Detecting Acromegaly: Screening for Disease with a Morphable Model

Qifeng Lu, Erik Learned-Miller, Angela Paisley,
Peter Trainer, Volker Blanz, Katrin Dedden, and Ralph Miller

Abstract

Acromegaly is a rare disorder which affects approximately 50 out of every million people. The disease is typically caused when a tumor of the pituitary gland secretes excess growth hormone, causing swelling of the hands, feet, and face, and eventually permanent changes due to abnormal bone growth in areas such as the jaw, the brow ridge, and the cheek bones. Because the disease is so rare and the onset of symptoms is slow, it is often missed by general practitioners and other health care professionals. This means that the disease often progresses beyond where it might if it were identified and treated earlier.

In this work, we consider a semi-automated approach to detecting the presence of acromegaly in people, using a supervised learning paradigm. We built a training set of images by taking frontal photographs of 24 acromegalic patients and photographs of 25 disease-free subjects. Because the recognition of acromegaly is dependent upon subtle features which are difficult to detect locally in the image, we decided upon a more global method of modelling that uses information across the entire photograph. In particular, we modelled each subject's face in an analysis-by-synthesis loop using the three-dimensional morphable face model of Blanz and Vetter [2]. The parameters of the morphable model capture many features of the three-dimensional shape of the subject's head from just a single photograph and are excellent features with which to classify subjects as either acromegalic or not. We report encouraging results of a classifier built from the training set of real human subjects.

1. Introduction

Figure 1 shows two men of the same age. For most observers, seeing either one of them alone on the street would not prompt any particular reaction. The man on the left would possibly be identified as slightly overweight but otherwise unremarkable while the man on the right appears perfectly healthy. It may be surprising to learn that these men are actually identical twins.

This photograph appeared recently in the New England Journal of Medicine as one of the journal's periodic "Medical Mysteries" with the caption, "Which twin is the patient?" [8]. As it turns out, the man on the left has a condition known as *acromegaly*, which is difficult for most laypeople, and even for most physicians, to diagnose.

Figure 1: *These men are identical twins. Which one is sick, and what is his condition? By fitting the 3D morphable face model of Blanz and Vetter [2] to each of these faces (see Figures 2 and 3), the system described in this paper correctly identified the man on the left as an acromegalic, and the man on the right as healthy. This "Medical Mystery" first appeared in the New England Journal of Medicine [8].*

Acromegaly is a rare disorder of the endocrine system which affects roughly 50 of every million people in the general population. Early detection is important in treating the disease successfully, but it is often missed because the signs are subtle and the condition is rare. Since many of the symptoms of the disease, such as swelling of the nose and growth of the jaw, affect facial appearance, the disease can be detected by experts (endocrinologists, for example) in many cases from a normal frontal photograph of a person. If a patient's appearance suggests that he or she may have the disease, additional laboratory blood tests may be performed to confirm its presence or absence. Because these tests are expensive and time consuming, it would clearly be valuable to have an inexpensive and automatic prescreening method.

The ultimate goal of our research is to develop an automated, voluntary prescreening system for acromegaly. For example, when obtaining a photograph for a driver's li-

Figure 2: *A 3D morphable model [2] was adapted in an analysis-by-synthesis loop to match the photograph of the man on the left of Figure 1. After a manual initialization (see text), the fitting of the 3D model was fully automatic. Although the model is not a perfect replica of the person in the photograph, it captures important properties of the face and skull which prove to be effective in distinghishing between patients that do or do not have acromegaly. Note that the 3D model captured the swelling of the man's nose, a strong indicator of acromegaly which is difficult to detect as an image feature using traditional feature detectors operating on the image. Although it is significantly more subtle, the 3D model also seems to have captured the coarseness of the man's lips, another indicator of possible acromegaly. Our classifier correctly classified this man as an acromegalic.*

Figure 3: *A rendering from the 3D model fit to the man on the right of Figure 1. Using our classifier on the parameters of the 3D model, we classified this man as a person without acromegaly.*

cense, one could voluntarily choose to be prescreened automatically for various conditions such as acromegaly. If the posterior probability of some disease exceeded a particular threshold, the system would recommend that the driver see a physician for further analysis. We believe that eventually, such systems could make a signficant contribution to detecting this and other conditions early, which would have a signficant impact on treating the disease. The effectiveness of treating acromegaly patients, like many other diseases, is heavily dependent upon how early it is detected [4].

