Face Recognition using Spherical Wavelets

Christian Lessig

Abstract

“If there is technological advance without social advance, there is, almost automatically, an increase in human misery, . . .”

Michael Harrington

In this report we evaluate the applicability of the SOHO wavelet basis for face recognition. We demonstrate that a representation of 3D face models using spherical wavelets allows to distinguish between subjects, and that the intra-subject variability due to facial expressions is smaller than inter-subject variations.

1 Introduction

Traditionally, two-dimensional images have been employed for face recognition. Recently however, approaches based on three-dimensional facial representations became popular [14, 2] due to improved capturing techniques (e.g. [12, 1]) and the promise of higher recognition rates by using signals in the native space $\mathbb{R}^3$ instead of projections $\mathbb{R}^3 \to \mathbb{R}^2$ [15, 27, 28, 40, 48]. The last Face Recognition Vendor Test (FRVT) demonstrated the viability of 3D face recognition systems and these techniques provided the same accuracy as 2D approaches [43]. These results are particularly promising given the young history of the field and the relatively small number of approaches which have been proposed in the literature.

In this report, we propose a novel technique for 3D face recognition which employs the recently developed SOHO wavelet basis [36, 37]. The input to our technique is a 3D facial mesh which is projected onto the sphere, and the resulting spherical signal is then transformed using the SOHO basis. The basis function coefficients provide the template which is used for recognition, and standard techniques for 1D digital signal analysis can be employed to compare faces.

Work related to our technique can be found both in the face recognition literature and in Computer Graphics. In the following we will focus on 3D techniques [14, 2] and refer the interested reader to the surveys by Samal and Iyengar [46], Chellapa et al. [21], and Zhao et al. [57] for a discussion of 2D approaches. Commercial systems, which outperformed techniques submitted by academic institutions in the FRVT 2006 [43], will also be excluded from the following discussion because in general little is known about the underlying technology.

*lessig@dgp.toronto.edu
Gordon [26] was one of the first to employ 3D information for face recognition. He assumed that certain facial regions such as the forehead and the cheeks are almost invariant under different poses and facial expressions, and concluded that these regions are particularly well suited for a face recognition system. He employed curvature to extract invariant regions and high-level shape descriptors projected into a feature space for comparison. Cartoux et al. [18] similarly used curvature to segment faces but these authors utilized the plane of bilateral symmetry for comparison. Tanaka et al. [49] projected the minimal and maximal principal curvature directions of a face onto the sphere and employed these Extended Gaussian Images for face recognition, and Moreno et al. [41] employed Gaussian curvature to segment faces and compute feature vectors. As pointed out by Abate et al. [2], approaches which directly employ curvature are not scale-invariant and cannot recognize the same face presented in different sizes.

Similar to the work by Gordon [26], Chua et al. [22] employed the forehead, the cheeks, and other regions invariant to pose and facial expressions. These author used however point signatures to identify invariant regions. Wang et al. [54] used sphere spin images, which are histograms derived from local shape descriptors computed at points of minimal curvature, and performed face recognition by computing correlation coefficients between the histograms of different subjects. Nagamine et al. [42] identified five feature points in the face to standardize pose and then employed various curves and profiles to characterize a face. In similar work, Beumier and Acheroy [9] used vertical face profiles with gray values for face recognition; and Wu et al. [55] employed horizontal profiles. Gaussian-Hermite moments have been used by Xu et al. [56] as local shape descriptors to compare faces, and Irfanoglu et al. [30] employed the Point Set Distance, which is a discrete approximation of the volume between two models, for face recognition.

Blanz and Vetter [11, 13, 10] proposed a full face recognition system including the 3D model generation. Their technique requires as input only a single 2D image and they estimate 3D mesh models using a morphing technique proposed in earlier work [12]. Different approaches to compare face models are discussed by these authors and Blanz and Vetter employed for example a distance computed from the parameters of the morphable models of different subjects. In similar work, Lu et al. [38] also created a 3D model of a subject from a single image using morphing. However they then generated synthetic images from the subject in different poses and with different illuminations, and employed the rendered images for face recognition using standard 2D techniques. Ansari and Abdel-Mottaleb [7] employed a frontal and a side view to obtain a morphable model, and they used 29 predefined feature points for recognition, exploiting that a morphable model guarantees proper alignment of the meshes. Approaches based on morphable models have been criticized for not being accurate. Abate et al. [2] argued that a single image is not sufficient to uniquely reconstruct a 3D representation, and it has been recognized that synthesized facial expressions might look realistic but often differ substantially from observed ones.

