Automated Classification System for Bone Age X-ray Images

Jinwoo Seok∗, Baro Hyun∗, Josephine Kasa-Vubu†, and Anouck Girard∗
∗Department of Aerospace Engineering
University of Michigan, Ann Arbor, Michigan 48109-2140
Email:{sjinu,bhyun,anouck}@umich.edu
†Department of Pediatrics-Division of Pediatric Endocrinology
University of Michigan Health System, Ann Arbor, Michigan 48109
Email: jzkv@med.umich.edu

Abstract—Bone Age (BA) determination using radiological images of left hands and wrists is important in pediatric endocrinology to correctly assess growth and pubertal maturation. In this paper, we propose a fully automated Greulich and Pyle Atlas (GP) bone age determination system using feature extraction and machine learning classifiers. The original contributions of this paper are as follows: (i) We use commercially available morphing tools to create a modified GP atlas that has images regularly spaced at three month intervals, (ii) We propose a novel Singular Value Decomposition (SVD) based feature extractor to create a feature vector. We use the Scale Invariant Feature Transform (SIFT) to extract features from the images then apply SVD to compose the feature vectors. Then, we train a Neural Network classifier using the generated feature vectors. Our preliminary results show that, even with a small number of training data sets, we obtain promising results. Future direction is discussed.

Index Terms—Medical and Health Systems, Image Classification, Scale Invariant Feature Transform (SIFT).

I. INTRODUCTION

A. Motivation

The assessment of growth and pubertal maturation is central to the practice of pediatric endocrinology. Estimation of the bone age with a radiological image of the left hand and wrist describes the degree of maturation of a child’s skeleton. For the past 50 years the assessment of a child’s bone age has been the result of a visual match with one of the standards compiled in the Greulich and Pyle Atlas (1959) [1], which provides a set of radiographic data sampled from children at different ages. While the Greulich and Pyle Atlas is widely used and accepted by most pediatric endocrinologists as a “gold standard”, this clinical tool has several limitations: (a) The Greulich and Pyle Atlas is derived from a limited number of children and provides an insufficient number of data spanning the entire growth spectrum, from early childhood to late adolescence; (b) the data was sampled randomly such that the temporal gap between two consecutive samples can be quite large (e.g., more than a year of spacing in clinical ranges of interest); (c) the data itself is extremely outdated and biased since the sample population is from the 1950s and is not representative of children in the U.S. today. In the last five decades, the U.S. population has been reshaped by a larger number of children of international ancestry and the nutritional environment has fostered an obesity epidemic with a profound impact on the growth and rate of physical maturation of children. Such limitations affect the accuracy of the bone age assessment, which is further hampered in instances when the unknown bone age relates to an intermediate value between two standards. There is some evidence that bone age (BA) assignments by different physicians can differ significantly; for instance, in a blind bone age rating study with four raters [2], among the 27 outliers, 6 original manual ratings were found to have a deviation larger than 1.9 years from the reference. Additional data can be found in [3, 4]. Such a reliance on subjective evaluation, large standard deviations for each standard, and the significant variability between raters makes the estimation of bone age using the Greulich and Pyle Atlas susceptible to poor prediction performance.

To make a bone age determination, a radiologist compares the patient’s x-ray to those contained in the reference atlas and determines which image in the atlas the patient’s x-ray is closest to. The patient’s x-ray is then assigned the bone age corresponding to that atlas image. From an engineering perspective, the method is human-centric and prone to bias in interpretation, that is, different clinicians may use different criteria when matching the patient’s x-ray to the atlas. In addition, the large gaps between atlas images in clinically important ranges are frustrating. As such, our research will attempt to modernize the methodology by:

1) Generating intermediate images for Greulich and Pyle, using nothing but their original data. We use commercially available morphing tools which allow us to create a modified atlas that will have images regularly spaced at three month intervals in the clinically significant ranges, allowing a more precise assessment of BA.

2) Using engineering techniques to calculate automatically how closely the patient’s x-ray correlates with the new (or existing) atlas images. This process is not simple, as many different factors come into play in BA determination. It requires relevant feature extraction and intelligence in making the determination.

The first bullet point is a new application of existing engineering techniques. The reason we can generate intermediate
images using commercially available morphing tools is as follows. Even though the growth rate is in general nonlinear over one’s childhood, we assume that it is linear over short time periods. The Greulich and Pyle images provide us with endpoint data, and we generate intermediate images by linear interpolation between the existing standards.