In this work, we consider a semi-automated approach to detecting the presence of acromegaly in people, using a supervised learning paradigm. We built a training set of images by taking 24 frontal photographs of acromegalic patients and 25 photographs of disease-free subjects. Because the recognition of acromegaly is dependent upon subtle features which are difficult to detect locally in the image, we decided upon a more global method of modelling that uses information across the entire photograph. In particular, we modelled each subject's face in an analysis-by-synthesis loop using the morphable models of Blanz and Vetter [2]. The parameters of the morphable model capture many features of the three-dimensional shape of a subject's head from just a single photograph and are excellent features with which to classify subjects as either acromegalic or not. We report encouraging results of a classifier built from the training set of real human subjects.

2 The Training Data

Because acromegaly is uncommon, photographs of acromegalics are not readily available. Thus, a major portion of the effort in this work was in acquiring a database of images of acromegaly patients, and a matching set of images from healthy subjects. Three of the acromegaly patients from our database are shown in Figure 4.

At the top is a patient with clear signs of the disease. Features such as a large jaw, protruding brow, frontal bossing (a distinctive protrusion of the forehead), swollen nose, prominent cheekbones, enlarged lips, and prominent naso-labial folds (creases in the skin of the cheek, visible here mainly in the top left patient) all make these patients strong candidates for further evaluation by an endocrinologist. While no one of these symptoms would necessarily indicate acromegaly (with the possible exception of the frontal bossing), taken together they are a strong indication of the disease.

The patient in the middle still has clear, although more subtle, signs of acromegaly. He is a good example of the type of patient that would almost certainly not be detected by a lay person, and might often be missed by a physician who is not a specialist in acromegaly. The patient has an enlarged jaw, some swelling of the nose, and probably of the lips, some slight enlargement of the cheekbones, in addition to minor naso-labial folds. Finally, at the bottom of the figure is a patient with very minor, if any, signs of the disease visible even to the expert.

In view of the wide range of symptoms and their strength, it seems unlikely that any test based purely on visual information will be able to diagnose this condition with perfect accuracy. Nevertheless, as a prescreening tool, we believe classifiers such as the one described below can be of great value.

One of the dangers in building a binary classifier is that details of the image acquisition environment will be leveraged by the classifier to "recognize" a condition such as acromegaly. That is, if some simple feature of the acquisition environment differs between one class and another, this may be used improperly to aid the classification. This was an especially important concern for us since we acquired images of acromegalics in one location and images of normals in another location. To minimize the probability that some exogenous factor affected the classification performance, we developed a protocol for taking photos. Our protocol specified

- the camera make and model to be used (Nikon Coolpix 5000 digital camera),
- the background for the photograph (a specific piece of colored fabric),
- the expression of the subject (relaxed and neutral expression, including a closed mouth and open eyes),
- the orientation of the patient (front-facing),
- and the general lighting conditions.

Ten pictures of each patient were taken so that images with closed eyes, accidental non-neutral facial expressions, blur due to movement, and other anomalies could be removed. Typically, at least eight of the pictures of each subject were of good enough quality to use. One of the defect-free photographs of each subject was chosen manually for inclusion in the final database.

Finally, as described below, the color and texture of a patient's face was *not* used directly in the final classification.

Figure 4: *Examples of acromegalics from our database. The symptoms decrease in severity from top to bottom. (See text.)*

This minimizes two important effects which could cause biases in the classification, that of lighting which is difficult to control carefully, and that of skin tone, which happens to be biased slightly toward darker tones in our set of acromegalics relative to our set of normals.

