and posed smiles. Results of CLBP based approaches show that expressive faces are less reliable than neutral faces for kinship verification. Based on the temporal appearance results, we can say that using only eye, cheek, and mouth regions (*Proposed: Spatio-temporal*) provides more reliable information for CLBP-TOP feature extraction, in comparison to the use of the whole face (CLBP-TOP: Smile Onset).

When we compare the average accuracy of all methods for spontaneous and posed smiles, it is seen that posed smiles provide 4% and 3% (relative) less accurate verification on the subsets and on the whole set, respectively. This can be explained by the learned characteristics of the posed expressions. However, the dynamics of posed smiles are still informative for kinship verification. Additionally, we have analyzed the average accuracy of all methods on each of the kinship subsets. Our findings show that the most accurate results are obtained for the sister-sister and brother-brother pairs. This result can be explained by the resemblance in terms of age and gender.

6. Conclusions

We have proposed a first exploration of facial dynamics for the difficult kinship verification problem, and ob-

tained results that advance the state-of-the-art in this area. While conducted on a database with fewer number of subjects than related studies, our experiments were performed on a high-resolution database with controlled conditions, and precise age ground-truth annotations. We evaluate our proposed method and contrast it with several baseline approaches from the literature. Our results show that incorporating facial expression dynamics allows the computer to perform kinship verification at rates higher than reported for humans.

Kinship verification is a relatively recent problem. An interesting related work in the biometrics literature is the studies on face recognition on identical twins. Jonathon Phillips and colleagues looked at 126 twin pairs, and verified that under controlled environments and small time differences between captured images, state-of-the-art face recognition algorithms had no trouble separating identical twins [17]. However, when one year passed between image capturing sessions, the performance quickly degraded. The implication is that the level of automatic analysis on faces is advanced to the point that very small deviations can be detected, but with time, these deviations accumulate, and the threshold with which the system judges similarity becomes difficult to tune. Kinship detection in a sense springs from the exact opposite paradigm with twin biometrics: instead of magnifying small differences, similarities are captured and magnified. Research into this interesting problem will advance both automatic face analysis, and our understanding of how humans evaluate facial appearance.

Acknowledgments

This research was part of Science Live, the innovative research programme of science center NEMO that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers. Additionally, this study was supported by the Dutch national program COMMIT and Boğaziçi University project BAP-6531. Authors would like to thank Prof. Ethem Alpaydın for his invaluable suggestions which initiated this work.

References

[1] M. F. Dal Martello and L. T. Maloney. Where are kin recognition signals in the human face? *Journal of Vision*, 6(12), 2006.

[2] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti. A smile can reveal your age: Enabling facial dynamics in age estimation. In *ACM Multimedia*, pages 209–218, 2012.

[3] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *ECCV*, pages 525–538, 2012.

[4] H. Dibeklioğlu, A. A. Salah, and T. Gevers. A statistical method for 2-d facial landmarking. *IEEE Trans. on Image Processing*, 21(2):844–858, 2012.

[5] P. Ekman. *Telling lies: Cues to deceit in the marketplace, politics, and marriage*. New York: WW. Norton & Company, 1992.

[6] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press Inc., San Francisco, CA, 1978.

[7] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. *J Nonverbal Behav*, 6:238–252, 1982.

[8] R. Fang, K. D. Tang, N. Snavely, and T. Chen. Towards computational models of kinship verification. In *IEEE ICIP*, pages 1577–1580, 2010.

[9] G. Guo and X. Wang. Kinship measurement on salient facial features. *IEEE Trans. on Instrumentation and Measurement*, 61(8):2322–2325, 2012.

[10] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. on Image Processing*, 19(6):1657–1663, 2010.

[11] N. Kohli, R. Singh, and M. Vatsa. Self-similarity representation of Weber faces for kinship classification. In *IEEE BTAS*, pages 245–250, 2012.

[12] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang. Neighborhood repulsed metric learning for kinship verification. In *CVPR*, pages 2594–2601, 2012.

[13] L. T. Maloney and M. F. Dal Martello. Kin recognition and the perceived facial similarity of children. *Journal of Vision*, 6(10), 2006.

[14] G. Peleg, G. Katzir, O. Peleg, M. Kamara, L. Brodsky, H. Hel-Or, D. Keren, and E. Nevo. Hereditary family signature of facial expression. *Proceedings of the National Academy of Sciences*, 103(43):15921–15926, 2006.

[15] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on PAMI*, 27(8):1226–1238, 2005.

[16] T. Pfister, X. Li, G. Zhao, and M. Pietikainen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *ICCV Workshops*, pages 868–875, 2011.

[17] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *IEEE FG*, pages 185–192, 2011.

[18] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise Bézier volume deformation model. In *CVPR*, number 1, pages 611–617, 1999.

[19] P. F. Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, pages 609–615, 1980.

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

[21] X. Zhou, J. Lu, J. Hu, and Y. Shang. Gabor-based gradient orientation pyramid for kinship verification under uncontrolled environments. In *ACM Multimedia*, pages 725–728, 2012.