B. Literature Review

Bone age (BA) is a measure of the maturation of a child’s skeleton, and, as such, a key clinical indicator of growth, used by pediatricians and pediatric endocrinologists. As a person’s body ages, from birth through childhood, puberty and adulthood, the size and shape of the bones of the skeleton changes. The bone age is meant to be the “average” age at which the skeleton reaches a certain degree of maturation. In combination with other measures, it can be used to predict future adult height. BA is determined by taking an x-ray of a specific portion of a child’s skeleton (usually the left hand), and comparing it with images contained in a reference atlas. For example, the “Radiographic Atlas of Skeletal Development of the Hand and Wrist” by Greulich and Pyle (GP) [1] contains 31 x-ray images for boys and 29 for girls, aged from newborns to 19 and 50 year-olds, respectively. The children accepted for study were selected on the basis of their freedom from gross physical or mental defects and on the promise of their parents to permit their continued participation until the completion of the project [1].

There have been attempts at automated BA detection such as CASAS [5], Peitka [6] and [7, 8]. Unfortunately, those systems were not deemed accurate enough to be used for clinical purposes. A next generation automatic BA determination system, called BoneXpert [9, 10], has been developed recently. It uses the Active Appearance Model (AAM) [11], which is a new method of matching statistical models of appearance to artificial images that AAM reconstructs from training sets using shape and appearance parameters. BoneXpert is formed of three layers [9]: Layer A reconstructs the outlines of 15 artificial but realistic bones (radius, ulna and 11 short bones) by using a manually annotated training set; Layer B predicts intrinsic BA based on chronological age (CA) and the x-ray data, and then layer C transforms the intrinsic BA into either GP BA or Tanner Whitehouse (TW) BA. To determine the intrinsic BA, the system uses three features, namely shape, intensities and texture. To validate the system, the authors used 1097 images and reported that the system accepted 99% of the validation images (14 images are rejected) and that the root mean square deviation between a manually obtained rating and the system’s was 0.72 years (95% CI 0.69-0.75) [2].

However, there are some possible points of contention: to validate their system, the authors used reference data that was re-rated by four human raters. But only 27 images among 1,097 images, which had a big gap between the original manual BA and the BoneXpert BA, were re-rated. This ignores the possibility that there may be errors in the validating data (that is, in the other 1,070 images). In addition, while the principal assumption of this system is that there is a relationship between CA and BA (CA is an input to the system); BoneXpert does not determine BA directly, and the CA - BA relationship is unclear from the publications.

In this study, we propose an automated bone age determination system based on GP’s bone age data. As such, we determine bone age directly from the x-ray images without any information of CA. We use a feature extraction method called Scale Invariant Feature Transform (SIFT), proposed by David G. Lowe in 1999 [12], which is a robust feature extraction technique that gives scaling and rotation invariant features. In addition, we use neural networks [13] to classify the bone age data appropriately. Details are given in the technical approach section III and IV. Fig. 1 shows the schematic overview of the automated classification system.

C. Original Contributions

The original contributions of this work are as follows: (i) we use commercially available morphing tools to create a modified atlas that has images regularly spaced at three month intervals in the clinically significant ranges, allowing a more precise assessment of BA, (ii) we propose a novel Singular Value Decomposition (SVD)-based feature extractor to create a feature vector out of the descriptors obtained from SIFT. Our preliminary results show that, in some cases, using the SVD-based feature extractor gives a better classification performance than using commonly available techniques, such as Bag Of visual Words (BoW) [14].

D. Organization

The organization of this paper is as follows. In Sec. II, we briefly discuss the image morphing technique that we used and in Sec. III, we present the novel SVD-based feature extractor along with other existing methodologies that we used in this study. In Sec. IV, we describe the classifiers. In Sec. V, we present our preliminary results and finally, in Sec. VI, we present the conclusions and future work.

II. IMAGE MORPHING

Morphing is an image processing methodology that is primarily used for the metamorphosis of an image into another.
The key idea is to obtain a sequence of intermediate images that, when put together with the original images, represents the change from one image to the other. The more intermediate images, the smoother the transition between the original images.

Among commercially-available image morphing software, we chose Abrosoft® FantaMorph 5 for its user-friendly interface. We have selected seven radiograph samples each from the female standards that are within the clinically significant range: female standards 14-20 (5 years 9 months - 12 years). Each sample was scanned with a resolution of 300 dpi and converted to jpeg format.

