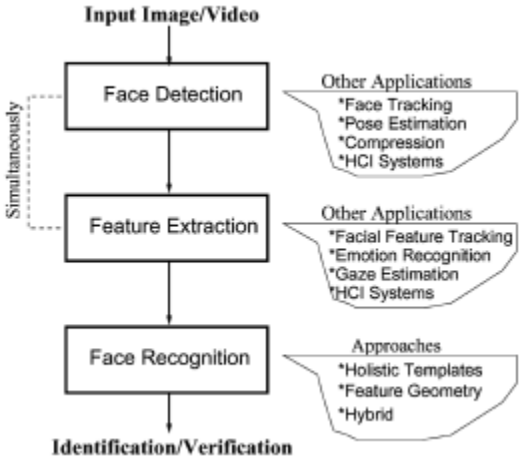
University of Waterloo
CS 489 Machine Learning W17

Literature Survey - Face Recognition

Liam Palmer (20534162) and Lucas Palmer (20534173)

Introduction

The goal of face recognition is to identify people within an image or video using a stored database of faces. The process has traditionally been broken down into three steps: face detection, feature extraction, and face recognition.



Each of these steps is critical in many algorithms that attempt to translate a raw image to a reliable identification. Sophisticated techniques have grown independently within each of these domains, as they are considered very important problems by themselves. Face detection is the process of finding the locations of faces present within images, and is used (independently from face recognition) in face tracking, compression, and many human-computer interaction systems. Feature extraction involves determining measurements that describe a given face, such as eye, nose, and mouth locations. This is used (independently from face detection) in many applications such as emotion recognition. Finally, face recognition is the process of translating various metrics from feature extraction into an identity estimation (corresponding to a given database of faces). These three steps, especially the latter two, constitute what is often referred to as "face recognition".

The research area of face recognition can be traced back to at least the 1950s in psychology. Research involving machine recognition of faces didn't emerge until the 1970s. Over the past couple decades, the problem of machine face recognition has become especially important due to the wide range of commercial and law enforcement applications. Some typical applications for machine facial recognition include virtual reality video games, human-computer interaction, voter registration, personal device login, video surveillance, and many others especially in the sectors of law enforcement, information security, and entertainment.

Although this is a literature survey, various technical aspects will be included. This technical information provides very important insights to the algorithms used for face recognition, and will allow the reader to understand the final analysis.

Techniques

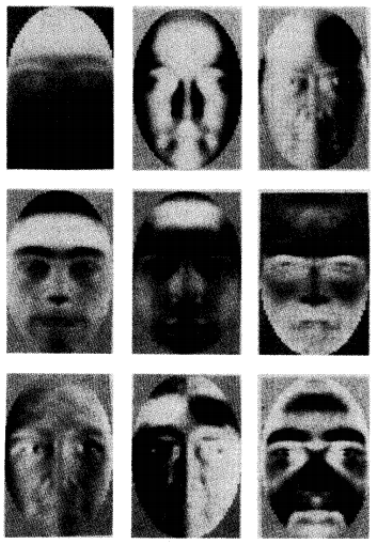
First, we will focus on **Holistic Matching Methods**. These methods treat the input facial image as raw input to a recognition system. Thus, no specific human features are extracted from the image, as the algorithm has no concept of eyes or other facial features. One of the most widely used holistic methods involves eigen-pictures and principle component analysis, developed by Kirby and Sirovich in 1990, which we present and discuss now. Specifically, we discuss the paper "Application of the Karhunen-Lokve Procedure for the Characterization of Human Faces" by M. Kirby and L. Sirovich [1].

The goal of this approach is to represent an image of a face in terms on a basis of more general images, which are referred to as eigen-pictures. If the eigen-pictures are basis images that span the images of all faces, one could represent new images in this basis and compare the eigen-picture coefficients for face recognition.