Similar to eigenfaces [35] which is the canonical algorithm for 2D face recognition, PCA has also been employed by different authors for 3D face recognition [39, 28, 3]. Another machine learning technique, Gaussian Mixture Models, has been used by Cook et al. [23] to compare faces.

A variety of authors employed a combination of 2D and 3D techniques. Chang et al. [20, 19] performed PCA separately on grayscale and depth images and combined the results using a weighted average. The authors demonstrated that using both modalities out-performs using either of them alone. In similar work, Tsalakanidou et al. [52] employed hidden markov models to combine 2D and 3D face recognition results. Wang et al. [53] used Gabor wavelets to characterize the 2D image and point signatures to compactly describe the 3D shape.

Bronstein et al. [16, 15, 17] developed a technique based on isometric embedding [25]
which is in spirit similar to the ISOMAP algorithm [50] and Locally Linear Embedding [45]. Given the original geodesic distances of the samples they find a mapping onto a low-dimensional space such that the total error of the geodesic distances after projection is minimized. Multidimensional scaling on the metric tensor of the original surface is employed for the embedding. The authors demonstrate that a representation using the geodesic distances is less sensitive to facial expressions and they therefore claim that their approach is (almost) expression invariant. Different embedding spaces are employed by the authors and Bronstein et al. report that the sphere $S^2$ provides the least distortion [15]. Our work has similarity with those of Bronstein et al. We also employ a projection of the facial mesh onto $S^2$ but we directly resample the mesh onto the sphere along the normal direction and therefore do not require expensive nonlinear optimization. In contrast to the Spherical Harmonics employed by Bronstein et al., we use spherical wavelets because the basis functions are localized, allow data-dependent approximation, do not suffer from ringing artifacts, and the transforms can be computed more efficiently. The SOHO wavelet basis is thereby the representation of choice because it is the only known spherical Haar wavelet basis which is symmetric and orthogonal, making it particularly well suited for the efficient representation and approximation of spherical signals; the readers is referred to the paper by Lessig and Fiume [37] and the thesis by Lessig [36] for a more detailed discussion of the SOHO basis.

In Computer Graphics, different techniques have been developed to compare and identify shapes. In early work, Horn [29] used the normal distribution across the surface as shape descriptor, while Ankerst et al. [6] employed the distribution of points on the surface of the model. Spherical Extent Functions, which describe the maximal extent of an object from the origin, have been employed by Saupe and Vranic [47] to characterize shapes. Kazhdan and coworkers [31, 32] employed a reflective symmetry descriptor which represents the symmetry of a model with respect to all possible planes through the origin. In later work, the same authors employed a discrete shell representation of a model and used Spherical Harmonics (SH) to compactly encode the shells [33, 34]. The SH basis is reflection and rotation invariant and the approach therefore overcomes the alignment problem which Kazhdan et al. considered as severe for previous techniques. Our work has also similarity with the shell technique by Kazhdan et al. We however do not require multiple shells. A face or a head is already approximately an embedding $S^2 \subset \mathbb{R}^3$ and we can therefore directly project the mesh onto $S^2$. Rotation- and reflection-invariance, which motivated the use of Spherical Harmonics by Kazhdan et al., is no severe problem in our application since for example the ears and the mouth, or the eyes and the mouth, naturally span a reference coordinate frame.