We manually specified a number of morphing points that were used as the reference points when morphing. These points can be chosen judiciously such that clinically interesting areas, such as the epiphysis and the capitate, have more points for better morphing resolution. As a result, we have images areas, such as the epiphysis and the capitate, have more points can be chosen judiciously such that clinically interesting points were used as the reference points when morphing. These were used as the reference points when morphing. These converted to jpeg format.

Experiments [15] suggest that equal to 10 is a good threshold.

III. IMAGE FEATURE EXTRACTION

Extracting image features is a key step for image classification. The performance of the classifier is dependent on the image features that were chosen. Thus, how to make a distinctive feature vector is very important for the purpose of improving classification performance. There are many features that can be extracted from an image, such as edges, textures and colors, but using one kind of feature alone may not yield good performance for the classifier. To improve the performance, we need to customize each classifier with respect to its purpose and possibly use a combination of existing methods.

A. Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) is one of the feature extraction methods that was introduced by David G. Lowe in 1999 [12], and was further improved in 2004 [15]. SIFT is a local-based feature extraction method whose feature descriptor is invariant to scaling and rotation, and partially invariant to viewpoint and illumination changes. In other words, SIFT is a robust method to identify objects in images. Key steps in the method are described below. Interested readers may consult [12] and [15] for more details.

1) Detection of scale-space extrema: SIFT detects key points using a cascade filtering approach, which is an efficient algorithm to identify candidate locations. Using the Gaussian function, \( G(x, y, \sigma) \), the scale space function of an image, \( L(x, y, \sigma) \), is defined as:

\[
L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),
\]

where \( I(x, y) \) is an input image, \( \sigma \) is a constant, and ‘∗’ is the convolution operation in \( x \) and \( y \) which are pixels in the imported images. To detect key points efficiently, we use scale-space extrema in the difference-of-Gaussian function (DoG) convolved with the image \( D(x, y, \sigma) \), which is the difference between two nearby spaces of an image scaled by the constant \( k \), i.e.,

\[
D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma).
\]

The difference-of-Gaussian function can be expressed as a close approximation of the scale-normalized Laplacian of Gaussian, \( \sigma^2 \nabla^2 G \), such that

\[
G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G.
\]

Comparing each sample point’s eight neighbors in the current image and nine neighbors in the scaled above and below image, we can detect the local maxima and minima of \( D(x, y, \sigma) \), which are the candidate key points.

2) Accurate keypoint localization: After obtaining candidate key points, the next step is to eliminate the unstable key points, which are those that have low contrasts or are poorly localized along the edge. To eliminate low contrast candidate key points, the Taylor expansion of the scale-space function, \( D(x, y, \sigma) \), shifted such that the origin is at the sample point, is used, i.e.,

\[
D(\hat{x}) = D + \frac{1}{2} \frac{\partial^2 D}{\partial x^2} \hat{x},
\]

where \( D(\hat{x}) \) is the function value at the extremum and \( \hat{x} = -\frac{\partial^2 D}{\partial x^2}^{-1} \frac{\partial D}{\partial x} \). Using a threshold determined experimentally [15], values of \( |D(\hat{x})| \) less than 0.03 are removed (assuming the image pixel values are in the range of \([0, 1])\).

Strong responses are detected by the difference-of-Gaussian function along the edge, although the edge may be poorly localized. A poorly-localized edge in the difference-of-Gaussian has large principal curvatures and can be calculated from a \( 2 \times 2 Hessian \) matrix. Note that the eigenvalues of \( H \) are proportional to the principal curvatures of \( D \). Let \( r \) be the ratio of the largest and smallest eigenvalue of \( H \). Then, according to the following equation,

\[
\frac{Tr(H)^2}{Det(H)} < \frac{(r + 1)^2}{r},
\]

the unstable key points are eliminated by a predefined \( r \). Experiments [15] suggest that \( r \) equal to 10 is a good threshold.
3) Orientation assignment: Assigning consistent orientation to the key point, which is based on local properties, makes the key point descriptor be represented relative to the orientation. Therefore, the key point descriptor is invariant to image rotation. From [15], we can obtain the gradient magnitude, \(m(x, y)\), and orientation, \(\theta(x, y)\). The orientation histogram is established from the gradient orientation around the key point. The highest peak in the histogram is chosen to be the orientation of the key point that is the dominant local gradient. In addition, any peak higher than 80% of the highest peak is also chosen to be an orientation of the key point because it benefits matching stability; that is, there can be multiple orientations with respect to the same location and scale.