There are three properties of our database that are not ideal, and which we plan to address in the near future. As noted above, all of the acromegalic photos were taken in one locale and all of the normals were taken in another locale. Despite our precautions, this could possibly lead to hidden biases in the classifier. Another issue is that the photos of acromegalics were taken somewhat closer or with a slightly different zoom, on average, than the photos of normals. While the morphable model software explicitly incorporates estimates of distance and zoom into its estimate of 3D shape (minimizing the effect of differences in perspective or fish-eye distortion), these are nevertheless systematic differences in the database which should be eliminated in the future. Finally, our database of normals consists only of white males, while our acromegalics are a signficantly more diverse population. While we do not believe that any of these issues had a significant impact on the classifier, we plan to remedy all of these problems with the image database in future work. Next, we turn to the task of modelling the faces in our database.

3. Modelling Faces

Many of the symptoms of acromegaly are difficult to capture using traditional local image features such as edges and image derivatives. Symptoms such as swelling of the nose and lips, protrusion of the brow and cheekbones, and growth of the jaw are very difficult to detect locally. Our initial work on this project focussed on measuring distances between various facial landmarks and computing the relative size of landmarks not correlated with disease, such as iris diameter, and landmarks like jaw width, that would be expected to be larger, on average, given that the patient had the disease.

One problem with this approach is that it is often difficult to choose landmarks that lead to consistent measurements of a patient's face. It is very difficult to define in a consistent and repeatable fashion measurements such as the "width of the jaw". Developing software to do this automatically is even more difficult.

Another problem with the landmark method is that it does not use all of the information available in the photograph. A patient's nose may be a normal width, according to a landmark based specification, and yet it may be clear to any observer that the patient's nose is actually swollen. In other words, landmark-based methods may not capture a significant symptom which is obvious to an observer.

It became clear that a system which could model the true three-dimensional shape of each subject's head could alleviate many of these problems. The 3D morphable models of Blanz and Vetter [2] seemed to be an ideal tool to tackle such a problem.

3.1 Morphable Models

In previous work, Blanz and Vetter developed a linear statistical model of the 3D geometry and texture (or surface color) of human heads from a set of 3D Cyberware (TM) laser scans of 200 individuals (100 male and 100 female).

In the morphable face model, facial surface data that were recorded with the laser scanner as a triangular mesh are represented in shape vectors that combine x, y, and z coordinates of all vertices:

$$\mathbf{v} = (x_1, y_1, z_1, ..., x_n, y_n, z_n)^T \in \mathcal{R}^{3n}.$$

Sampled at a spacing of less than one millimeter, each surface is represented by $n = 75972$ vertices. Linear combinations of shape vectors will only produce realistic novel faces if corresponding points, such as the tip of the nose, are represented by the same vector components across all individual shape vectors. This is achieved by establishing dense correspondence between different scans, and forming vectors $\mathbf{v}_i$ in a consistent way. Along with shape, the morphable face model also represents texture, but texture is discarded in our work here.

There are two key features of the morphable model work that made it an ideal tool for our application. First, the 3D head model represents shape variations in terms of the common modes (principal components) of shape deformation of healthy human heads. Each parameter of the geometry-part of the model describes a common mode of variation in (densely) registered human heads. While parameters such as jaw size are not explicitly coded into the model, the statistical model must be able to represent such variability as linear combinations of its parameters in order to achieve good approximations of the subjects in the original Cyberware scan database. That is, as long as there were some significant relative differences in jaw size among the initial subjects, the model would need to encode this variability in some form to make good fits. We hypothesized that the analysis of these sorts of natural parameters of face variation would allow us to sort subjects into groups of acromegalics and healthy patients.

