UNIVERSITY OF CALIFORNIA
Los Angeles

A Part-Based, Multiresolution, TensorFaces Approach to Image-Based Facial Verification

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Eric Kim

2016

ABSTRACT OF THE THESIS

A Part-Based, Multiresolution, TensorFaces Approach to Image-Based Facial Verification

by

Eric Kim

Master of Science in Computer Science

University of California, Los Angeles, 2016

Professor Demetri Terzopoulos, Chair

In the field of computer vision, multilinear (tensor) algebraic approaches to image-based face recognition have attracted interest in recent years. Previously, these methods have operated uniformly over the entire facial image at uniform resolution. In this thesis, we present a multiresolution, region-based multilinear method. By computing multiple multilinear models of various facial features, such as the eyes, nose, and mouth, in appropriate spatially-localized regions, we achieve a representation that, using the same amount of training data, is more discriminative for the purpose of facial verification. Adding a multiresolution image pyramid as well as a mixture-of-experts weighting scheme further improves performance. We report encouraging experimental results on two datasets, one consisting of synthetic images, the other of real-world images.

The thesis of Eric Kim is approved.

Song-Chun Zhu

Alan Yuille

M. Alex O. Vasilescu

Demetri Terzopoulos, Committee Chair

University of California, Los Angeles

2016

Insert dedication here.

TABLE OF CONTENTS

List of Figures

List of Tables

ACKNOWLEDGMENTS

(Acknowledgments omitted for brevity.)

CHAPTER 1

Introduction

Face recognition is an important computer vision problem vital to a large number of applications, such as surveillance technologies, keyless biometric systems, human-computer interaction, etc. Despite being a well-studied problem that has attracted decades of involved research and development, it remains a big challenge for automated machine vision systems to accurately recognize faces deformed by different facial expressions that have been imaged from arbitrary viewpoints and under various lighting conditions. This is known as the unconstrained face recognition problem. Additional (nuisance) factors that hinder face recognition, but which need to be taken into account, include aging effects, facial hair, weight gain/loss, makeup, eyeglasses, etc. Variation in the appearance of a person's face in an image due to these factors in addition to pose, illumination, and facial expression must be explicitly modeled, and then disregarded, so as to achieve robust, unconstrained face recognition.

TensorFaces, a multilinear (tensor) algebraic approach to facial image representation and recognition, which was introduced by Vasilescu and Terzopoulos (Vasilescu and Terzopoulos, 2002a, 2003), computes person-specific facial signatures by explicitly decomposing and representing facial images in terms of the various casual factors associated with image formation—the person-specific facial geometry and reflectance, the illumination (*i.e.*, the location and types of light sources), and the imaging conditions (*i.e.*, viewpoint and camera characteristics). By performing a multilinear (Tucker) decomposition of a training image data tensor, the model is able to decouple and explicitly account for the causal factors that generate the observed image data, thus yielding *unique* person-specific signatures that are invariant to illumination, viewpoint, expression, and other nuisance factors.

Conventional TensorFaces represent facial appearance monolithically; that is, without isolating and explicitly representing important facial features. Motivated by the TensorFaces model, we introduce in this thesis a novel multilinear model that extracts person-specific facial signatures, not only monolithically over the entire image, but also as a combination of signatures computed for the eyes, nose, mouth, and other important facial features in appropriately localized regions of the facial image. We demonstrate that our part-based approach enhances the discriminative power of our multilinear model relative to the monolithic TensorFaces approach using the *same* amount of training data.

Since the important facial features are localized, they can be analyzed at a higher resolution than the remainder of the image. Adding a multiresolution image pyramid as well as a mixture-of-experts weighting scheme further improves the performance of our method.

We report the experimental results of applying our multiresolution region-based multilinear method to two datasets: one consisting of synthetic images (*University of Freiburg 3D Morphable Faces* (Blanz and Vetter, 1999)) and one consisting of real-world images (*Labeled Faces in the Wild* (Huang et al., 2007)).

1.1 Thesis Overview

The remainder of this thesis is organized as follows: Chapter 2 surveys relevant prior work. Chapter 3 reviews the mathematics of multilinear algebraic analysis of images and the TensorFaces model. Chapter 4 presents the technical details of our part-based, multiresolution, TensorFaces approach to image-based facial verification. Chapter 5 presents our experiments with our new technique and reports our results. Chapter 6 presents our conclusions and discusses promising avenues for future work.

CHAPTER 2

Related Work

In this chapter, we review relevant prior work, focusing on multilinear analysis, part-based approaches, ensemble learning, and deep learning,

2.1 Multimodal and Multilinear Analysis

Multimodal factor analysis in a tensor framework was introduced by Tucker (1966) and further developed by Kapteyn et al. (1986), Magnus and Neudecker (1988), and de Lathauwer et al. (2000). Multilinear (tensor) models have been previously applied to face recognition by Vasilescu and Terzopoulos (2002b) and Kan et al. (2016) and to face verification from video by Lee and Kwak (2011).

Multilinear analysis has been applied to other problems, such as image rendering (Vasilescu and Terzopoulos, 2004), and motion analysis (Vasilescu, 2002). Vlasic et al. (2005) automatically transfer facial animations from one person to another by multilinear analysis. They learn their model by decomposing a training-data tensor organized as the Cartesian product of (16 identities $\times$ 5 expressions $\times$ 5 visemes $\times$ 30K mesh vertices). With this multilinear model, they are able to transfer an individual's facial movements to a different target individual.

Researchers have also found connections between deep learning and multilinear analysis. For instance, Cohen et al. (2015) prove that shallow networks realize a CP (rank-1) decomposition, and that deep networks realize a Hierarchical Tucker decomposition (Hackbusch and Kühn, 2009).

2.2 Part-Based Approaches

The aforementioned facial representation approaches are monolithic inasmuch as they approach the representation problem using the entire image, rather than by decomposing the face into local parts. To further improve facial recognition systems, researchers have used part-based models to separately analyze facial parts, such as eyes, nose, mouth, *etc*.

Wong et al. (2012) and Gao et al. (2010) show that computing regional descriptors gives more robust verification performance compared to using the whole image. Following this direction, Berg and Belhumeur (2013) and Duan et al. (2013) detect facial parts and extract feature descriptors for each part. Similar to our work, Li and Hua (2015) build a hierarchical part-based face representation by recursively subdividing the face into subparts. Unlike our representation, these authors extract feature descriptors from each part that are not necessarily invariant to the variations present in an image. In our work, we develop a compositional region-based signature that uniquely captures the identity of the person while remaining invariant to factors such as viewpoint and illumination.

Xu et al. (2008) parse face images using a hierarchical compositional model and localize individual facial features (*i.e*., left-eye, right-eye, *etc*) via a collection of deformable templates within an *AND-OR* graph framework. Similar to our work, the parsing is done in a coarse-to-fine manner, where facial features requiring fine details reside at high-resolution levels of a Gaussian pyramid. Once the face has been parsed, the system can generate high-quality sketches of the facial image that capture structural details, including the wrinkles of the skin. In their system, facial analysis is a coupled process—for instance, the result of localizing and analyzing the left-eye affects the analysis of the local patches surrounding the left-eye. This allows their system to perform an adaptive analysis that can better cope with variations in facial features. In our own work, however, different parts of the face are analyzed independently.

2.3 Ensemble Learning

Combining a set of weak learners to create a single strong classifier is not a new idea in face recognition. Su et al. (2009) take an approach very similar to our own, where the image representation consists of a composite of global and local image regions. They learn a separate Fischer linear discriminant classifier for each region, and the final classifier is a weighted sum of the individual classifiers. Our work differs in several important ways. First, rather than using features derived from Gabor wavelets, we apply multilinear analysis to compute image features. Second, we employ a multiresolution image pyramid to extract features at different scales, as different facial features require different levels of detail.