4) The local image descriptor: Using Gaussian windows, image descriptors relative to each key point are computed; they are distinctive, yet invariant to the changes of images. In order to compute the key point descriptors, image gradient magnitudes and orientations are sampled around key points. To make the descriptor be invariant to rotations, image gradient magnitudes and orientations have to be rotated relative to the key point orientation. Then the Gaussian window is used to weigh the magnitude of each sample point. In order to prevent rapid changes in key point descriptors and give less weight to gradients that are apart, a Gaussian window is used. A \(4 \times 4\) sub-sampled region forms 8 orientation histograms in which each orientation has a magnitude. By choosing \(16 \times 16\) sampled gradients around the key point, \(4 \times 4\) orientation histograms are generated, which means a total of \(4 \times 4 \times 8 = 128\) features are generated for each key point.

We apply the SIFT method using VL-SIFT which is available as an open source code by the group VLFeat [16]. An example of SIFT descriptors for a BA determination x-ray is shown in Fig. 3.

B. Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is one of the matrix factorization methods that allows reducing the size while keeping the characteristics of a matrix. Note that it is important to reduce the size of the feature vector to reduce the computational load for training the classifier. Given an \(m \times m\) matrix \(A\), the expression of its SVD is

\[
A = U\Sigma V^T, \tag{6}
\]

where \(U\) is an \(m \times m\) matrix, \(V\) is an \(n \times n\) matrix and \(\Sigma\) is the singular values of matrix \(A\) which is an \(m \times n\) non-negative real diagonal matrix. We use SVD on the feature matrix, which is generated from SIFT, to make a shorter feature vector that still holds the features’ characteristics. The reason for using SVD to make feature vectors is that it is relatively straight-forward to apply and gives the singular values, which contain information of the features generated from SIFT, in a descending order. Ordering the elements of the feature vector in descending order gives an advantage when using a training data set for a neural network.

![Figure 3. Feature descriptors using VL-SIFT. We can see key points, orientations and descriptors. The yellow circles indicate the key points, and the lines in the circles indicate orientation. The green boxes and arrows indicate the descriptors.](image339x508 to 560x730)

C. Bag of Visual Words

The Bag of Visual Words [14] is one of the methods that can generate feature vectors. It is inspired by "Bag of Words" which is commonly used in documentation classification. Using K-means clustering, it groups the local features into a desired number of groups and makes the feature vector by counting the number of each group (the dimension of the feature vector). We apply K-means clustering using VL-SIFT which is available as an open source code by the group VLFeat [16].

IV. CLASSIFICATION BY NEURAL NETWORKS

We use neural networks for bone age determination, mainly because of their strong performance when the machine has limited information (in our case, training data). Currently, we have access to 24 images from the GP data and, among them, 19 images that are clinically interesting are used for morphing; we only have one image for each class for training. We use neural networks as a proof of concept. At this time we are gathering patient data which will be used to train future versions of our classification system.

Artificial Neural Networks (ANNs) [13] are inspired by natural neural networks. When inputs come in to the ANNs, they are weighted, and outputs are constructed based on the weighted inputs. To make the right decisions, ANNs need to be trained, and using a set of inputs and an initialized set of desired outputs, one can train ANNs. The training process begins with ANNs initializing and assigning random weights to the input. ANNs adjust weight to get the desired output.
by considering the errors between the actual output and the desired output. The back propagation algorithm is a commonly used tool to train ANNs. The back propagation algorithm uses a weighted sum given as

\[ A_j(x, w) = \sum_{i=0}^{n} x_i w_{ij} \]

where \( x_i \) is an input, \( w_{ij} \) is the respective weight, \( i \) is the number of inputs and \( j \) is the number of hidden layers (an indication of the complexity of the ANN/model). From the weighted sum, we can obtain the output

\[ Y_j(x, w) = \frac{1}{1 + e^{A_j(x, w)}}. \]

By minimizing the difference of \( Y_j(x, w) \) and the desired output, the weights \( (w_{ij}) \) can be calculated. We used 24 training data (i.e., feature vectors) to get the desired output which is an \( I_{24 \times 24} \) matrix where each column represents one training data’s desired output. The neural network is trained to match each feature vector to each column in the identity matrix using Eq. (8).

