Robust Face Recognition via Sparse Representation

-- A Q&A about the recent advances in face recognition

and how to protect your facial identity

Allen Y. Yang (yang@eecs.berkeley.edu)
Department of EECS, UC Berkeley

July 21, 2008

Q: What is this technique all about?
A: The technique, called *robust face recognition via sparse representation*, provides a new solution to use computer program to classify human identity using frontal facial images, i.e., the well-known problem of face recognition.

Face recognition has been one of the most extensively studied problems in the area of artificial intelligence and computer vision. Its applications include human-computer interaction, multimedia data compression, and security, to name a few. The significance of face recognition is also highlighted by a contrast between human's high accuracy to recognize face images under various conditions and the computer's historical poor accuracy.

This technique proposes a highly accurate recognition framework. The extensive experiment has shown the method can achieve similar recognition accuracy as human vision, for the first time. In some cases, the method has outperformed what human vision can achieve in face recognition.

Q: Who are the authors of this technique?
A: The technique was developed in 2007 by Mr. John Wright, Dr. Allen Y. Yang, Dr. S. Shankar Sastry, and Dr. Yi Ma.

The technique is jointly owned by the University of Illinois and the University of California, Berkeley. A provisional US patent has been filed in 2008. The technique is also being published in the IEEE Transactions on Pattern Analysis and Machine Intelligence [Wright 2008].

Q: Why is face recognition difficult for computers?
A: There are several issues that have historically hindered the improvement of face recognition in computer science.

1. High dimensionality, namely, the data size is large for face images.

When we take a picture of a face, the face image under certain color metrics will be stored as an image file on a computer, e.g., the image shown in Figure 1. Because the human brain is a massive parallel processor, it can quickly process a 2-D image and match the image with the other images learned in the past. However, the modern computer algorithms can only process 2-D images sequentially, meaning, it can only process an image pixel-by-pixel. Hence although the image file usually only takes less than 100 K Bytes to store on computer, if we treat each image as a sample point, it sits in a space of more than 10-100 K dimension (that is each pixel owns an individual dimension). Any pattern recognition problem in high-dimensional space (>100 D) is known to be difficult in the literature.

Fig. 1. A frontal face image from the AR database [Martinez 1998]. The size of a JPEG file for this image is typically about 60 Kbytes.

2. The number of identities to classify is high.

To make the situation worse, an adult human being can learn to recognize thousands if not tens of thousands of different human faces over the span of his/her life. To ask a computer to match the similar ability, it has to first store tens of thousands of learned face images, which in the literature is called the training images. Then using whatever algorithm, the computer has to process the massive data and quickly identify a correct person using a new face image, which is called the test image.

Fig. 2. An ensemble of 28 individuals in the Yale B database [Lee 2005]. A typical face recognition system needs to recognition 10-100 times more individuals. Arguably an adult can recognize thousands times more individuals in daily life.

Combine the above two problems, we are solving a pattern recognition problem to carefully partition a high-dimensional data space into thousands of domains, each domain represents the possible appearance of an individual's face images.

3. Face recognition has to be performed under various real-world conditions.

When you walk into a drug store to take a passport photo, you would usually be asked to pose a frontal, neutral expression in order to be qualified for a good passport photo. The store associate will also control the photo resolution, background, and lighting condition by using a uniform color screen and flash light. However in the real world, a computer program is asked to identify humans without all the above constraints. Although past solutions exist to achieve recognition under very limited relaxation of the constraints, to this day, none of the algorithms can answer all the possible challenges, including this technique we present.

To further motivate the issue, human vision can accurately recognize learned human faces under different expressions, backgrounds, poses, and resolutions [Sinha 2006]. With professional training, humans can also identify face images with facial disguise. Figure 3 demonstrates this ability using images of Abraham Lincoln.