Viola and Jones (2004) combine many weak-learners trained on simple Haar-like features in the AdaBoost framework (Freund and Schapire, 1997) to train an efficient, robust face detector.

Another method of combining classifier outputs is to proceed in a "divide-and-conquer" manner—partition the input space and assign different classifiers to each partition. Gutta et al. (2000) use an ensemble of radial-basis networks and decision trees to classify the gender, ethnicity, and pose of face images.

2.4 Deep Learning

In recent years, deep learning has been successfully applied to face verification (Taigman et al., 2014) due to the availability of large amounts of images from social networks, and high performance computing; *e.g.*, distributed and GPU computing (Abadi et al., 2015; Collobert et al., 2011; Bastien et al., 2012; Bergstra et al., 2010). The very resources that make deep learning a viable approach today are also its shortcomings. First, this abundance of data will result in multiple representations per person. In the limit, there can be as many representations as there are images rather than a unique signature per person, making online classification challenging. Second, while facial images uploaded on social networks may be representative of the world population, they are not typically representative of the appearance of an individual, since uploaded images are often

nearly-frontal, well-lit, and pass a minimum human vanity/aesthetic threshold.

Some current methods for face verification focus on using deep neural networks to automatically build a hierarchical feature representation in support of verification. For instance, DeepFace (Taigman et al., 2014) implements a 9-layer deep neural network trained on a proprietary, labeled dataset of 4,000,000 facial images. On the other hand, in Huang et al. (2012) train a deep network on facial descriptors, such as Local Binary Patterns (LBP), rather than on pixel intensities. Also, several variants of deep network architectures have been applied (Sun et al., 2013; Chen et al., 2015; Xiong et al., 2016). However, these approaches rely on large training datasets to learn numerous network parameters, whereas our method has a more compact representation and can be trained on a small representative dataset.

CHAPTER 3

Multilinear Analysis of Facial Images

In this chapter, we present a brief discussion of relevant multilinear concepts.

3.1 TensorFaces

TensorFaces (Vasilescu and Terzopoulos, 2002a, 2003; Vasilescu, 2009, 2011) is a multilinear (tensor algebraic) approach to facial analysis and recognition. By explicitly factoring out variations present in the dataset, TensorFaces learns a representation that is invariant to factors such as viewpoint and illumination (Figure 3.1).

3.1.1 Tensor Decomposition

The data tensor $\mathcal{D} \in \mathbb{R}^{I_{\text{P}} \times I_{\text{V}} \times I_{\text{L}} \times I_{\text{x}}}$ of training facial images, where I_{P}, I_{V}, I_{L}, I_{x} denote the number of people, viewpoints, illuminations, and image pixels, is decomposed to compute a representation in which the causal factors are made explicit. In particular, we compute a rank-$(R_{\text{P}}, R_{\text{V}}, R_{\text{L}})$ tensor decomposition (or *Tucker* decomposition (Tucker, 1966; Kroonenberg and De Leeuw, 1980; De Lathauwer et al., 2000)):

$$\mathcal{D} = \mathcal{T} \times_{\text{P}} \mathbf{U}_{\text{P}} \times_{\text{V}} \mathbf{U}_{\text{V}} \times_{\text{L}} \mathbf{U}_{\text{L}}, \tag{3.1}$$

where $\mathcal{T} \in \mathbb{R}^{R_{\text{P}} \times R_{\text{V}} \times R_{\text{L}} \times I_{\text{x}}}$ is the *extended core tensor*, $\times_m$ denotes the mode-m product, and $\mathbf{U}_{\text{P}} \in \mathbb{R}^{I_{\text{P}} \times R_{\text{P}}}$, $\mathbf{U}_{\text{V}} \in \mathbb{R}^{I_{\text{V}} \times R_{\text{V}}}$, and $\mathbf{U}_{\text{L}} \in \mathbb{R}^{I_{\text{L}} \times R_{\text{L}}}$ are the orthonormal matrices containing the basis vectors for the person, viewpoint, and illumination spaces respectively, with R_{P}, R_{V}, R_{L} being the

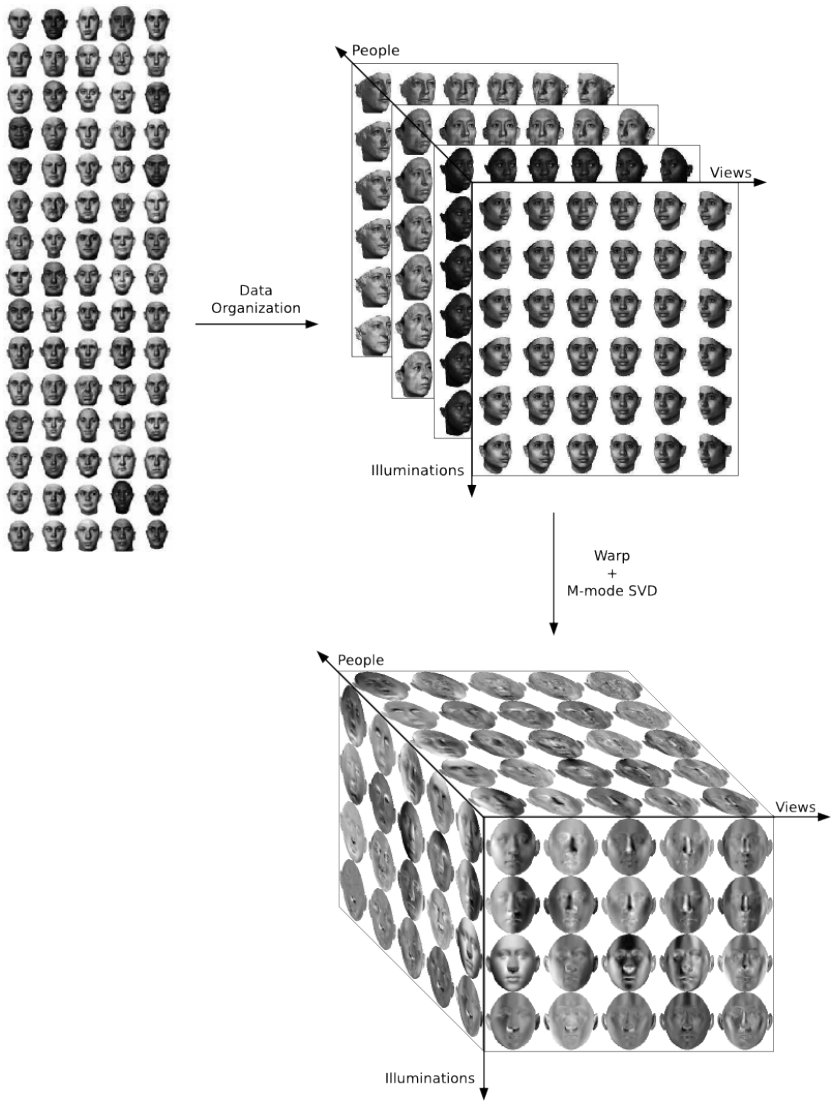


Figure 3.1: Overview of the TensorFaces approach. (Top left) Facial images rendered from a subset of the *University of Freiburg 3D Morphable Models* dataset (Blanz and Vetter, 1999). (Top right) The training image ensemble is organized into a three-way data array, where each axis corresponds to a different factor, specifically identity, viewpoint, and illumination. (Bottom) A partial visualization of the TensorFaces basis. Displayed are the principal axes of variation across each of the factors.

ranks of the orthonormal spaces.

Each row of $\mathbf{U}_{\text{P}}$ contains a *person signature* that is not only unique for each person portrayed in $\mathcal{D}$, but is also invariant to factors such as viewpoint and illumination.