An image of a face is represented as a pixel-map $\rho(x, y)$, which maps coordinates (x, y) in an image to a grey-scale intensity ranging from $0-255$. $x = 0$ is the vertical midline of the pixel representation. If we have M images in the face database (training set), we first extend this set to the $2M$ images $\cup_{n=1}^{M}(\rho_n(x, y) \cup \rho_n(-x, y))$ where ρ_n is the nth image in the face database. We have simply extended the initial training set to include the initial images reflected about the midline $x = 0$. The average face is then given by, $\overline{\rho} = \frac{1}{2M}\sum_{n=1}^{M}(\rho_n(x, y) + \rho_n(-x, y))$. We can then translate the initial M training images to M mean subtracted images $\phi_n = \rho_n - \overline{\rho}$, often referred to as caricatures (this has proven to be effective in practice).

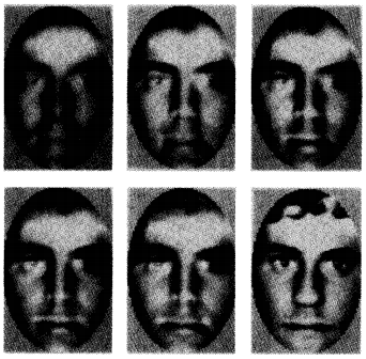
From left to right: sample face, average face, caricature of sample face.

We now consider analytical methods to translate the M images ϕ_n into a basis. Let it be noted that by taking a digital photograph of a human face in the first place, we create an upper bound on the dimensionality of this space, namely the number of pixels in the image (in this analysis 128x128). We aim to improve on this crude estimate when choosing our basis functions. We consider the Karhunen-Lohe eigen-function expansion to generate basis functions. Specifically, our basis elements will consist of the eigenfunctions u_n of the equation $Cu_n = \lambda_n u_n$ with $C(x, y, x', y') = \frac{1}{2M}\sum_{n=1}^{M}(\phi_n(x, y)\phi_n(x', y') + \phi_n(-x, y)\phi_n(-x', y'))$. To make sense of the eigen-function equation, one may think of u_n as a 128x128 2D vector, while C is a 128x128x128x128 four-tensor that multiples with u_n to create a 128x128 vector. After various linear algebra (see Kirby and Sirovich [1]), this problem can be reduced to solving the eigen-vector problem with an MxM matrix, which can be done with various computational techniques in $O(M^3)$ time. Because of the way we enhanced the training set of images, all eigen-functions are even or odd about the vertical midline, as seen to the right.



The first nine eigen-pictures, ordered by highest eigenvalue from left to right, top to bottom.

Thus, for any member of the training set ρ, we can write the following: $\rho = \overline{\rho} + \sum_{n=1}^{2M} a_n u_n$ with $a_n =< u_n, \rho - \overline{\rho} >$ (<> denotes the usual dot product). Truncating this sum to only several terms, with the eigenfunctions of greatest eigenvalues taking precedence, can still accurately reconstruct an image. Specifically, performing this truncation can accurately reconstruct images that do not reside in the training set, as seen to the right.



Approximations of the exact caricature (lower right) using 10, 20, 30, 40, and 50 eigenvectors.

Given that we have constructed a small set of basis functions that seem to accurately represent the space of human face images, we can consider the step of face recognition for new face images. This is often done by taking a new face image and projecting it into the computed basis (ie. computing the coefficients on the eigen-vectors using $a_n =< u_n, \rho - \overline{\rho} >$). Images of the same face will likely have very similar coefficients in the basis. Thus, face similarity can be measured by comparing basis coefficients with some measure (ie. Euclidean Norm). This method has achieved upwards of 70% test accuracy when using only 29 eigenvectors on a training set on 100 images.

The idea of eigen-faces has been expanded greatly since Kirby and Sirovich [1], notably by Pentland, Moghaddam, and Starner [2]. They expanded the training set to 7562 images of approximately 3000 people, used a representative 128 faces to construct the eigen-faces, and used the first 20 eigen-faces for the basis. They also considered technical variations such as eigen-features (where eigen-faces are expanded to eigen-eyes, eigen-noses, and eigen-mouths), and modular eigenspaces. By taking more modern approaches and significantly increasing the data resources, Pentland, Moghaddam, and Starner [2] achieved an accuracy of 95% for the traditional eigen-face method, and slightly higher for the eigen-features and modular eigenspace variations. However, it should be noted that all images in this analysis appear to be taken with little variation in viewpoint and lighting (ie. the facial images are very structured).