The second appealing feature of the morphable model work is that its developers showed how an analysis-by-synthesis method could produce the approximate 3D shape and texture of a person's head from a single photograph. After a manual initialization process (described below), a fully automatic procedure adapts the parameters of the morphable model until a rendered image of the model matches a given photograph as closely as possible under a soft constraint that makes the parameters of the resulting 3D head as likely as possible under the statistical morphable model. That is, the analysis-by-synthesis loops strives to minimize an error of the form

$$E(\alpha) = \sum_{x,y} ||\mathbf{I}_{input}(x, y) - \mathbf{I}_{model}(x, y; \alpha)||^2 + f(\alpha),$$

where α is a vector of parameters estimated for the particular head under consideration and $f(\alpha)$ is the negative log likelihood of the parameters under the original linear statistical model of heads.

We were given direct access to the software of Blanz and Vetter, and used it to develop 3D models from photographs such as those shown in Figures 3 and 2. To intialize the model fitting process, the user clicks on a number (seven to 12) of feature points in the image and the corresponding features on the 3D model using an interactive tool. These point correspondences are enforced in the first iterations of the algorithm. Their weight is gradually reduced to zero during the optimization. The pose, illumination, shape and texture coefficients are all set to the default values at the beginning of the fitting process. Additional details of the initialization and fitting process are described elsewhere [3].

In addition to global estimates of head shape, a secondary procedure was used in which smaller parts of the face were extracted and estimated separately. This procedure gives greater detail and accuracy for the followings facial regions, or groups of regions:

- the nose;
- the eyes, eyebrows, and brow;
- the mouth; and
- all other features, such as the cheeks, chin, forehead, ears, and neck.

Thus, in addition to a single global geometric shape estimate for each head, there were four additional "parts" estimates, yielding a total of five individual geometric 3D shape estimates per face. In our experiments, these additional detailed shape analyses improve the performance of the classifier significantly, as discussed below.

4. Experiments

After fitting the morphable model to each face in our database, and to each of the subregions described above, we retrieved 199 "geometry" parameters for each of the five pieces of the head estimate, for a total of 995 parameters per photograph. Parameters measuring texture were discarded as it was decided in advance that the small benefit they might add[1] would be outweighed by the general increase in variance of the results.

We used, somewhat arbitrarily, the first 99 parameters from each part of the geometric head shape estimate, for a total of 495 parameters per head in our final experiments. With only 24 examples from the acromegalic class and 25 examples from the normal class, we needed to mitigate overfitting as much as possible. As a result we decided on a leave-one-out classification paradigm, and decided to use linear or second-order polynomial kernel support vector machines for classification.

We wish to make an additional, often overlooked point, with regard to overfitting. When working with such a small data set, it is important not only to control the capacity of the classifier being used, but also to limit the number of classifiers tried on the test data. Since we could not afford (given the limited training set size) to set aside a validation set with which to experiment with SVM kernels, slack variable parameters, or other classifier types, we chose almost all of the classifier parameters up front. In particular, we decided in advance to use 99 parameters from each part of the geometric shape estimate, and to use linear or quadratic SVMs. While we show results (in Figure 6) for using varying numbers of components from the morphable model to suggest a general trend, we chose in advance (before seeing the data) to select the one with 99 parameters as our final accuracy estimate. We did allow ourselves the luxury of experimenting with quadratic and linear kernels, and found that linear kernels performed better.

[1]In general the skin tone is at best a very weak indicator of acromegaly.

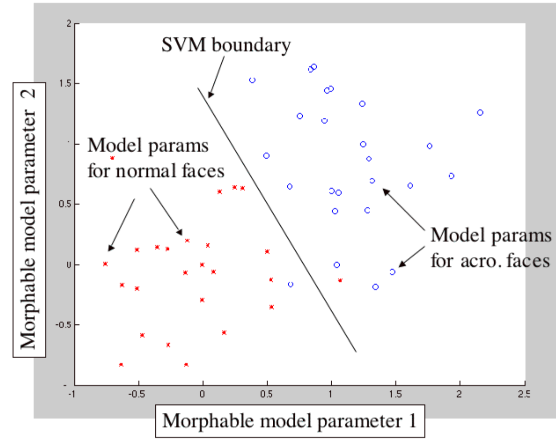

Figure 5: *This figure shows schematically the data used in our classification experiments. Each red or blue point represents the parameters (or coefficients) of the morphable model fit to each subject's face. After holding out one point, an SVM was trained using the remaining points, resulting in a classification boundary used to classify the held out point.*