The multilinear model represents each training image $\mathbf{d}_{pvl}$ via its associated set of person, view, and illumination signatures $\mathbf{p}_{\text{P}}$, $\mathbf{v}_{\text{V}}$, $\mathbf{v}_{\text{L}}$:

$$\mathbf{d}_{PVL} = \mathcal{T} \times_{\text{P}} \mathbf{p}_{\text{P}}^{\text{T}} \times_{\text{V}} \mathbf{v}_{\text{V}}^{\text{T}} \times_{\text{L}} \mathbf{l}_{\text{L}}^{\text{T}}. \tag{3.2}$$

3.1.2 Multilinear Projection

To compute the person signature of a test facial image, we must project the image into the multilinear space computed in (3.1). Following the multilinear projection algorithm from (Vasilescu, 2011), we first compute the response tensor $\mathcal{R}$ for input image $\mathbf{d}$:

$$\mathcal{R} = \mathcal{T}^{+_{\text{x}}} \times_{\text{x}}^{\text{T}} \mathbf{d}, \tag{3.3}$$

where $\mathcal{T}^{+_{\text{x}}}$ is the pseudo-inverse of the extended core tensor $\mathcal{T}$ in the pixel-mode. Next, we decompose $\mathcal{R}$ into its coefficient vectors via an alternating least-squares algorithm:

$$\mathcal{R} \approx \mathbf{r}_{\text{P}} \circ \mathbf{r}_{\text{V}} \circ \mathbf{r}_{\text{L}}, \tag{3.4}$$

where $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$ denotes the outer product. the vectors $\mathbf{r}_{\text{P}}, \mathbf{r}_{\text{V}}, \mathbf{r}_{\text{L}}$ are the estimated person, viewpoint, and illumination signatures, respectively, for image $\mathbf{d}$. The person signature $\mathbf{r}_{\text{P}}$ is used for recognition purposes.

3.1.3 Dimensionality Reduction

To perform dimensionality reduction in the multilinear space, we compute an optimal rank-$(\tilde{R}_{\text{P}}, \tilde{R}_{\text{V}}, \tilde{R}_{\text{L}})$ approximation of the training data tensor

$$\tilde{\mathcal{D}} = \mathcal{T} \times_{\text{P}} \mathbf{U}_{\text{P}} \times_{\text{V}} \mathbf{U}_{\text{V}} \times_{\text{L}} \mathbf{U}_{\text{L}}, \tag{3.5}$$

where $\mathbf{U}_{\text{P}} \in \mathbb{R}^{I_{\text{P}} \times \tilde{R}_{\text{P}}}$, $\mathbf{U}_{\text{V}} \in \mathbb{R}^{I_{\text{V}} \times \tilde{R}_{\text{V}}}$, and $\mathbf{U}_{\text{L}} \in \mathbb{R}^{I_{\text{L}} \times \tilde{R}_{\text{L}}}$ are the reduced-rank orthonormal mode matrices for the people, viewpoint, and illumination modes respectively, $\mathcal{T} \in \mathbb{R}^{\tilde{R}_{\text{P}} \times \tilde{R}_{\text{V}} \times \tilde{R}_{\text{L}} \times I_{\text{x}}}$ is the extended core tensor, and $1 \leq \tilde{R}_{\text{P}} \leq R_{\text{P}}$, $1 \leq \tilde{R}_{\text{V}} \leq R_{\text{V}}$, and $1 \leq \tilde{R}_{\text{L}} \leq R_{\text{L}}$.

To estimate the optimal rank-$(\tilde{R}_{\text{P}}, \tilde{R}_{\text{V}}, \tilde{R}_{\text{L}})$ approximation, we minimize the following error function via an alternating-least squares algorithm:

$$e = ||\mathcal{D} - \tilde{\mathcal{D}}|| + \sum_{m=1}^{M} \mathbf{\Lambda}_m ||\mathbf{U}_m^{\text{T}} \mathbf{U}_m - \mathbf{I}||, \tag{3.6}$$

where $\mathcal{D}$ is the data tensor, $\tilde{\mathcal{D}}$ is the reduced-rank approximation, $\mathbf{\Lambda}_m$ are Lagrange multiplier matrices, $||\mathbf{U}_m^{\text{T}} \mathbf{U}_m - \mathbf{I}||$ represents the orthogonality constraint on each $\mathbf{U}_m$, and m refers to the mode (people, viewpoints, and illuminations). The details are presented in Chapter 3.2.1 of (Vasilescu and Terzopoulos, 2007).

CHAPTER 4

Laplacian Pyramid TensorFaces

We now present our novel contributions to performing face verification within a hierarchical, scale-sensitive multilinear framework. Our work is primarily motivated by the following intuitions: First, facial images should be analyzed in a local part-based manner, rather than only via global analysis. This improves our model's representation power in a combinatorial manner while using the *same* amount of training data. Second, when comparing two faces, different parts of the face require different levels of detail. Third, not all facial parts should contribute equally when determining if two faces are similar.

4.1 Overview

The offline training process comprises two stages: preprocessing and model fitting. During the preprocessing stage, we align each image to a canonical coordinate system and determine the occluded regions of the face. After performing illumination normalization, we perform a hierarchical subdivision of each face image, and fit a separate TensorFaces model to each facial part. These models are used to compute a multilinear *composite facial signature* that is unique to each person, while remaining invariant to factors such as viewpoint and illumination.

To perform face verification, given a pair of test images, we preprocess the images to bring them into correspondence to our canonical coordinate system, and determine the occluded regions. We then compute the composite facial signature $\mathbf{s}$ for each of the normalized images, and apply a weighted nearest-neighbor classifier to determine if the images contain the same person.

Figure 4.1: The occlusion detection algorithm. (a) The input image. (b) The result of warping the input image to the template shape using only the *unoccluded* markers. Note that the occluded landmarks (in blue) have been nudged to the border of the face by the warp. (c) The unoccluded region (in red) is the interior of the warped facial landmarks. The pixels outside of the unoccluded region will not be considered during analysis.

4.2 Occlusion Detection and Alignment

To gracefully handle cases where regions of the face are not visible, we must automatically determine the occluded regions of the face (Figure 4.1). These occluded regions are handled separately during analysis.

4.2.1 Occluded Landmarks

We first determine which facial landmarks in the input image are occluded. For instance, if the left-ear is not visible, then the left-ear landmarks should be tagged as occluded.

To make this determination, we compare the Delaunay triangulation of the input image against the triangulation of a frontal, neutral-expression template face shape. If the area of a triangle in the input triangulation is significantly smaller than the area of the corresponding triangle in the template shape, then the triangle is flagged.

After all triangles are compared, a landmark is declared as occluded if it belongs to any flagged triangle.

4.2.2 Warping

To align each image to a canonical coordinate system, we warp each image to the template shape via a piecewise-affine warp, using only the *unoccluded landmarks* to compute the warp field. To perform this warp, we compute a triangulation on the input facial landmarks, and then warp each triangle to the corresponding triangle of the template shape via an affine transformation.

Once the image has been brought into correspondence to our canonical coordinate system, we next determine which regions of the face are unoccluded. First, we apply the image-to-template warping function to the occluded markers (Section 4.2.1). The unoccluded region is defined as the interior of the warped facial landmarks, and will not be used for recognition purposes (Figure 4.1).