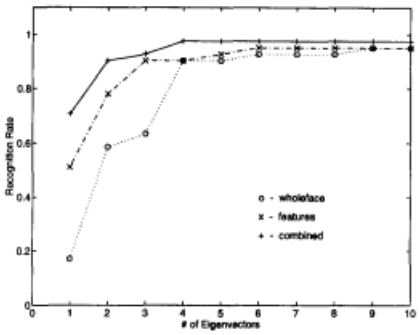


Figure 7: Recognition rates for eigenfaces, eigenfeatures and the combined modular representation.

Next, we will focus on **Feature-Based Matching Methods**. In these methods, facial structures such as the eyes, nose, and mouth are extracted from the image and analyzed before being passed to some structural classifier. In principle, these methods can be made invariant to scale, rotation, and lighting variations, which is very important in many applications. Here, we focus on the paper "Feature-Based Face Recognition Using Mixture-Distance" by I. Cox, J. Ghosn P. Yianilos [3]. The analysis of a feature-based matching method is often broken down into definition of a feature set, extraction of a feature set, and the recognition algorithm. However, Cox, Ghosn and Yianilos manually generated the feature set and focused on the recognition algorithm performance.

In this pure geometry method, the feature set consists of 30 distances derived from 35 measured locations. Because our focus is on the recognition algorithm, we consider the following problem. Given a database of facial feature vectors $Y = \{y_i\}$, and the facial feature vector q of an unidentified person, what is the identity of this person? First, we think of

any facial feature vector as a product of two phases P and O. The P phase (for platonic) generates an idealized feature vector for a specific human, which represents the actual facial measurements of that person. The O phase (for observation) generates the vectors we ultimately observe by accounting for intraclass variation (variation within images of the same person) which can be generated by a variety of factors including viewpoint, lighting, and emotion expressions. We assume that the O process is a zero mean Gaussian process that is distinct for each person. Thus, the probability that an observed feature vector q was generated by the platonic p is $Pr(q|p) = O_p(q-p)$ (the Gaussian process O for person p applied to the difference in feature vectors q and p). We also assume that the database feature vectors are platonic, that is, a query is an observation of the database elements. Thus, once we generate the processes O_p, we can guess the identity by $argmax_p(Pr(q|p)) = argmax_p(O_p(q-p))$ where q is the query and p represents the set of platonic feature vectors in the database.

We use the mixture-distance method to generate O_p for each feature vector p in the database. A finite mixture model M is a sum of probability models M_i weighted by non-negative mixing parameters c_i that sum to 1. Namely, $M(x) = \sum_{k=1}^{m} c_k M_k(x)$ with $\sum_{k=1}^{m} c_k = 1$. Let $N_{\Sigma,\mu}(x)$ denote the multivariate normal density with covariance Σ and mean μ. When each M_i is a Gaussian distribution, $M(x)$ is referred to as a Gaussian Mixture. Given vectors $x_1, x_2, ..., x_m$, the task of estimating the parameters of a Gaussian Mixture (μ_i, Σ_i, c_i) is well studied, and we adopt the Expectation Maximization method here. Once an n-element Gaussian Mixture Model M is generated by the Expectation Maximization method using the set $Y = \{y_i\}$ (feature vectors in the database), we obtain that $Pr(q|y_i, M) = \sum_{k=1}^{n} Pr(q - y_i|\overline{M_k})Pr(M_k|y_i)$ for every database feature vector y_i with $Pr(x|\overline{M_k}) = Pr(x + \mu_k|M_k)$. (For a lengthy calculation, see [3].) Thus, $Pr(q|y_i, M)$ is a sum of zero mean Gaussian distributions, which implies it is a zero mean distribution itself, and thus can be used to define the distribution O_i for each y_i.