2 Face Recognition using the SOHO Wavelets

The input to our algorithm for 3D face recognition is a facial mesh model. Given the model, first the projection $M_j \rightarrow \tilde{M}_j$ of every mesh triangle $M_j = \{p^j_1, p^j_2, p^j_3\}$, with $p^j_i \in \mathbb{R}^3$, onto the sphere is computed with

$$\tilde{M}_j = \{\tilde{p}^j_1, \tilde{p}^j_3, \tilde{p}^j_3\} = \left\{ \frac{p^j_1}{\|p^j_1\|}, \frac{p^j_2}{\|p^j_2\|}, \frac{p^j_3}{\|p^j_3\|} \right\},$$

where $\| \cdot \|$ denotes the $\ell_2$ norm, and $\tilde{p}^j_i \in S^2$. Next, the partition tree $T = \{T_{l,k}\}$, over which the SOHO basis is defined, is constructed up to some level $l_k$, and then the mesh is resampled onto the domains $T_{l_k,j}$. For every $T_{l_k,j}$ we generate $M$ random samples $s_m \in S^2$ and then determine

$$d^m_{l_k,j} = \max_{j \in J_{s_m}} \|p^j_{s_m}\|,$$
where $J_{sm}$ is the set of all projected mesh faces such that $s_m \in \tilde{M}_j$, and $p_{sm}^j$ is the closest point to $s_m$ in $M_j$. The normalized average

$$d_{lk,j} = \frac{1}{M} \sum_{m=1}^{M} \frac{d_{lk,j}^m}{d_{max}}$$

over all samples $s_m$ yields the final depth value for $T_{lk,j}$. The maximal distance $d_{max}$ is defined as

$$d_{max} \equiv \max_{i,j} \|p_i^j\|,$$

and leads to the scale-invariance of our representation. Note that the resampling of the facial mesh onto the partitions makes the technique robust against variations in the mesh model, for example a variable number of faces. An example for the projection of a head model is shown in Fig. 1.

The resampling of the facial mesh onto the $T_{lk,j}$ yields a spherical signal with scaling function coefficients $\lambda_{lk,j} \equiv d_{lk,j}$. The forward wavelet transform with the SOHO basis is employed to obtain wavelet basis function coefficients, and these coefficients provide the template which is used for recognition. Note that, according to the general theory of wavelets, we can expect that most of the basis function coefficients are very small or zero, and that a very small number of the coefficients is sufficient to characterize a signal.

3 Evaluation

The 3D face recognition algorithm proposed in Sec. 2 has been evaluated with 16 facial mesh models from four different individuals and four different facial expressions per subject. The models have been generated with Poser [24] and we employed Maya [8] to triangulate and simplify the quad meshes. The models employed for the experiments had approximately 1,250 vertices for the man in the top row of Fig. 2; 1,300 vertices for the woman in the second row; 2,600 vertices for the girl in the third row; and 1,050 vertices for the boy in the last row of Fig. 2. For the first two and the last model we employed a very low number of vertices to test the applicability of our algorithm for systems where the input data comes from non-intrusive capturing techniques; and the third model has been used to assess the influence of the mesh resolution on the algorithm.

For the SOHO wavelet basis an octahedron was used as base polyhedron, and we chose $l_k = 5$ resulting in 8,192 partitions $T_{lk,j}$. The first four octants correspond to the upper
hemisphere and the last four octants to the lower hemisphere. All experiments were performed with the Matlab programming environment [51].

In Fig 3 the magnitude of the basis function coefficients for the eighth octant before and after approximation are shown for the models in the first column in Fig. 2. For the approximation we retained 512 or 6.25% of the 8,192 coefficients. The results show that only very small coefficients are affected by the approximation and that the peaks, which characterize the signals, are retained almost unaffected. The possibility of such an approximation provides a three-fold advantage. Firstly, only a small fraction of the basis function coefficients has to be stored for each template, which is important for large databases; secondly, the approximation makes comparisons of templates, and therefore the recognition problem, more efficient; and thirdly, the small coefficients correspond mostly to noise and using only the large coefficients improves the robustness of the system. In the following we will only consider templates with the $k = 512$ largest basis function coefficients.

Fig. 4 shows the templates for all 16 models, and the basis function coefficient spectra of the last four octants are shown in detail in Fig. 5 to Fig. 8. In Fig. 4 it can be seen that the first four octants contain almost no information. With the projection of a face mesh onto $S^2$ shown in Fig. 1 it becomes clear that all features such as ears, eyes, and nose lie almost entirely in the lower hemisphere, and that the upper hemisphere therefore only encodes the head shape. The very small number of significant basis function coefficients thereby confirms that a sphere is a suitable embedding for a head. The results also demonstrate that the head shape is not discriminating between subjects, at least not with the projection currently employed by us. The detail plots in Fig. 5 to Fig. 8 show the basis function coefficients corresponding to the characteristic features of the faces such as the eyes, the nose, and the mouth. The graphs demonstrate that the intra-subject variability is smaller than the inter-subject variations and that each face has a characteristic spectrum of peaks in the set of wavelet basis function coefficients. Facial expressions thereby correspond to a local scaling of basis function coefficients but in general do not alter the position of the peaks. We attribute the discriminating power of our algorithm and its approximate invariance to facial expressions to the localization of the SOHO basis functions in space and frequency.