4.2.3 Illumination Normalization

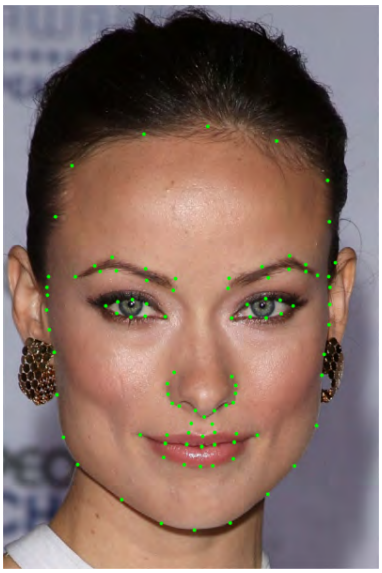
Finally, we normalize the image intensities in three stages:

1. First, we shift the image intensities such that the median intensity value is 0.5, and all intensities less than/greater than the median are stretched from $[0, 0.5]$ and $[0.5, 1.0]$ respectively. This intensity transformation corrects for global brightness changes.

2. Next, we perform contrast normalization via an adaptive contrast histogram equalization algorithm. Rather than performing contrast correction on the entire image, we normalize the contrast on subtiles of the image, and use interpolation to avoid tiling artifacts. Histogram clipping is used to avoid over-saturated regions.

3. Finally, we reapply the intensity normalization from Step 1.

We also standardize the image intensity values by subtracting the mean and dividing by the standard deviation, where the mean and standard deviation are computed from the training set.



(a) Facial landmark legend.

(b) Facial features.

Figure 4.2: (a) The facial landmark legend used for aligning images. (b) Our system builds a multilinear representation for each facial feature; *i.e.*, eyes, nose, mouth, *etc*.

4.3 Hierarchical Spatial Subdivision

In previous work, TensorFaces used a single global model to analyze facial images. By extending this approach and analyzing each facial part separately, we gain a combinatorial increase in representational power while using the *same* amount of training data. Furthermore, this representation is unique for each person, and it is invariant to factors such as viewpoint and illumination.

To implement this idea, we manually subdivided the template face shape into a series of facial parts, such as eyes, nose, mouth, *etc*. (Figure 4.2). Furthermore, we build a spatial hierarchy by recursively subdividing each part region (Figure 4.3). By doing so, we treat the face as a composite of parts, and analyze each part independently from the other parts.

During training, we perform a spatial hierarchical subdivision of the facial images, and fit a separate TensorFaces model to each facial part (Figure 4.5). These models are used to compute part-based multilinear signatures.

This part-based representation is not only far more expressive than a single global representation, but is also more robust to occlusion. For instance, suppose we have a person registered in

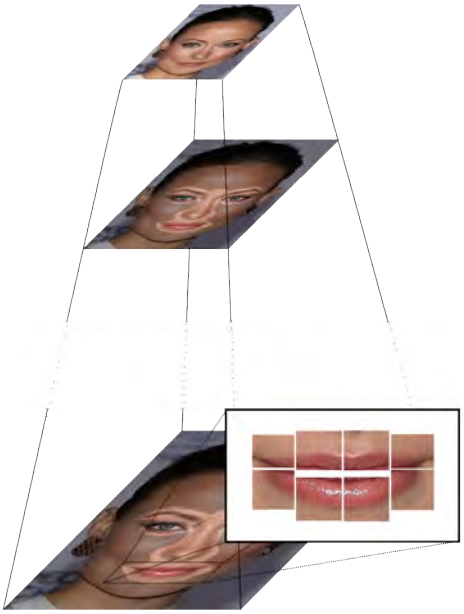


Figure 4.3: The hierarchical spatial subdivision. Each facial part is recursively subdivided into smaller subregions. For instance, the mouth part is divided into smaller subparts to capture the contours of the upper and lower lip.

our facial database, and a previously unseen incoming image of the same person arrives, where the person is wearing sunglasses. A monolithic approach, such as TensorFaces, would not gracefully handle the occlusion, as the presence of sunglasses significantly alters the global appearance of the face image. On the other hand, our part-based representation gracefully handles this occlusion. While the signatures for the eyes will be unreliable, the other facial parts will be unoccluded and unaffected for recognition purposes.

4.4 Scale-Sensitive Hierarchical Subdivision

With the current spatial subdivision, we make the implicit assumption that all parts of the face should be analyzed at the same level of detail. This assumption is incorrect. On the one hand, when analyzing global facial features, high-frequency details such as the individual hairs of the eyelashes should be suppressed, and the overall facial shape should dominate. On the other hand, when comparing two different eyes, the fine details of the eyelashes should become more important for recognition. In other words, small local regions should focus on modeling the *fine details* of

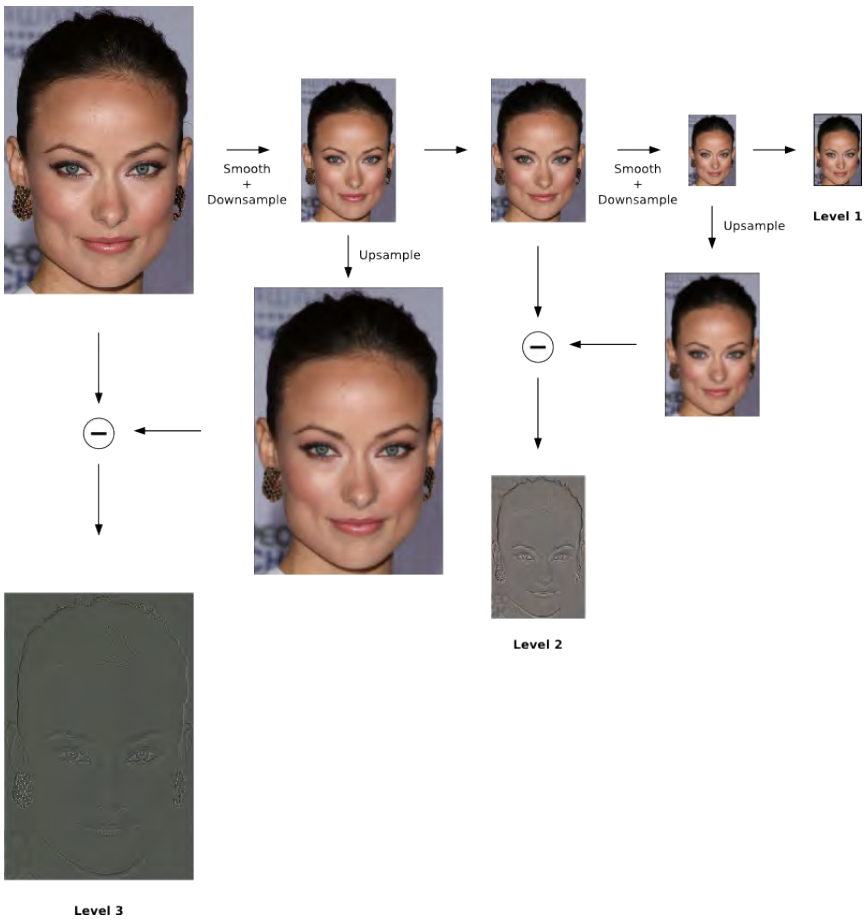


Figure 4.4: A visualization of the Laplacian pyramid construction.

the face, whereas large spatial regions should capture the *overall features* of the face.

To implement this *coarse-to-fine* approach, we compute a six-level pyramid (Laplacian or Gaussian) on both the training and test images (Figure 4.4). Each level of the pyramid corresponds to a level of our part hierarchy. For instance, the highest level (blurriest) corresponds to the part containing the entire face. The lowest level (sharpest) corresponds to the smallest local part regions; *e.g.*, the subparts of the lower/upper lips (Figure 4.3).

Our compositional model operates at different spatial regions and scales, allowing the final combined model to have a robust, unique person representation that is invariant to factors such as viewpoint and illumination.

