35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Automated Down Syndrome Detection using Facial Photographs*

Qian Zhao, Kenneth Rosenbaum, Kazunori Okada, Dina J. Zand, Raymond Sze, Marshall Summar, and Marius George Linguraru

Abstract—**Down syndrome, the most common single cause of human birth defects, produces alterations in physical growth and mental retardation; its early detection is crucial. Children with Down syndrome generally have distinctive facial characteristics, which brings an opportunity for the computer-aided diagnosis of Down syndrome using photographs of patients. In this study, we propose a novel strategy based on machine learning techniques to detect Down syndrome automatically. A modified constrained local model is used to locate facial landmarks. Then geometric features and texture features based on local binary patterns are extracted around each landmark. Finally, Down syndrome is detected using a variety of classifiers. The best performance achieved 94.6% accuracy, 93.3% precision and 95.5% recall by using support vector machine with radial basis function kernel. The results indicate that our method could assist in Down syndrome screening effectively in a simple, non-invasive way.**

I. Introduction

Down syndrome is a chromosomal condition caused by the presence of a third copy of chromosome 21. It is the most common chromosomal abnormality and it affects one out of every 300 to 1,000 infants worldwide depending on factors such as prenatal testing and maternal age [1, 2]. Patients with Down syndrome have an increased risk for developmental disabilities, heart defects, respiratory and hearing problems and the early detection of the syndrome is fundamental for managing the disease.

Down syndrome may be diagnosed before or after birth. Biochemical screening and cytogenetic diagnostic tests can be performed prenatally. After birth, Down syndrome is most commonly identified on the basis of the presence of a number of minor physical variations and minor malformations including upslanting palpebral fissures, small ears, protruding tongue, a flat facial profile, epicanthal fold (texture feature) and extremity variations. These differences may be subtle and are influenced by the length of gestation, the effects of labor and delivery and the geographical backgrounds of the family, frequently making a rapid, accurate diagnosis difficult. In many centers, access to pediatric specialists including geneticists can provide additional guidance while waiting for chromosomal confirmation which is still likely to take 48-72 hours. Numerous scoring systems have also been devised to help with the diagnosis [3]. For infants born in non-academic centers, more rural settings and internationally, access to specialists is much more limited or not readily available. Dysmorphologists estimate that the accuracy of a clinical diagnosis of Down syndrome prior to cytogenetic results approximates 50%-60% and is likely to be lower in many instances[4]. The development and implementation, therefore, of automated remote image detection for the diagnosis of Down syndrome and other dysmorphic syndromes has the potential for dramatically improving the diagnostic rate and providing early guidance to families and involved professionals.

Photography and image analysis could serve as a readily available and powerful tool for automated computer-aided diagnosis of Down syndrome. Some efforts have been made to diagnose genetic syndromes using face recognition techniques. In [5], the authors investigated the disease-specific facial patterns for ten syndromes, excluding Down syndrome, using Gabor wavelet features. Their latest study [6] classified 14 syndromes with 21% accuracy, representing a ratio of three over a random choice. However, their method did not distinguish the syndromes from a healthy population. Moreover, the method is not fully automated that requires manually labeled landmarks. The authors in [7] detected Down syndrome based on Gabor wavelet transform with high accuracy. In [8], local binary pattern (LBP) was used to recognize Down syndrome with template matching based method. However, both paper in [7, 8] performed pre-processing manually to standardize the images.

In this study, we propose a novel method to detect Down syndrome using non-standardized frontal facial photographs of patients and machine learning techniques. First, facial landmarks are located automatically using a constrained local model (CLM). Then geometric features are extracted from these anatomical landmarks and texture features based on LBP are extracted around each landmark using size-variant windows. Down syndrome specific features are selected after feature extraction. Finally, we compare four classifiers, support vector machines (SVM) with radial basis function (RBF) kernel, linear SVM, k-nearest neighbor (k-NN), and random forest (RF), in terms of accuracy, precision and recall.