Give the results presented in this section we conclude that the algorithm proposed in Sec. 2 is in principal applicable for 3D face recognition.

4 Future Work

In Sec. 3 we demonstrated the applicability of the SOHO wavelet basis for 3D face recognition. Significant work remains however until the technique described in Sec. 2 can be employed in a practical system.

Of particular importance is the development of an automatic algorithm to compare templates. One approach, which is commonly used to compare digital signals, would be to compute correlation coefficients. One could however also imagine to employ machine learning techniques such as hidden markov models or neuronal networks. Another alternative would be to use string matching algorithms which have been employed successfully in Computational Structural Biology [4, 5]. It is thereby currently unclear to which extent the magnitude of the coefficients should be included in the comparison and further research is required to answer this question. A more thorough investigation of the approximate invariance to facial expression, which we reported in Sec. 3, is also necessary. We currently assume that it occurs for similar reasons as in the work by Bronstein et al. [16, 15, 17] but we would like to understand the observations in more detail.

In our experiments we employed the octahedron as base polyhedron for the SOHO basis. It
is also possible to use the tetrahedron or the icosahedron and previous results demonstrate that the choice of the base polyhedron can significantly affect the performance [36]. Future experiments should therefore include the tetrahedron and the icosahedron and determine which one is best suited for face recognition. Similarly, the method currently used to project the face model onto the sphere has not been compared to alternative approaches. Although its simplicity has advantages, in particular in terms of computation time, one could also imagine to employ the approach proposed by Bronstein et al. [16, 15, 17], or one of the many techniques for mapping mesh models onto the sphere which can be found in the literature (e.g. [44]).

The results discussed in Sec. 3 demonstrate that at most four octants contain significant information about the face or more generally the head. In practice it therefore seems to be sufficient to employ the octant(s) with the most significant and most discriminating information. This would reduce storage requirements for the template and speed-up comparisons. The efficacy of our face recognition system could most likely also be improved by exploiting texture information, as done by many other state-of-the-art face recognition systems [43, 20, 19, 53, 52].

Crucial for every system is a meaningful evaluation. The data set employed for the FRVT 2006 is available upon request and we believe this would provide a good basis for a comparison to other systems.

5 Conclusion

In this report we investigated the applicability of the SOHO wavelet basis for 3D face recognition. The proposed algorithm first projects a facial mesh model onto the sphere, and then performs a wavelet transform using the SOHO wavelets to obtain basis function coefficients which serve as template. Our results demonstrate that the template is characteristic for a face, and that intra-subject variability due to facial expressions is smaller than inter-subject variations. We also showed that it is sufficient to employ only a small subset of the $k$ largest basis function coefficients, which significantly reduces the storage requirements for a template, increases the robustness of the algorithm, and reduces the comparison time. With the obtained results, we believe it would be worthwhile to further develop the proposed algorithm to a full 3D face recognition system.

Acknowledgement

We thank Patricio Simari for the Matlab OBJ loader.
References


Figure 2: Test models used for the experiments presented in Section 3.
B Evaluation Results

Figure 3: Spectrum of the wavelet basis function coefficients for the last octant before (left) and after approximation (right) where 6.25\% of all 8,192 coefficients were retained non-zero. Clearly visible is that the approximation only affects very small coefficients which can be identified as noise but that the peaks which characterize the signals are almost unaffected.
Figure 4: Full basis function coefficient spectrum for the 16 models in the test set.
Figure 5: Basis function coefficient spectrum for octant 5 for the 16 models in the test set.
Figure 6: Basis function coefficient spectrum for octant 6 for the 16 models in the test set.
Figure 7: Basis function coefficient spectrum for octant 7 for the 16 models in the test set.
Figure 8: Basis function coefficient spectrum for octant 8 for the 16 models in the test set.