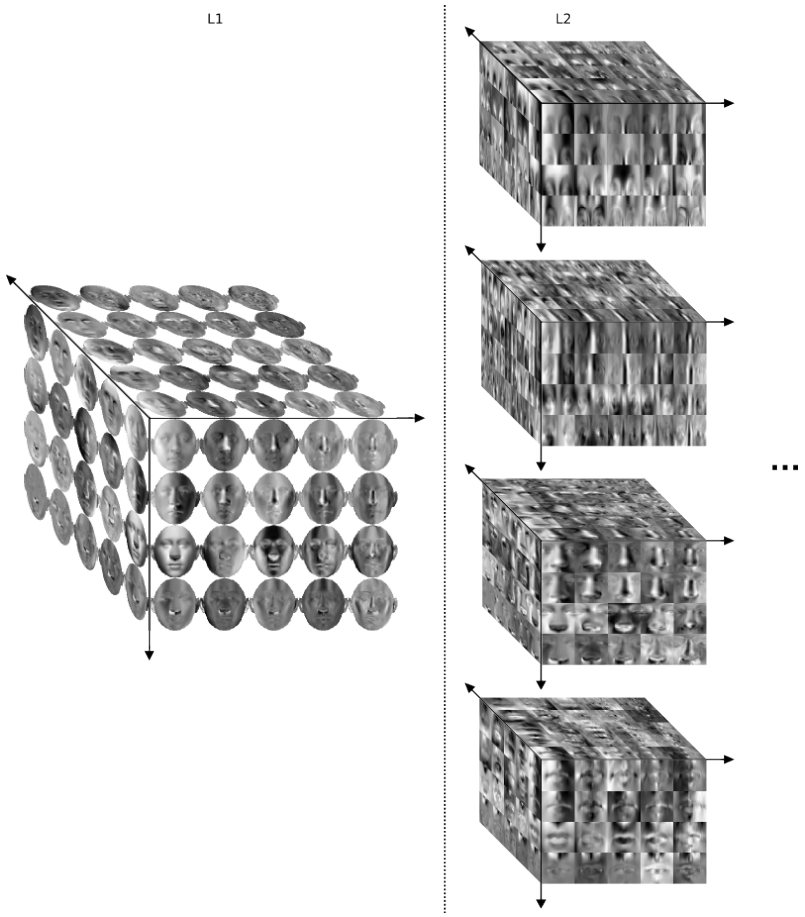


Figure 4.5: The TensorFaces basis vectors for each facial part, starting at Level 1: *full region*. At Level 2, from top to bottom: *eyebrows*, *eyes*, *nose*, and *mouth*.

4.5 Composite Facial Signature

We now describe how to compute the composite facial signature $\mathbf{s}$ of an input vectorized image $\mathbf{d}$. First, we perform a scale-sensitive hierarchical subdivision of the input image. For each facial part t, we compute its signature $\mathbf{s}_p$ using the multilinear projection algorithm (3.4) (Vasilescu and Terzopoulos, 2007). The final model representation for the set P of facial parts is the *composite facial signature* $\mathbf{s} = [\mathbf{s}_1^{\mathrm{T}} \ldots \mathbf{s}_P^{\mathrm{T}} \ldots \mathbf{s}_{|P|}^{\mathrm{T}}]^{\mathrm{T}}$, where the superscripts denote transposition; *i.e.*, the concatenation of the $|P|$ part signatures.

To ensure that occluded portions of the face are correctly handled in the analysis (Section 4.2), we impose the following constraint on $\mathbf{s}$. Let $occluded(p) \in [0, 1]$ denote the percentage of visible

pixels for facial part p. Each facial part signature $\mathbf{s}_p$ is modified as follows:

$$\mathbf{s}_p = \begin{cases} \mathbf{s}_p & \text{if } \textit{occluded}(p) \geq \alpha_{\textit{invalid}} \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{4.1}$$

The scalar $\alpha_{\textit{invalid}} \in [0, 1]$ specifies the desired tolerance for occluded pixels within a facial part. Setting $\mathbf{s}_p = \mathbf{0}$ amounts to ignoring facial part p during analysis.

4.6 Verification

Given two vectorized facial images $\mathbf{d}_1$ and $\mathbf{d}_2$ that have been preprocessed and brought into correspondence (Section 4.2), we compute their composite facial signatures $\mathbf{s}_1, \mathbf{s}_2$. To compute the similarity between $\mathbf{s}_1$ and $\mathbf{s}_2$, we compute the normalized cosine-similarity

$$f(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{|P_u|} \sum_{p \in P_u} \frac{\mathbf{s}_{1,p}^{\mathrm{T}} \mathbf{s}_{2,p}}{||\mathbf{s}_{1,p}||\,||\mathbf{s}_{2,p}||}. \tag{4.2}$$

In the summation, P_u is the set of overlapping unoccluded facial parts between images $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\mathbf{s}_{1,p}$ and $\mathbf{s}_{2,p}$ denote the signatures of facial part p in images $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

Finally, we define the decision function as

$$\textit{ismatch}(\mathbf{s}_1, \mathbf{s}_2) = \begin{cases} 1 & \text{if } f(\mathbf{s}_1, \mathbf{s}_2) \geq \tau \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

The decision threshold $\tau \in [0, 1]$ is empirically determined on a validation set separate from the training and testing sets.

4.7 Weighted Signatures

Considering that each facial part operates in a different spatial region and possibly at a different scale, it is unlikely that each facial part signature $\mathbf{s}_p$ should contribute equally to the final verification score (4.2). For instance, the eyes and mouth likely contain more discriminative information than the cheek. Following this intuition, we weight the contribution of each facial part to the final verification score, as follows:

$$f(\mathbf{s}_1, \mathbf{s}_2; \mathbf{w}) = \sum_{t \in P_u} w_p \frac{\mathbf{s}_{1,p}^{\mathrm{T}} \mathbf{s}_{2,p}}{||\mathbf{s}_{1,p}||\,||\mathbf{s}_{2,p}||}, \tag{4.4}$$

where $\mathbf{w} \in \mathbb{R}^{|P|}$ is the vector of scalar weights for each facial part, $w_p \in \mathbb{R}$ is the scalar weight for facial part p in the set P of facial parts, and $\mathbf{s}_{1,p}$ and $\mathbf{s}_{2,p}$ are the signatures for facial part p in images $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

We learn the weights $\mathbf{w}$ by minimizing the following objective function over a validation set that is separate from the training and testing set:

$$\underset{\mathbf{w}}{\text{minimize}} \; \|\mathbf{M} - \sum_{p \in P} w_p \mathbf{S}_p \mathbf{S}_p^{\mathrm{T}}\|_{\mathrm{Fro}}, \tag{4.5}$$

where $\mathbf{S}_p \in \mathbb{R}^{N \times \tilde{R}_{\mathrm{p}}}$ contains the facial part signatures of the validation set for part p, N is the number of validation images, $\tilde{R}_{\mathrm{p}}$ is the dimension of the facial part signature $\mathbf{s}_p$, the part weights vector $\mathbf{w} \in \mathbb{R}^{|P|}$ contains the scalar weights $w_p \in \mathbb{R}$ for each facial part p, and $\mathbf{M} \in \mathbb{R}^{N \times N}$ is the desired *true–false matching matrix* that stores the identities of each image in the validation set. Each row of $\mathbf{S}_p$ is a signature for facial part p, with the signatures for occluded facial parts set to $\mathbf{0}$. The desired *true–false matching matrix* $\mathbf{M}$ is defined as

$$\mathbf{M}(i,j) = \begin{cases} 1 & \text{if images } i \text{ and } j \text{ depict the same person} \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