II. Methods

The dataset consists of 100 frontal facial photos with 50 Down syndrome patients and 50 healthy individuals acquired with a variety of cameras and under variable illumination. The age of individuals ranges approximately from 0 to 10. The

*This project was supported by a philanthropic gift from the Government of Abu Dhabi to Children's National Medical Center. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the donor.

Qian Zhao is with Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Medical Center, Washington DC 20010 USA (corresponding author to provide phone: 202-476-1285; fax: 202-476-1270; e-mail: qzhao@ cnmc.org).

Raymond Sze and Marius George Linguraru are with Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Medical Center, Washington DC 20010 USA.

Kenneth Rosenbaum, Dina J. Zand, and Marshall Summar are with Division of Genetics and Metabolism, Children's National Medical Center, Washington DC 20010 USA.

Kazunori Okada is with Computer Science Department, San Francisco University, 1600 Holloway Ave. San Francisco, CA 94132 USA.

dataset includes multiple ethnicities and both genders. Due to the non-standardized photography acquisition and highly variable light illumination effects, each image was normalized to have an average intensity 128 and a standard deviation 20. The method, a supervised learning scheme, can be divided into four parts: landmark detection, feature extraction, classification, and evaluation.

A. Automated Landmark Detection

Dryden and Mardia [9] categorize landmarks into three categories: anatomical landmarks, mathematical landmarks, and pseudo-landmarks. We first define 44 anatomical landmarks. As the error in landmark detection decreases approximately quadratically as the number of landmarks increases [10], we add 37 more pseudo-landmarks by interpolation. It results in the final 81 landmarks as shown in Fig.1 (a).

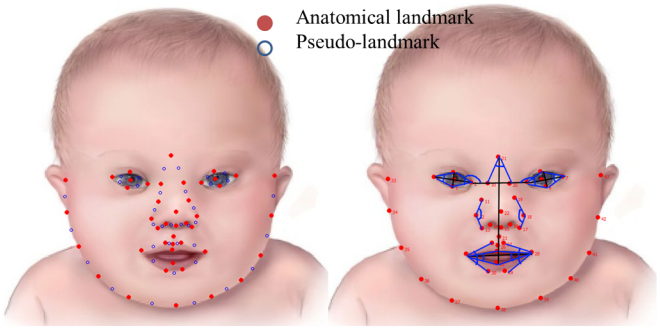


Fig.1.Facial landmarks: (a) Face annotated using 44 anatomical landmarks and 37 pseudo-landmarks; (b) The illustration of geometric landmarks.

A mathematical representation of an n-point shape in k dimensions could be to concatenate each dimension into a kn-vector. The vector representation for planar shapes (i.e. $k=2$) would then be

$$\boldsymbol{x}=\left[x_1,y_1,x_2,y_2\cdots x_n,y_n\right]^T. \tag{1}$$

We use a modified CLM [11, 12] to locate landmarks automatically. Our method optimizes CLM with respect to 2-D coordinates directly instead of optimizing shape parameters. The process can be subdivided into two parts which are model building and searching. For the model building, CLM consists of a shape model and a patch model. The shape model describes how face shape can vary and it is built with principal component analysis (PCA). The patch model describes how the image around each facial feature point should look like. With these two models, both face morphology and textures are described.

To study the shape variation of each landmark throughout the training dataset, all shapes need to be aligned to each other first. The alignment in this study is done by Procrustes analysis. Then, each aligned shape of the training set is represented by the vector $\boldsymbol{x}$, where $\boldsymbol{x}$ now contains the new coordinates resulting from alignment. Then the shape model is built using PCA describing shape variations in the training data. After PCA, we can represent each landmark vector as a linear combination of the principal components.

$$\boldsymbol{x}=\bar{\boldsymbol{x}}+P\boldsymbol{b} \tag{2}$$

where $\bar{\boldsymbol{x}}$ is the mean shape, P is the eigenvector matrix, and $\boldsymbol{b}=\left[b_1\ b_2\ldots b_K\right]^T$ is a shape parameter vector. To obtain acceptable or allowable shapes, the shape parameters vary in the range $-3\sqrt{\lambda_j}<b_j<3\sqrt{\lambda_j}$, $j\in\left[1,K\right]$, where λ_j is the j^{th} eigenvalue Fig.2 shows the first three principal modes of PCA.