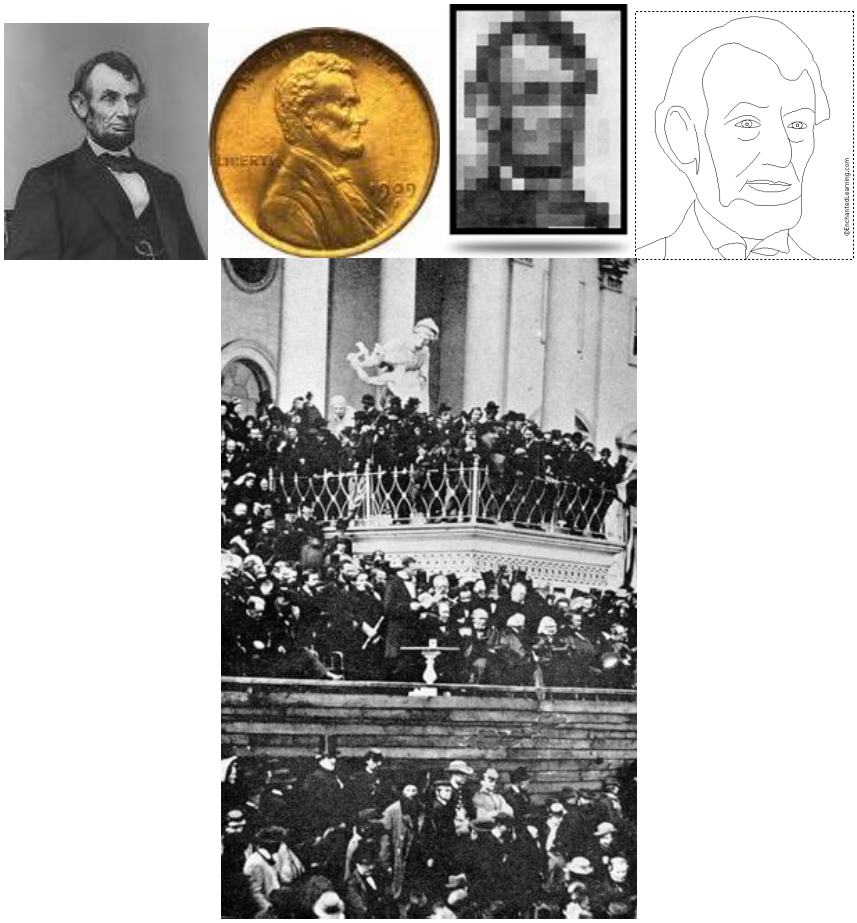

Fig. 3. Images of Abraham Lincoln under various conditions (available online). Arguably humans can recognize the identity of Lincoln from each of these images.

A natural question arises: Do we simply ask too much for a computer algorithm to achieve? For some applications such as at security check-points, we can mandate individuals to pose a frontal, neural face in order to be identified. However, in most other applications, this requirement is simply not practical. For example, we may want to search our photo albums to find all the images that contain our best friends

under normal indoor/outdoor conditions, or we may need to identify a criminal suspect from a murky, low-resolution hidden camera who would naturally try to disguise his identity. Therefore, the study to recognize human faces under real-world conditions is motivated not only by pure scientific rigor, but also by urgent demands from practical applications.

Q: What is the novelty of this technique? Why is the method related to sparse representation?

A: The method is built on a novel pattern recognition framework, which relies on a scientific concept called sparse representation. In fact, sparse representation is not a new topic in many scientific areas. Particularly in human perception, scientists have discovered that accurate low-level and mid-level visual perceptions are a result of sparse representation of visual patterns using highly redundant visual neurons [Olshausen 1997, Serre 2006].

Without diving into technical detail, let us consider an analogue. Assume that a normal individual, Tom, is very good at identifying different types of fruit juice such as orange juice, apple juice, lemon juice, and grape juice. Now he is asked to identify the ingredients of a fruit punch, which contains an unknown mixture of drinks. Tom discovers that when the ingredients of the punch are highly concentrated on a single type of juice (e.g., 95% orange juice), he will have no difficulty in identifying the dominant ingredient. On the other hand, when the punch is a largely even mixture of multiple drinks (e.g., 33% orange, 33% apple, and 33% grape), he has the most difficulty in identifying the individual ingredients. In this example, a fruit punch drink can be represented as a sum of the amounts of individual fruit drinks. We say such representation is *sparse* if the majority of the juice comes from a single fruit type. Conversely, we say the representation is not sparse. Clearly in this example, sparse representation leads to easier and more accurate recognition than nonsparse representation.

The human brain turns out to be an excellent machine in calculation of sparse representation from biological sensors. In face recognition, when a new image is presented in front of the eyes, the visual cortex immediately calculates a representation of the face image based on all the prior face images it remembers from the past. However, such representation is believed to be only sparse in human visual cortex. For example, although Tom remembers thousands of individuals, when he is given a photo of his friend, Jerry, he will assert that the photo is an image of Jerry. His perception does not attempt to calculate the similarity of Jerry's photo with all the images from other individuals. On the other hand, with the help of image-editing software such as Photoshop, an engineer now can seamlessly combine facial features from multiple individuals into a single new image. In this case, a typical human would assert that he/she cannot recognize the new image, rather than analytically calculating the percentage of similarities with multiple individuals (e.g., 33% Tom, 33% Jerry, 33% Tyke) [Sinha 2006].