As (4.5) is linear in the unknowns $\mathbf{w}$, we solve for the optimal $\mathbf{w}$ by computing the least-squares

solution of an equivalent linear system. We express (4.5) as

$$\underset{\mathbf{w}}{\text{minimize}} \; \|\text{vec}(\mathbf{M}) - \mathbf{A}\mathbf{w}\|_2, \tag{4.7}$$

where $\text{vec}(\mathbf{M})$ vectorizes $\mathbf{M}$ by stacking each column into a single vector, $\mathbf{w}$ contains the weights for each facial part, and $\mathbf{A}$ is the system matrix containing the vectorized $\mathbf{S}_p\mathbf{S}_p^{\mathrm{T}}$ terms:

$$\mathbf{A} = \begin{bmatrix} \vdots & \cdots & \vdots & \cdots & \vdots \\ \text{vec}(\mathbf{S}_1\mathbf{S}_1^{\mathrm{T}}) & \cdots & \text{vec}(\mathbf{S}_p\mathbf{S}_p^{\mathrm{T}}) & \cdots & \text{vec}(\mathbf{S}_p\mathbf{S}_p^{\mathrm{T}}) \\ \vdots & \cdots & \vdots & \cdots & \vdots \end{bmatrix} \tag{4.8}$$

We directly solve the system in (4.7) via the pseudoinverse

$$\mathbf{w} = \mathbf{A}^{\dagger}\text{vec}(\mathbf{M}). \tag{4.9}$$

4.8 Feature Weights

Rather than weight the composite facial signature at a part-level granularity, an alternate approach is to introduce a generalized feature weighting $\mathbf{W}$, as follows:

$$f(\mathbf{s}_1, \mathbf{s}_2; \mathbf{W}) = \mathbf{s}_1^{\mathrm{T}}\mathbf{W}\mathbf{s}_2, \tag{4.10}$$

where $\mathbf{W} \in \mathbb{R}^{(|P|\tilde{R}_{\mathrm{P}})\times(|P|\tilde{R}_{\mathrm{P}})}$ is the matrix of feature weights, $(|P|\tilde{R}_{\mathrm{P}})$ is the dimension of a composite signature, and $\mathbf{s}_1$ and $\mathbf{s}_2$ denote the composite signatures for images $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

We learn the weights $\mathbf{W}$ over a validation set that is separate from the training and test sets:

$$\underset{\mathbf{W}}{\text{minimize}} \; \|\mathbf{M} - \mathbf{S}\mathbf{W}\mathbf{S}^{\mathrm{T}}\|_{\text{Fro}}, \tag{4.11}$$

where $\mathbf{M} \in \mathbb{R}^{N\times N}$ is the true–false matching matrix (4.6), $\mathbf{S} \in \mathbb{R}^{N\times(|P|\cdot\tilde{R}_{\mathrm{P}})}$ is the matrix containing the composite facial signatures for all N validation images, and $\mathbf{W}$ is the matrix of feature weights.

While one could directly solve for $\mathbf{W}$ by solving (4.11), we found that this led to severe overfitting. Instead, by constraining $\mathbf{W}$ to be diagonal, the resulting weights generalized well to test images:

$$\underset{\mathbf{w}_{\text{feat}}}{\text{minimize}} \; \|\mathbf{M} - \mathbf{S}\,\text{diag}(\mathbf{w}_{\text{feat}})\,\mathbf{S}^{\text{T}}\|_{\text{Fro}}, \tag{4.12}$$

where $\mathbf{w}_{\text{feat}} \in \mathbb{R}^{(|P|\cdot\tilde{R}_{\text{P}})}$ contain the feature weights, and $\text{diag}(\mathbf{w}_{\text{feat}})$ constructs a square, diagonal matrix with $\mathbf{w}_{\text{feat}}$ as its main diagonal.

To efficiently find the optimal diagonal $\mathbf{w}_{\text{feat}}$, we can reduce the optimization in (4.12) to a series of simple, independent problems. First, we express (4.12) as the system

$$\mathbf{S}\,\text{diag}(\mathbf{w}_{\text{feat}}) = \mathbf{MS}(\mathbf{S}^{\text{T}}\mathbf{S})^{\dagger}. \tag{4.13}$$

Let $\mathbf{A} = \mathbf{S}$ and $\mathbf{B} = \mathbf{MS}(\mathbf{S}^{\text{T}}\mathbf{S})^{\dagger}$. Then, the i-th element of $\mathbf{w}_{\text{feat}}$ is

$$(\mathbf{w}_{\text{feat}})_i = \mathbf{b}_i^{\text{T}} \frac{\mathbf{a}_i}{\mathbf{a}_i^{\text{T}}\mathbf{a}_i}, \qquad i = 1, \ldots, (|P| \cdot \tilde{R}_{\text{P}}), \tag{4.14}$$

where $(\mathbf{w}_{\text{feat}})_i \in \mathbb{R}$ is the i-th entry of $\mathbf{w}_{\text{feat}}$, and where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the i-th columns of $\mathbf{A}$ and $\mathbf{B}$, respectively.

We chose not to apply $\mathbf{w}_{\text{feat}}$ in our experiments, opting instead to use the facial part weights $\mathbf{w}$.

CHAPTER 5

Experiments

We evaluated our approach on two datasets—one containing synthetic images, another containing real-world images. For all experiments, we used a scale-sensitive hierarchical subdivision containing 91 facial parts with 6 levels. If a facial part p has even a single invalid pixel, then we discard its facial part signature $\mathbf{s}_p$; *i.e.*, we set $\alpha_{invalid} = 0$ in (4.1).

In all experiments, we show the performance impact of each novel contribution of this work. First, we explore the effect of hierarchically subdividing the face and analyzing each facial part separately, but without computing a multiscale representation; *i.e.*, each part is represented at the original image resolution (Section 4.3). This is shown under the "Pixels" column of Table 5.1. We then examine the impact of analyzing each facial part in a scale-sensitive manner (Section 4.4) using two different multiresolution pyramids: the Gaussian pyramid, and the Laplacian pyramid. Finally, we report the performance impact of weighting each facial part's contribution to the verification score, under the "Weight" columns (Section 4.7). The performance of Principal Components Analysis (PCA) and TensorFaces are also shown as baseline results.

5.1 University of Freiburg 3D Morphable Faces

The University of Freiburg 3D Morphable Faces dataset (Blanz and Vetter, 1999) is a dataset containing 100 subjects. Each subject was rendered from 15 viewpoints and under 15 illumination conditions, generating a total of $(100 \times 15 \times 15) = 22,500$ synthetic images (Figure 5.1). All images were manually annotated with facial landmarks (Figure 4.2).

The dataset was randomly partitioned into disjoint train, validation, and test sets, where the