The patch model is built using linear SVM due to its powerful discrimination and computational simplicity. For each landmark, we extract m patch samples, some of which are negative and some are positive examples. All m patches have the same patch size, N pixels. We concatenate each patch into a vector. We assign an output value for SVM, $y^{(i)}=\{-1,\ 1\}\ \ i=1,2,\ldots,m.$ For CLM, we can write SVM output as a linear combination of the input vector

$$y^{(i)}=\boldsymbol{w}^T\cdot x^{(i)}+\theta, \tag{3}$$

where $\boldsymbol{w}^T=\left[w_1\ w_2\ldots w_N\right]$ represent the weight for each input pixel, and θ is a bias. The weights matrix could be used as the patch model to capture the intensity information.

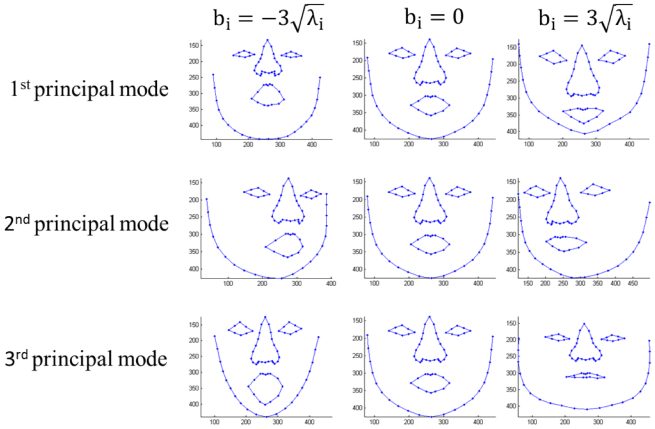


Fig.2. Mean shape deformation using 1st, 2nd and 3rd principal mode, $b_i=-3\sqrt{\lambda_i}, b_i=0, b_i=3\sqrt{\lambda_i}$.

After building the CLM model, we use it to search each landmark around their local region. First, we use Viola-Jones face detector [13] to detect the face, eyes and tip of the nose in the image. The images are cropped according to the rectangle that contains the face, and scaled to 512×512 pixels. Then we make an initial estimation of each landmark location using transformed mean shape. The transform matrix is determined by the detected positions of eyes and tip of nose. Each landmark is searched in the local region of its current position using SVM. We denote the SVM response image with $R(x,y)$ which is fitted with a quadratic function

$$r(x,y)=a(x-x_0)^2+b(y-y_0)+c, \tag{4}$$

which can also be written in matrix format

$$r(x,y)=\boldsymbol{v}H\boldsymbol{v}^T-2F\boldsymbol{v}^T+ax_0^{\ 2}+by_0+c, \tag{5}$$

where $\boldsymbol{v}=\left[x\ y\right]$, $H=\begin{bmatrix}a & 0\\ 0 & b\end{bmatrix}$ and $F=\left[ax_0\ by_0\right]$. The parameters can be solved by minimizing a following objective function

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}=(a,b,c)} \sum_{x,y} \left[R(x,y) - r(x,y) \right]^2 . \quad (6)$$

Finally, the optimal landmark positions are found by optimizing quadratic functions and shape constraints. The joint objective function is

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{H}\boldsymbol{x} - 2\boldsymbol{F}\boldsymbol{x} - \beta\left(\boldsymbol{x} - PP^T\boldsymbol{x}\right)^T \left(\boldsymbol{x} - PP^T\boldsymbol{x}\right), \quad (7)$$

$$\text{subject to } -3 < P_{nr}^T \cdot \boldsymbol{x} < 3 \quad (8)$$

where $\boldsymbol{H} = \begin{bmatrix} H_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_n \end{bmatrix}$, $\boldsymbol{F} = \begin{bmatrix} F_1 & \cdots & F_n \end{bmatrix}$

$P_{nr} = \left[p_1 / \sqrt{\lambda_1} \ldots p_K / \sqrt{\lambda_K} \right]$, and p_i is the ith eigenvector.