Q: What are the conditions that the technique applies to?

A: Currently, the technique has been successfully demonstrated to classify frontal face images under different expressions, lighting conditions, resolutions, and severe facial disguise and image distortion. We believe it is one of the most comprehensive solutions in face recognition, and definitely one of the most accurate.

Further study is required to establish a relation, if any, between sparse representation and face images with pose variations.

Q: More technically, how does the algorithm estimate a sparse representation using face images? Why do the other methods fail in this respect?
A: This technique has demonstrated the first solution in the literature to explicitly calculate sparse representation for the purpose of image-based pattern recognition. It is hard to say that the other extant methods have failed in this respect. Why? Simply because previously investigators did not realize the importance of sparse representation in human vision and computer vision for the purpose of classification.

For example, a well-known solution to face recognition is called the *nearest-neighbor* method. It compares the similarity between a test image with all individual training images separately. Figure 4 shows an illustration of the similarity measurement. The nearest-neighbor method identifies the test image with a training image that is most similar to the test image. Hence the method is called the nearest neighbor. We can easily observe that the so-estimated representation is not sparse. This is because a single face image can be similar to multiple images in terms of its RGB pixel values. Therefore, an accurate classification based on this type of metrics is known to be difficult.

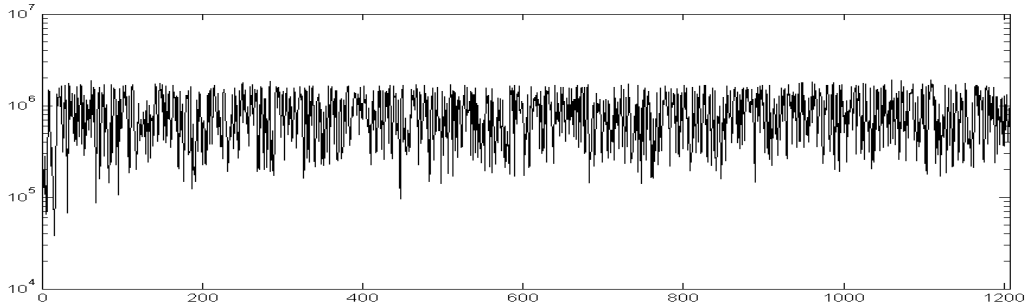

Fig. 4. A similarity metric (the y-axis) between a test face image and about 1200 training images. The smaller the metric value, the more similar between two images.

Our technique abandons the conventional wisdom to compare any similarity between the test image and individual training images or individual training classes. Rather, the algorithm attempts to calculate a representation of the input image w.r.t. all available training images as a whole. Furthermore, the method imposes one extra constraint that the optimal representation should use the smallest number of training images. Hence, the majority of the coefficients in the representation should be zero, and the representation is sparse (as shown in Figure 5).

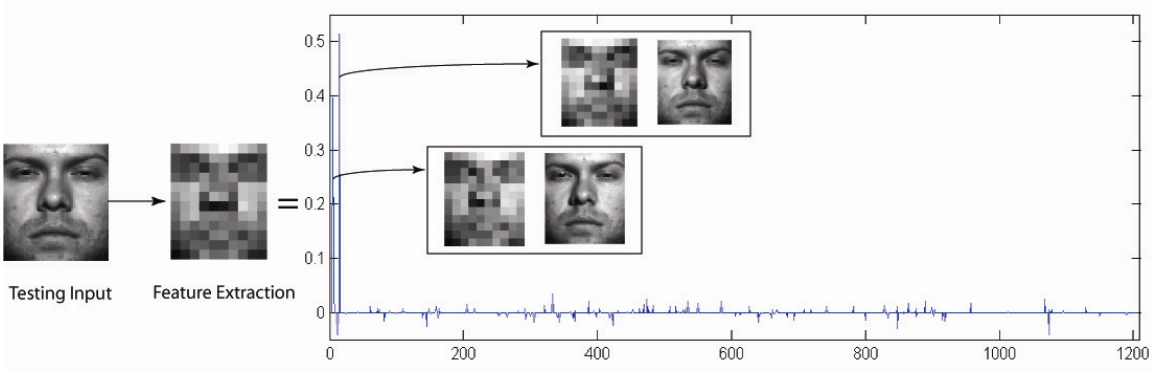

Fig. 5. An estimation of sparse representation w.r.t. a test image and about 1200 training images. The dominant coefficients in the representation correspond to the training images with the same identity as the input image. In this example, the recognition is based on downgraded 12-by-10 low-resolution images. Yet, the algorithm can correctly identify the input image as Subject 1.