Empirically, this method was able to achieve accuracy results in excess of 95% when altered with several variations such as the number of mixtures in M and the underlying models of each M_k. Once again, the problem of extracting facial features from images is done manually in this experiment, and only deals with predominantly frontal images.

Wiskott et al. [4] provides an analysis of face recognition with Elastic Bunch Graph Matching. This strategy performs feature extraction computationally using Gabor Wavelet functions, and accommodates a great variation in pose and facial position. It records as high as 99% accurate when comparing different frontal images with similar lighting and texture, and as low as 27% accuracy when comparing profile images with half-profile images (strong variation in face orientation).

Finally, we present an example of a **Hybrid Method**, which uses a combination of holistic methods and feature-based methods. Jennifer Huang [5] uses Support Vector Machines and 3D Morphable Models to achieve 90% accuracy in recognition with image databases that vary considerably in terms of both image orientation and image lighting. This is an example of component based analysis, where smaller components of an image such as the eyes, nose, or mouth are analyzed in a more holistic way (with support vector machines in this case).

Despite some very high accuracies in many constrained cases, none of the previous methods perform very well when recognizing faces in unconstrained images. Facial recognition in unconstrained images is very useful in many applications, especially in sophisticated surveillance and law domains. A technology developed at Facebook called **DeepFace** is arguably one of the best performing algorithms in unconstrained facial recognition, reaching an accuracy of 97.35% on the "Labelled Faces of the Wild" dataset (almost reaching human performance). This dataset consists of over 4 million

images consisting of over four thousand distinct identities.

DeepFace's algorithm consists of two main components. First, 3D face modelling and affine transformations are used to align faces from images to a certain standard. Second, a representation for a given face is computed from a nine-layer deep neural network with over 120 million parameters. This strategy benefits greatly from having vast amounts of training data, and likely won't be nearly as effective when working with significantly smaller training sets. Due to the advanced 3D face modelling technologies used, in addition to the holistic, data-oriented nature of deep neural networks, DeepFace is best classified as a highly advanced hybrid algorithm.

Comparison Summary

The most significant measures of success we have discussed include both accuracy in constrained and unconstrained environments, in addition to time constraints.

In the context of small and constrained data sets on the order of 100-1000 images, holistic matching methods have the advantage of high accuracy and fast computation. Feature based matching methods also provide highly accurate and fast classification on the order of 90%. The two hybrid techniques discussed would likely not be practical in this case, due to the fact that little advantages are obtained from feature-based analysis (due to the constraints) and that the machine learning component often relies on significantly larger test data sets to become highly accurate. DeepFace's neural network with 120 million parameters likely has too much expressivity to be reliably trained on a small data set. In the context of large constrained data sets, DeepFace's high expressivity would likely produce the highest accuracy. Some holistic matching methods, such as the one discussed in [1], would not be feasible due to the runtime constraints ([1] was cubic in the number of images).

Next, we consider the context of unconstrained data sets. Most holistic matching methods fail greatly in this context because treating the image as a whole becomes unreliable when the same person is represented in many vastly different whole images. As it was evident in the analysis of [1], structured images allowed a basis to be extracted from the data set. This task becomes significantly more complicated when structure is lost. It appears evident that at least some form of feature based matching is essential for high accuracy, as feature extraction essentially aligns a face and generates some more structured representation. Hybrid methods like DeepFace will surely perform best in this case, especially if there is lots of data in the training set.

In summary, the effectiveness of methods depends greatly on the application qualities. For identifying citizens of the United States in surveillance video for criminal investigations, DeepMind would likely be the best option due to the vast number of possible identifications and large training set that can be used to train a neural network offline (a one-time processing job). For testing identity within a small subset of people, a more lightweight holistic or feature-based technique would likely be the most practical.