To make the search more efficient and robust, we perform a multi-resolution search. Before the search begins, we build an image pyramid and repeat the CLM search at each level, from coarse to fine resolution. The start shape for the first search is the shape generated from the Viola-Jones face detector. The start shape at each subsequent level is the best face found by the search at the level below.

B. Feature Extraction

Before the feature extraction, all images are aligned to the first image using Procrustes analysis, but keeping their own scales. Thus, the effect of translation and in-plane rotation is removed, but the image resolution is reserved.

The geometric features consist of three types: horizontal distances, vertical distances and corner angles. We define four horizontal distances, seven vertical distances and 13 angles, as shown in Fig.1 (b). The horizontal and vertical distances are both normalized by their baselines, respectively. The horizontal baseline is the distance between left corner of left eye and right corner of right eye, and the vertical baseline is the distance between the forehead point and lower lip point. The normalized geometric features are invariant to scale, translation and rotation.

The local texture features are extracted based on LBP histogram [14]. First, a LBP histogram is extracted from the region of interest (ROI) around each of the 33 inner facial landmarks, covering important facial features. Then six first-order statistical measurements are computed from the histogram, which are the mean, variance, skewness, kurtosis, energy and entropy, to capture the LBP image property. Finally, the feature vectors in all ROIs are concatenated to form the local texture features for the image. Thus, the local texture features with spatial information characterize both the local and global facial textures.

The uniform LBP, originated from LBP is a more general approach, in which the number of neighboring sample points is not limited:

$$LBP_{P,R}^{riu2}(x,y) = \begin{cases} \sum_{p=0}^{P-1} s\left(g_p - g_c\right) & ,if \quad U\left(LBP_{P,R}\right) \le 2 \\ P+1 & ,otherwise \end{cases}, \quad (9)$$

$$U\left(LBP_{P,R}\right) = \left| s\left(g_{P-1} - g_c\right) - s\left(g_0 - g_c\right) \right| + \sum_{p=1}^{P} \left| s\left(g_p - g_c\right) - s\left(g_{p-1} - g_c\right) \right| \quad (10)$$

where $s(\cdot)$ is a sign function, g_p corresponds to the grey values of P equally spaced pixels on a circle of radius R, and g_c is the grey value of the central pixel. In this study, we set $P = 8$ and $R = 1$.

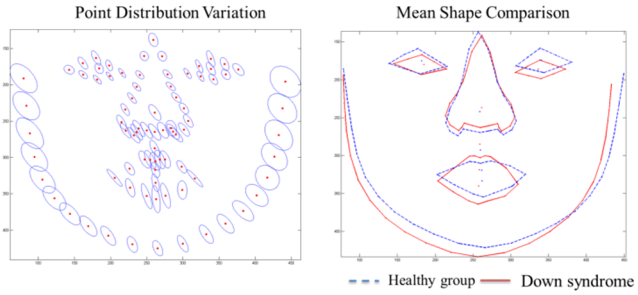
We combine the geometric and local texture features by concatenation. The dimensions for the original geometric and local texture features were 24 and 198, respectively, adding to 222 combined features. The features are then ranked based on the area under the receiver operating characteristic (ROC) curve and the random classifier slope. The optimal dimension for each type of features is found based on ROC by empirical exhaustive search.

The SVM [15] with RBF kernel, linear SVM, k-NN [16], and RF classifiers [17] are compared in this study. In this study, we set $k = 3$ for k-NN classifier and trained the RF with 150 trees.