We used the publicly available support vector machine (SVM) package called SVM Light, which is described elsewhere [6]. Using a leave-one-out paradigm, we trained an SVM using all but one of the training samples, and tested on the remaining sample. Figure 5 shows schematically the data used to train each SVM as the coefficients of the 3D morphable model for each class. Note that the SVM classification boundary will be different, in general, for different test data points, since the training set changes each time. Of course, here we show only two dimensions, rather than the full set of 495 dimensions used in these experiments.

The accuracy using SVM Light with all 495 parameters from the five geometric parts estimates was 85.7% in this leave-one-out scheme. Interestingly, none of the normals were classified as acromegalics, while seven of the acromegalics were classified as normals. The original photographs for all of the misclassified examples are shown in Figure 7. These are all acromegalic patients who were classified as normals. It is encouraging to note that five of the seven patients have very subtle signs of the disease that might easily be missed by an expert. Two of the misclassified examples, however, do show obvious signs of the disease. It is

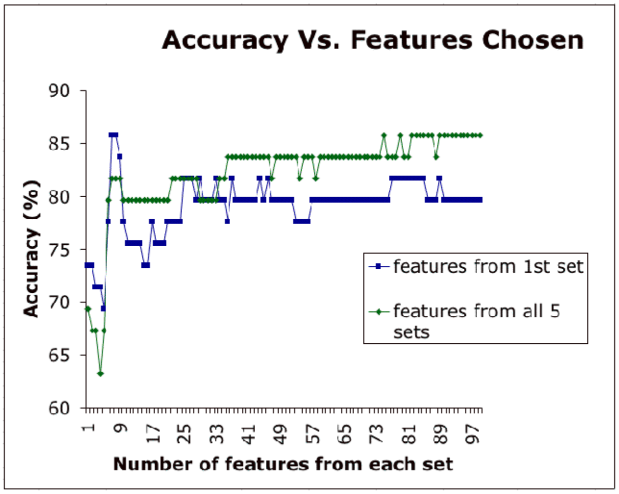

Figure 6: *This figure shows the leave-one-out classification accuracy (percentage of correct classifications) of a linear kernel SVM versus the number of features (parameters) of the morphable model that were used. The blue curve shows performance using only the global estimate of the head, while the green curve shows the performance using components from each of the five geometric shape estimates (one global and four parts based estimates).*

informative to examine the estimated 3D models for these patients, which are shown in Figure 8. In particular, for the patient on the upper right of the figure, who has severe acromegalic symptoms, the match by the 3D model is not very good. This may be in part because the face differs so much from a statistically common face that it is difficult for the principal component model to represent it well. For the patient on the bottom of the left column, another with clear signs of the disease, the 3D model looks fairly accurate, so it is not clear why he was not classified as an acromegalic. One possibility is that relative jaw size, in this database, may have been the best indicator of acromegaly, and this individual's symptoms are better described as frontal bossing, increased distance between the eyes, and large cheekbones. It is possible that increasing the size of our database would allow us to classify such examples correctly.

5. Related Work

There are a number of efforts in the medical community to do automatic statistical analysis of face shape. See for example [5, 1]. These methods, and others of which we are aware, require a specialized apparatus to acquire three-dimensional information about the subject. Stereo methods [1] and full 3D scans [5] are common methods for the acquisition of 3D information.

A key feature of this work is that while the original statistical morphable model required the acquisition of 3D laser scans, it can be applied to other database, like ours, consisting only of regular two-dimensional photographs. Such systems could in theory be widely deployed, analysing standard photographs to screen for a variety of conditions. It is relatively easy, in such a set-up, to implement an "opt-in" or "opt-out" system so that the analysis is completely voluntary.