V. SIMULATION RESULTS

A brief explanation of the steps forming a classification decision is as follows. First, we import images to Matlab. Then, we apply the SIFT algorithm using VL-SIFT to get each image’s key points and local image descriptors. Then, we apply SVD to each local image descriptor to get reduced feature vectors which represent each radiographic picture. After we obtain feature vectors, we train the ANN. For testing, we apply the same steps to obtain a feature vector for the test data. Then we input the test data to the trained ANN, and obtain the output performance. We used built-in functions for the Neural Network (Matlab® commands: newff, train) and the Singular Value Decomposition (Matlab® command: svd).

A. Data Set

We use 24 GP female standard images: 1 through 27 excluding 13, 21 and 27 because they include a ring on a finger and 21 because one of the fingers is partially cut off. In addition, 19 morphing images were generated for the range from 6 years to 11 year 9 months (Bone Age), which is the clinically interesting age range for girls.

B. Simulation Results

At first, we use GP’s female standard hand radiographic data only without the morphing data to train the network. For a testing data set, we use 19 morphing images which we generated, to validate the classifier and also the effectiveness of the morphing images. We did several tests using the same classifier and the same test set to confirm the consistency of the classifier’s performance. Two results are shown in Fig. 4. The diagonal line in Fig. 4 indicates the correct answer. The test result 1, marked with circles in the figure, shows that the classifier works well as most the answers are closely aligned to the diagonal line. Only one result shows radical misclassification; in that case the GP number was classified as 22 when the true GP number is 15.25. There are three results showing moderate errors, which assign GP number of 21 to data for which the true GP number are 18.75, 19.25 and 19.5. Note that some circles which deviate from the diagonal line represent round-off errors due to the classification label being a natural number while the input is not (e.g., classifying the GP number as 16 when the true GP number is 16.25). In other cases, (test result 2, marked with crosses in the figure), the classifier performs less well. The large error for this test is mainly due to the fact that there was only one training data per class; this is generally not considered sufficient to train classifiers.

In a second phase, we used both GP’s female standard hand radiographic images (excluding 13, 21 and 27) and 19 morphing images which we generated from 6 years to 11 year 9 months to train the network. Note that a total of 43 images were used to train the network. Each class is represented by its Bone Age. Fig. 5 illustrates the classifier’s performance using the same training data for testing. The diagonal line indicates the correct answer. The results in Fig. 5 show that the morphing images can effectively serve as a class and that all the classes are classified appropriately including the morphing images. Note that while the results show proof-of-concept of our approach, specifically that artificially morphed (generated) images can be represented as classes, more evaluations are required for fair judgement of the performance, e.g., evaluation based on different sets of training and testing data.

VI. CONCLUSIONS AND FUTURE WORK

We propose an automated bone age determination system. To accomplish this purpose, first, we generate intermediate images between GP’s bone age using existing morphing technique that can generate regular three-month intervals in the clinically-interesting range. This allows more precise bone age determi-
try other classifiers, such as Support Vector Machines (SVM).

In addition, we will disregard other features as the information from non-ROI may not be considered. We denote the important parts as ROI.

To determine bone age, we believe that the entire hand should be considered. We denote the important parts as ROI. To determine bone age, we believe that the entire hand should be considered. We denote the important parts as ROI.

As future work, we plan to use more x-ray image data such as the images using the Scale Invariant Feature Transform (SIFT) algorithm, and obtain feature vectors by applying Singular Value Decomposition (SVD). We train Neural Networks using feature vectors, and then we simulate the trained classifier to obtain the results as described in Sec. V.

This paper shows proof-of-concept of our approach, and there is much work to be done for future improvements. In this study, we only have morphing images in the clinically interested range which is 5 years to 12 years for female standards and 6 years to 13 years for male standards. We will generate more morphing images to make every gap between two classes in the GP atlas be three months.

As future work, we plan to use more x-ray image data such that the classifiers can be trained better. We will design feature vectors to combine several methods such that the feature vector can distinctly represent each class. In order to design that kind of feature vector, we will introduce Regions of Interest (ROI). To determine bone age, we believe that the entire hand information is not needed, but rather only some important parts should be considered. We denote the important parts as ROI. We will use our new feature vector to form selected ROI and disregard other features as the information from non-ROI may lower the performance of the classifier. In addition, we will try other classifiers, such as Support Vector Machines (SVM).

REFERENCES