Analysis

The state of the art face recognition algorithms are most certainly a combination of advanced feature extraction (often 3D modelling, used to combat unconstrained data sets) and very expressive machine learning techniques (deep neural networks, convolutional neural networks) that thrive on large data sets. These state of the art techniques are often being developed at large technology companies with significant research resources such as Facebook (DeepFace) and Google (FaceNet).

Now we discuss some general open problems in the domain of face recognition. To start, many of the most important algorithms in successful face recognition, especially deep and convolutional neural networks, are not very well explained with theory. For example, the 120 million parameter deep neural network called DeepFace developed by Facebook was not created and optimized with strong theory. Trial and error is ofter a major factor when trying to tune these machine learning algorithms to achieve higher accuracies. In fact, any open problems in classification machine learning are likely open problems in face recognition as well, as the two domains are tightly linked (face recognition techniques are just classification machine learning algorithms aside from any facial specific computations and feature extraction). As face recognition methods push towards human capabilities on many large data sets, being able to obtain and test on even larger data sets is often a roadblock.

The question of how humans themselves achieve face recognition is still an open problem in many fields such as psychology and computer science. Once again, this problem manifests itself in every machine learning problem application, as we do not know the intricacies or workings of the human brain when it is applied to any substantial task.

Conclusion

Overall, most face recognition methods are contained within the broad categories of Face Detection, Feature Extraction, and Face Recognition. Each of these categories are interesting problems on their own, and thus plenty of research has been conducted in each. Holistic methods process a facial image as a whole, and most likely do not even consider the image as a face (on the algorithm level). Feature extraction methods recognize an image as a face, and extract many human-like qualities from the image (eye locations, shape of nose etc.). Hybrid methods contain both feature-based and holistic strategies. The most advanced and successful face recognition algorithms consist of advanced feature extraction, often consisting of 3D modelling, followed by highly expressive machine learning classification algorithms such as deep and convolutional neural networks. This is evident in both DeepFace (developed by Facebook) and FaceNet (developed by Google), both of which have set record benchmarks in many categories.

We recommend future research be focused on general machine learning classification techniques and feature extraction algorithms. These are the two main components of the most successful face recognition methods today. More advanced general machine learning classification techniques would have an immediate application in face recognition and many other domains. Feature extraction helps alleviate arguably the biggest difficultly in face recognition which is unconstrained data; that is, the same human face can express a multitude of shapes and emotions while being positioned anywhere in 3D space relative to a camera.

Additional biological and psychological research would likely be valuable if it allowed us to copy off the brain when developing our own strategies. However, it is not likely that biological advancements will make significant improvements to machine face recognition any time soon.

Papers:

[1] M. Kirby and L. Sirovich. Application of the Karhunen-Lokve Procedure for the Characterization of Human Faces. IEEE Transactions on pattern analysis and machine intelligence. VOL. 12, NO. I, JANUARY 1990
URL: http://members.cbio.mines-paristech.fr/ jvert/svn/bibli/local/Kirby1990Application.pdf

[2] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In Computer Vision and PatternRecognition, pages 84-91, 1994.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=323814

[3] I. Cox, J. Ghosn, and P. Yianilos. Feature-Based Face Recognition Using Mixture-Distance. NEC Research Institute, Technical Report 95-09.
URL: https://www.researchgate.net/publication/3637747_Feature-Based_Face_Recognition_Using_Mixture-Distance

[4] L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg. Face Recognition by Elastic Bunch Graph Matching. Institut fur Neuroinformatik, Internal Report 96-08.
URL: http://www.cs.unsyiah.ac.id/ frdaus/PenelusuranInformasi/tugas2/data/irini96-08.pdf

[5] J. Huang. Component-based Face Recognition with 3D Morphable Models. Department of Electrical Engineering and Computer Science, MIT.
URL: https://pdfs.semanticscholar.org/079a/0a3bf5200994e1f972b1b9197bf2f90e87d4.pdf

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. URL: https://www.cs.toronto.edu/ ranzato/publications/taigman_cvpr14.pdf