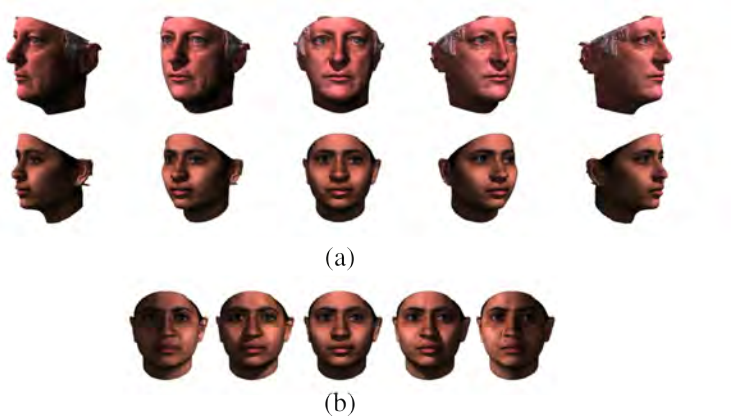


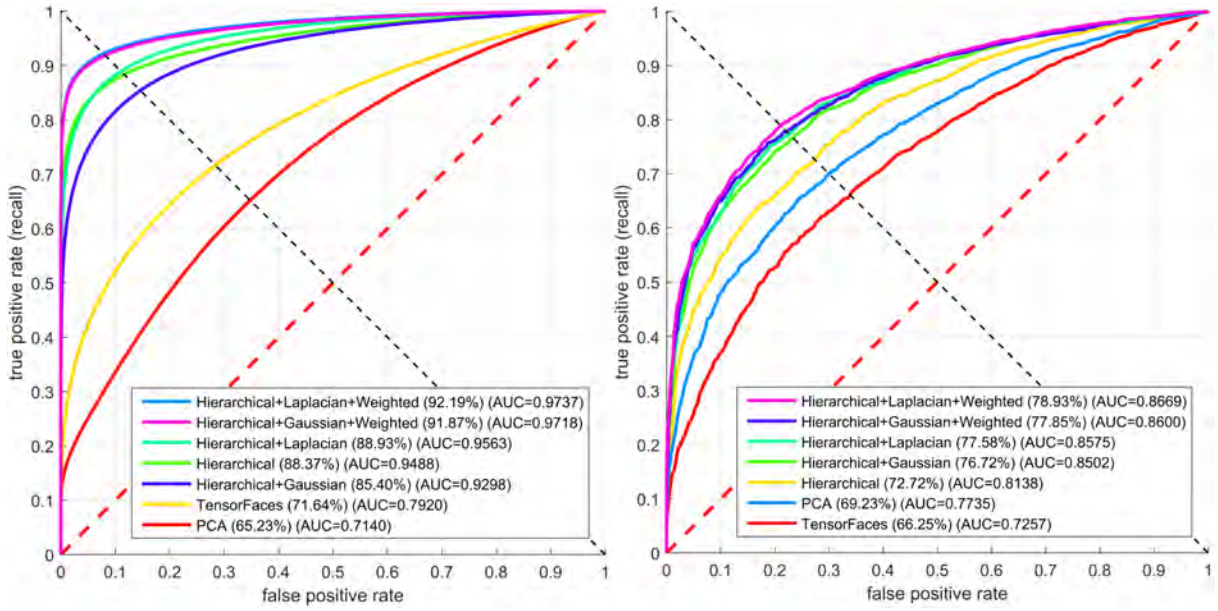
Figure 5.1: The University of Freiburg 3D Morphable Faces dataset (Blanz and Vetter, 1999). (a) Two subjects from the dataset, imaged from viewpoints $\theta = \{-60°, -30°, 0°, +30°, +60°\}$. (b) The second subject imaged under illumination conditions: $\delta = \{-60°, -30°, 0°, +30°, +60°\}$.

train and test sets do not overlap in identity, viewpoints, or illumination. 90 of the subjects were divided among the train and test sets, while preserving the gender and ethnicity distribution. The remaining 10 people were used for the validation set.

For all multilinear experiments, we performed dimensionality reduction, and retained the top 100, 6, and 2 basis vectors for the people, viewpoint, and illumination axes, respectively; *i.e.*, $\tilde{R}_{\text{P}} = 100$, $\tilde{R}_{\text{V}} = 6$, and $\tilde{R}_{\text{L}} = 2$ (Section 3.1.3). To obtain a fair comparison against PCA, we kept the top $45 \cdot 6 \cdot 2 = 540$ PCA basis vectors.

We evaluated our approach by testing on all possible test image pairs. Receiver Operating Characteristic (*ROC*) curves are shown in Figure 5.2. In Table 5.1, the reported accuracy is one minus the *equal error rate*. This is computed by finding the validation threshold τ that results in equal false-positive and false-negative rates, and this threshold is then used to compute the total test classification accuracy.

When we apply a hierarchical spatial subdivision to the facial images, we obtain a significant increase in verification performance relative to standard TensorFaces. By analyzing the face images in terms of local facial parts, the representational power of our model increases combinatorially, and we are able to recognize a greater number of unseen people with the same amount of training data. A second major jump in performance occurs when we utilize a spatial pyramid in



(a) University of Freiburg 3D Morphable Models

(b) Labeled Faces in the Wild

Figure 5.2: (Left) ROC curves for the University of Freiburg 3D Morphable Faces dataset. (Right) ROC curves for the Labeled Faces in the Wild dataset. The average accuracies are listed next to each method, along with the area under the curve (*AUC*). *Hierarchical* refers to using a Hierarchical TensorFaces model to separately analyze facial parts. *Gaussian*, *Laplacian* refers to using a Gaussian/Laplacian pyramid with a Hierarchical TensorFaces model to analyze facial parts at different scales. *"Weighted"* refers to using a weighted composite signature (Section 4.7).

Test Dataset	PCA	TensorFaces	Hierarchical TensorFaces				
			Pixels	Gaussian Pyramid	Gaussian (Weighted)	Laplacian Pyramid	Laplacian (Weighted)
Freiburg	65.23%	71.64%	90.50%	88.17%	94.17%	90.96%	93.98%
LFW	69.23% ±1.51	66.25% ±1.60	72.72% ±2.14	76.72% ±1.65	77.85% ±1.83	77.58% ±1.45	78.93% ±1.77

Table 5.1: Empirical results reported for PCA, TensorFaces and Hierarchical TensorFaces. "Pixels" denotes that each facial part is analyzed separately, but without any multiresolution pyramid. "Gaussian"/"Laplacian" use a multiresolution pyramid to analyze facial features at different scales. "Weighted" denotes that we use a weighted composite signature.
Training on Freiburg: 6 views (±60°, ±30°, ±5°) and 6 illuminations (±60°, ±30°, ±5°)
Test on Freiburg: 9 views (±50°, ±40°, ±20°, ±10°, 0°) and 9 illuminations (±50°, ±40°, ±20°, ±10°, 0°)
Test on LFW: We report the mean accuracy and standard deviation across standard literature partitions (Huang et al., 2007), following the *Unrestricted, labeled outside data* supervised protocol.

Figure 5.3: The *Labeled Faces in the Wild* (LFW) dataset (Huang et al., 2007).

combination with the weighted signatures.

5.2 Labeled Faces in the Wild

To validate the effectiveness of our system on real-world images, we report results using the "Labeled Faces in the Wild" dataset (*LFW*) (Huang et al., 2007). This dataset contains 13,233 facial images of 5,749 people (Figure 5.3). The photos are unconstrained (*i.e.*, "in the wild"), and include variation due to pose, illumination, expression, and occlusion. The dataset consists of 10 train/test splits of the data. We report the mean accuracy and standard deviation across all splits (Table 5.1, Figure 5.2). We follow the supervised *"Unrestricted, labeled outside data"* paradigm.

We used the *Dlib* (King, 2009) implementation of the automated landmark detector of Kazemi and Sullivan (2014) to obtain facial landmarks for the *LFW* images. This landmark detector applies an ensemble of regression trees to detect facial landmarks based off of image intensities. Because the *Dlib* landmark detector does not output all of the landmarks in our legend (Figure 4.2), we only use the landmarks in common when computing the warp field.

During training, we fit our models to the entirety of the 3D Morphable Faces dataset. For each of the 10 test splits, we define the validation set as the *LFW* images whose identities have no overlap with the identities of the current test split's image pairs. The validation set images are used to determine the verification threshold τ (4.3) and the weights for each facial part signature (Section 4.7). To further improve verification results, we applied a heuristic way of avoiding overfitting to the validation set by using only the positive entries in $\mathbf{w}$ when computing similarity

between images (4.4).

To determine the verification threshold τ (4.3), we find the τ that achieves the best accuracy on the *LFW* validation image pairs. Specifically, we compute the τ such that the false-positive rate and false-negative rate are equal; *i.e.*, the *equal error rate*.

For dimensionality reduction, we retained the top 100, 6, and 5 basis vectors for the people, viewpoint, and illumination modes, respectively; *i.e.*, $\tilde{R}_{\text{P}} = 100$, $\tilde{R}_{\text{V}} = 6$, and $\tilde{R}_{\text{L}} = 5$ (Section 3.1.3). For a fair comparison with PCA, we kept the top $100 \cdot 5 \cdot 6 = 3000$ PCA basis vectors.