While there is an enormous literature on face recognition, it is less common to use face databases to categorize images into categories. A good example of such an application is the classification of faces as male or female [7]. We are not aware of any previous work in trying to identify acromegaly. The combination of the rareness of the disease, the visual nature of many of the symptoms, and the benefits from early diagnosis [4] make this disease an ideal candidate for this type of analysis.

6. Discussion

Our ultimate goal is to deploy a system in the real world that can be used to help real sufferers of this disease. To achieve this goal, we must address each of the following issues.

- First, we must improve the photographic database. In addition to gathering more pictures to improve the reliability and robustness of our results, we must address the issues previously discussed regarding diversity and parameter control in the database.

- A real system must either use a full-time operator to perform manual functions or be fully automatic. Since the whole point of our work is to eliminate expensive and tedious work on large numbers of healthy patients, we would greatly prefer a fully automatic process. More research will be required to fully automate the fitting of a morphable model to a face, but this is an active area of research, and we believe that this is an attainable goal in a high percentage of cases.

- It is not clear what accuracy will be required to deploy such a system in the field. If we set a threshold of disease too low, physicians would be overwhelmed with requests for appointments for "potentially positive" subjects. If we set the threshold too high, virtually no patients will be screened as true positives. This problem is exacerbated by the rarity of the disease. It is worth noting, however, that even if a system for prescreening misses a large number of acromegaly patients, it will still provide an important service if it can detect some of them without too many false positives. Thus, a classifier such as the one reported here may be already getting close to practical levels of usefulness by screening many of the acromegalics without setting off any false alarms. Of course, we hope that by increasing the size of our database we will be able to improve the performance of our classifier further.

One final idea we wish to pursue is to perform a regression on the "degree of acromegaly" rather than on a binary random variable representing whether the subject has the disease or not. That is, we really want to teach the algorithm to estimate the degree of symptoms rather than estimating whether a patient has the disease or not. We should be able to make better use of our positive and negative examples by having experts rate the severity of symptoms on a scale from say, 1 to 100, and then trying to place new subjects appropriately on this scale (regression).

References

[1] Nicola D'Appuzo, "Measurement and modeling of human faces from multiimages," *International Archives of Photogrammetry and Remote Sensing*, Volume 34, Number 5, Commission V, WG, V/6. Pages 241–246, 2003.

[2] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3D faces." *SIGGRAPH'99 Conference Proceedings.* 1999.

[3] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3D morphable model." *IEEE Transactions on Pattern Analysis and Machine Intelligence,*" Volume 25, Number 9. pages 1063–1074. 2003.

[4] Pamela U. Freda, "Advances in the diagnosis of acromegaly." *The Endocrinologist,* Volume 10, pages 237-244, 2000.

[5] Peter Hammond, Tim J. Hutton, Michael A. Patton, and Judith E. Allanson. "Delineation and visualization of congenital abnormality using 3D facial images." Presented at the worshop on *Intelligent Data Analysis in Medicin and Pharmacology* at *MEDINFO 2001*, London, Semptember, 2001.

[6] Thorsten Joachims. "Learning to classify text using support vector machines: Methods, Theory, and Algorithms." Dissertation. Kluwer, 2002.

[7] B. Moghaddam and M.-H. Yang. "Learning gender with support faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **24**, 5, pp. 707–711, 2002.

[8] Pieters Nieuwlaat, "A medical mystery – Which twin is the patient?" *New England Journal of Medicine*, Vol. 351, page 68, 2004.

Figure 7: *This figure shows the photographs of six out of seven subjects misclassified by our linear SVM classifier. It turns out that all of the misclassified examples were acromegalics. Five of the seven would generally be considered to have mild or difficult-to-see acromegalic symptoms (all but the upper right and lower left patients).*

Figure 8: *This figure shows the frontal view of the 3D reconstruction estimated for each of the misclassified subjects listed in Figure 7. It may be that the upper right patient was misclassified as a normal (rather than an acromegalic) because the morphable model did not successfully match his facial features.*