Q: How does the technique handle severe facial disguise in the image?
A: Facial disguise and image distortion pose one of the biggest challenges that affect the accuracy of face recognition. The types of distortion that can be applied to face images are manifold. Figure 6 shows some of the examples.

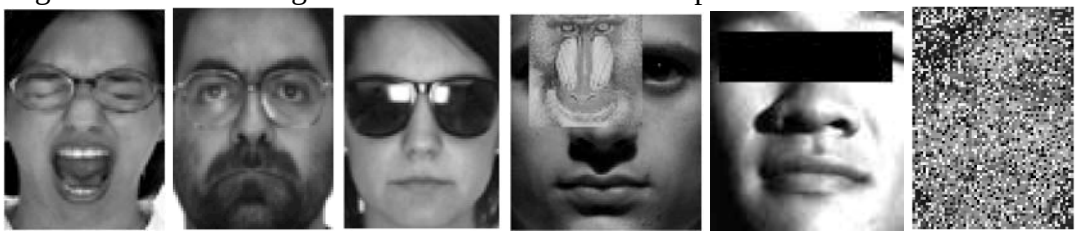
Fig. 6. Examples of image distortion on face images. Some of the cases are beyond human's ability to perform reliable recognition.

One of the notable advantages about the sparse representation framework is that the problem of image compensation on distortion combined with face recognition can be rigorously reformulated under the same framework. In this case, a distorted face image presents two types of sparsity: one representing the location of the distorted pixels in the image; and the other representing the identity of the subject as before. Our technique has been shown to be able to handle and eliminate all the above image distortion in Figure 6 while maintaining high accuracy. In the following, we present an example to illustrate a simplified solution for one type of distortion. For more detail, please refer to our paper [Wright 2008].

Figure 7 demonstrates the process of an algorithm to recognize a face image with severe facial disguise by sunglasses. The algorithm first partitions the left test image into eight local regions, and individually recovers a sparse representation per region. Notice that with the sunglasses occluding the eye regions, the corresponding representations from these regions do not provide correct classification. However, when we look at the overall classification result over all regions, the nonocclused regions provide a high consensus for the image to be classified as Subject 1 (as shown

in red circles in the figure). Therefore, the algorithm simultaneously recovers the subject identity and the facial regions that are being disguised.

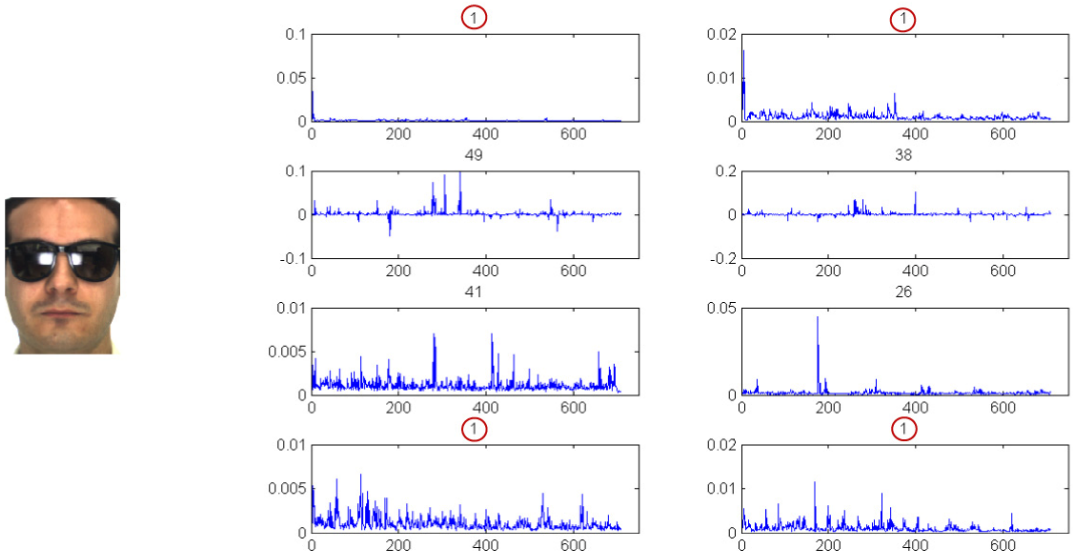

Fig. 7. Solving for part-based sparse representation using local face regions. Left: Test image. Right: Estimation of sparse representation and the corresponding classification on the titles. The red circle identifies the correct classiciation.