There is a significant increase in performance when we move from the global TensorFaces model to a part-based model ("Pixels"). Because our training set has relatively few subjects—100 people—it is unlikely that a person from the *LFW* dataset can be well-represented by a global model. However, when we move to a local part-based model, our representation power increases dramatically, leading to significantly increased verification performance.

Finally, we obtain an additional boost in performance by analyzing facial images in a multiresolution pyramid in combination with weighted signatures.

CHAPTER 6

Conclusion

6.1 Summary

In this thesis, we developed a hierarchical scale-sensitive multilinear approach to face verification. Our extensions to the original TensorFaces model (Vasilescu and Terzopoulos, 2003) leads to a combinatorial gain in representational power while using the *same* amount of training data, as well as a more robust *composite facial signature*.

To summarize, the major contributions of this work are:

- The introduction of a *composite, multilinear* facial signature that hierarchically analyzes different facial parts.
- The usage of a multiresolution pyramid within our hierarchical multilinear framework to ensure different facial parts are analyzed at different levels of detail.
- The addition of a weighted composite facial signature that weights the contribution of different facial parts to the final verification score.

We reported empirical results on both synthetic and real-world data, and demonstrated its effectiveness over a previous multilinear approach to face verification.

6.2 Future Work

The following are interesting avenues for future work:

1. *Metric learning:* We intend to further investigate the estimation of a more general feature weighting $\mathbf{W}$. Equation (4.11) is an instance of *metric learning*, where the goal is to estimate a matrix $\mathbf{W}$ that improves recognition results. A good $\mathbf{W}$ will group feature vectors of the same person closely together, while simultaneously keeping feature vectors of other people apart. For instance, Guillaumin et al. (2009) solve for a positive-definite $\mathbf{W}$ to improve face recognition performance. We intend to investigate this idea further in future work.

2. *Automatic part selection:* Currently, we manually define the spatial hierarchical subdivision based on our intuition for which facial features would be useful for recognition; *i.e.*, eyes, nose, mouth, *etc*. Determining the discriminative spatial regions in an automatic way would allow our system to be easily applicable to a wider range of recognition problems.

   For instance, Felzenszwalb et al. (2010) automatically detect object subpart regions by greedily choosing rectangular areas that contain high-energy edge histograms. This heuristic manner of automatically detecting parts can easily be applied to facial images, as the areas with high gradient activity also pertain to important facial features such as the eyes, nose, mouth, *etc*.

3. *Robustness to noisy facial landmarks:* In our system, the quality of the facial landmarks is a possible point of failure. Without properly aligned images, the system will not behave as desired.

   While designing more-accurate facial landmark detectors is certainly desirable, another approach is to extract image features that are robust to slight alignment errors. For instance, edge orientation histograms are a popular image representation in modern recognition systems; *e.g.*, *SIFT* (Lowe, 1999) and *HOG* (Dalal and Triggs, 2005). These histogram image features construct a representation that has limited robustness to several types of low-level geometric and photometric transformations, including translation, in-plane rotations, and contrast changes.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 5

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: New features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop. 5

Berg, T. and Belhumeur, P. N. (2013). POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 955–962. 4

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: A CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation. 5

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co. 2, 8, 22, 23

Chen, J. C., Ranjan, R., Kumar, A., Chen, C. H., Patel, V. M., and Chellappa, R. (2015). An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 360–368. 6

Cohen, N., Sharir, O., and Shashua, A. (2015). On the expressive power of deep learning: A tensor analysis. *CoRR*, abs/1509.05009. 3

Collobert, R., Kavukcuoglu, K., and Farabet, C. (2011). Torch7: A MATLAB-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376. 5

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE. 28

de Lathauwer, L., de Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications*, 21(4):1253–1278. 3

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). On the best rank-1 and rank-$(R_1, R_2, \ldots, R_n)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342. 7

Duan, C.-H., Chiang, C.-K., and Lai, S.-H. (2013). Face verification with local sparse representation. *IEEE Signal Processing Letters*, 20(2):177–180. 4

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645. 28

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139. 5

Gao, H., Ekenel, H. K., Fischer, M., and Stiefelhagen, R. (2010). Multi-resolution local appearance-based face verification. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1501–1504. 4

Guillaumin, M., Verbeek, J., and Schmid, C. (2009). Is that you? Metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 498–505. IEEE. 28

Gutta, S., Huang, J. R., Jonathon, P., and Wechsler, H. (2000). Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *Neural Networks, IEEE Transactions on*, 11(4):948–960. 5

Hackbusch, W. and Kühn, S. (2009). A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15(5):706–722. 3

Huang, G. B., Lee, H., and Learned-Miller, E. (2012). Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. 6

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst. 2, 24, 25

Kan, M., Shan, S., Zhang, H., Lao, S., and Chen, X. (2016). Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194. 3

Kapteyn, A., Neudecker, H., and Wansbeek, T. (1986). An approach to n-mode component analysis. *Psychometrika*, 51(2):269–275. 3

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1867–1874, Washington, DC, USA. IEEE Computer Society. 25

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758. 25

Kroonenberg, P. M. and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97. 7

Lee, M. W. and Kwak, K. C. (2011). Face verification using tensor representation from video. In *Information Science and Service Science (NISS), 2011 5th International Conference on New Trends in*, volume 1, pages 151–154. 3

Li, H. and Hua, G. (2015). Hierarchical-pep model for real-world face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4055–4064. 4

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee. 28

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York, New York. 3

Su, Y., Shan, S., Chen, X., and Gao, W. (2009). Hierarchical ensemble of global and local classifiers for face recognition. *Image Processing, IEEE Transactions on*, 18(8):1885–1896. 5

Sun, Y., Wang, X., and Tang, X. (2013). Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496. 6

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. 5, 6

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311. 3, 7

Vasilescu, M. and Terzopoulos, D. (2007). Multilinear projection for appearance-based recognition in the tensor framework. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. 10, 17

Vasilescu, M. A. O. (2002). Human motion signatures: Analysis, synthesis, recognition. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 456–460 vol.3. 3

Vasilescu, M. A. O. (2009). *A Multilinear (Tensor) Algebraic Framework for Computer Graphics, Computer Vision, and Machine Learning*. PhD thesis, University of Toronto, Department of Computer Science. 7

Vasilescu, M. A. O. (2011). Multilinear projection for face recognition via canonical decomposition. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 476–483. 7, 9

Vasilescu, M. A. O. and Terzopoulos, D. (2002a). Multilinear analysis of image ensembles: TensorFaces. In *Computer VisionECCV 2002*, pages 447–460. Springer. 1, 7

Vasilescu, M. A. O. and Terzopoulos, D. (2002b). Multilinear image analysis for facial recognition. In *Pattern Recognition, International Conference on*, volume 2, pages 20511–20511. IEEE Computer Society. 3

Vasilescu, M. A. O. and Terzopoulos, D. (2003). Multilinear subspace analysis of image ensembles. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–93. IEEE. 1, 7, 27

Vasilescu, M. A. O. and Terzopoulos, D. (2004). TensorTextures: Multilinear image-based rendering. In *ACM Transactions on Graphics*, pages 336–342. 3

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154. 5

Vlasic, D., Brand, M., Pfister, H., and Popović, J. (2005). Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433. ACM. 3

Wong, Y., Harandi, M. T., Sanderson, C., and Lovell, B. C. (2012). On robust biometric identity verification via sparse encoding of faces: Holistic vs local approaches. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. 4

Xiong, C., Liu, L., Zhao, X., Yan, S., and Kim, T. K. (2016). Convolutional fusion network for face verification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):517–528. 6

Xu, Z., Chen, H., Zhu, S.-C., and Luo, J. (2008). A hierarchical compositional model for face representation and sketching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):955–969. 4