Q: What is the quantitative performance of this technique?
A: Most of the representative results from our extensive experiment have been documented in our paper [Wright 2008]. The experiment was based on two established face recognition databases, namely, the Extended Yale B database [Lee 2005] and the AR database [Martinez 1998].

In the following, we highlight some of the notable results. On the Extended Yale B database, the algorithm achieved 92.1% accuracy using 12-by-10 resolution images, 93.7% using single-eye-region images, and 98.3% using mouth-region images. On the AR database, the algorithm achieves 97.5% accuracy on face images with sunglasses disguise, and 93.5% with scarf disguise.

Q: Does the estimation of sparse representation cost more computation and time compared to other methods?
A: The complexity and speed of an algorithm are important to the extent that they do not hinder the application of the algorithm to real-world problems. Our technique uses some of the best-studied numerical routines in the literature, namely, L-1 minimization to be specific. The routines belong to a family of optimization algorithms called convex optimization, which have been known to be extremely efficient to solve on computer. In addition, considering the rapid growth of the technology in producing advanced micro processors today, we do not believe there is any significant risk to implement a real-time commercial system based on this technique.

Q: With this type of highly accurate face recognition algorithm available, is it becoming more and more difficult to protect biometric information and personal privacy in urban environments and on the Internet?
A: Believe it or not, a government agency, a company, or even a total stranger can capture and permanently log your biometric identity, including your facial identity, much easier than you can imagine. Based on a Time magazine report [Grose 2008], a resident living or working in London will likely be captured on camera 300 times per day! One can believe other people living in other western metropolitan cities are enjoying similar "free services." If you like to stay indoor and blog on the Internet, your public photo albums can be easily accessed over the nonprotected websites, and probably have been permanently logged by search engines such as Google and Yahoo!.

With the ubiquitous camera technologies today, completely preventing your facial identity from being obtained by others is difficult, unless you would never step into a downtown area in big cities and never apply for a driver's license. However, there are ways to prevent *illegal* and *involuntary* access to your facial identity, especially on the Internet. One simple step that everyone can choose to do to stop a third party exploring your face images online is to prevent these images from being linked to your identity. Any classification system needs a set of training images to study the possible appearance of your face. If you like to put your personal photos on your public website and frequently give away the names of the people in the photos, over time a search engine will be able to link the identities of the people with the face images in those photos. Therefore, to prevent an unauthorized party to "crawl" into your website and sip through the valuable private information, you should make these photo websites under password protection. Do not make a large amount of personal images available online without consent and at the same time provide the names of the people on the same website.

Previously we have mentioned many notable applications that involve face recognition. The technology, if properly utilized, can also revolutionize the IT industry to better protect personal privacy. For example, an assembly factory can install a network of cameras to improve the safety of the assembly line but at the same time blur out the facial images of the workers from the surveillance videos. A cellphone user who is doing teleconferencing can activate a face recognition function to only track his/her facial movements and exclude other people in the background from being transmitted to the other party. All in all, face recognition is a rigorous scientific study. Its sole purpose is to hypothesize, model, and reproduce the image-based recognition process with accuracy comparable or even superior to human perception. The scope of its final extension and impact to our society will rest on the shoulder of the government, the industry, and each of the individual end users.

References
[Grose 2008] T. Grose. *When surveillance cameras talk*. Time (online), Feb. 11, 2008.

[Lee 2005] K. Lee et al.. *Acquiring linear subspaces for face recognition under variable lighting*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, 2005.

[Martinez 1998] A. Martinez and R. Benavente. The AR face database. CVC Tech Report No. 24, 1998.

[Olshausen 1997] B. Olshausen and D. Field. *Sparse coding with an overcomplete basis set: A strategy employed by V1?* Vision Research, vol. 37, 1997.

[Serre 2006] T. Serre. *Learning a dictionary of shape-components in visual cortex: Comparison with neurons, humans and machines*. PhD dissertation, MIT, 2006.

[Sinha 2006] P. Sinha et al.. *Face recognition by humans: Nineteen results all computer vision researchers should know about*. Proceedings of the IEEE, vol. 94, no. 11, November 2006.

[Wright 2008] J. Wright et al.. Robust face recognition via sparse representation. (in press) IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.