III. Experiments

A. Automated landmark detection

Although we built the shape model to include both the normal and abnormal cases, we also compared the mean shape of the Down syndrome and healthy groups. Fig. 3 (a) shows the statistical point distribution of the training data indicating large variation of our dataset and Fig. 3 (b) shows the mean shapes of the two groups. They agree with the clinical findings about Down syndrome including small nose, wide-open mouth, etc.



(a) (b)

Fig.3. Shape model: (a) point distribution of the training data; (b) mean shape of Down syndrome and healthy groups.

The performance of landmark detection was evaluated by a normalized error

$$\varepsilon = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N_{\mathrm{I}}} \sum_{x \in \mathrm{I}} d\left(\tilde{x}, x\right) \Big/ d_{pupil} \right), \quad (11)$$

where I is the set of inner face is points, $d\left(\tilde{x}, x\right)$ is the Euclidean distance between inner face points located by the automatic search and the corresponding manually landmarks, d_{pupil} is the distance between two pupils, and N is the number of images. The error is 5.37%+/-1.15%.

B. Down Syndrome Detection

Leave-one-out validation is performed throughout the dataset. The performance is evaluated using accuracy, precision, and recall.

The optimal dimensions of the features were selected with respect to area under curve of the receiver operating characteristic curve. These dimensions are 14, 9 and 9 for geometric, local texture and combined features, respectively.

The experimental results are shown in Table I. We noted that local texture features and combined features both achieved the best performance using SVM classifier with RBF with 94.6% accuracy, and high precision and recall. The high recall rate of 95.5% is preferable in a clinical setting, as a screening tool should miss as few syndromes as possible.

The local texture features outperformed the geometric features by 6.6%, probably due to small errors in landmark detection. For the geometric features, the highest accuracy and precision were obtained by linear SVM, while the highest recall was achieved by kNN.

We performed statistical significance tests using Fisher's exact test for different features and classifiers. The difference between geometric and texture features using SVM with RBF was significant (p=0.03). The average time for analyzing one new case was 23.31s using MATLAB on a Windows 8 core workstation with 12GB RAM.

TABLE I. PERFORMANCE COMPARISON OF FEATURES AND CLASSIFIERS

	Accuracy				**Precision**				**Recall**			
	SVM-RBF	SVM-linear	k-NN	RF	SVM-RBF	SVM-linear	k-NN	RF	SVM-RBF	SVM-linear	k-NN	RF
Geometric	0.826	**0.880**	0.837	0.870	0.818	**0.867**	0.784	0.864	0.818	0.886	**0.909**	0.864
Texture	**0.946**	0.870	0.902	0.891	**0.933**	0.833	0.889	0.870	**0.955**	0.909	0.909	0.909
Geometric +Texture	**0.946**	0.870	0.902	0.891	**0.933**	0.833	0.889	0.886	**0.955**	0.909	0.909	0.886

The performances of geometric, texture and combined features to detect Down syndrome are compared in terms of accuracy, precision and recall. SVM-RBF stands for support vector machine with radial basis function kernel classifier. SVM-linear stands for linear SVM classifier. And RF stands for random forest. The bold numbers indicate the highest performance for each type of feature set.

IV. CONCLUSION

We introduced a novel method for the automated detection of Down syndrome using non-standardized facial photos of patients. Facial landmarks were located using a constrained local model. Then geometric features, based on anatomical landmarks, and local texture features, based on local binary patterns, were extracted. After feature selection, several classifiers were employed to discriminate between the Down syndrome and normal cases. The results showed a substantial improvement of the performance of the method to 94.6% accuracy when using local texture features over geometric features. The SVM with RBF kernel was used as classifier. The results demonstrate the robustness of the technique to analyze highly variable photographic data, and the potential for computer-aided diagnosis for Down syndrome from home photography. Data collection is on-going for more comprehensive validation of our method. In future work we will include features from side-view images and investigate more effective methods for feature fusion.

REFERENCES

[1] G. de Graaf, *et al.*, "Changes in yearly birth prevalence rates of children with Down syndrome in the period 1986–2007 in the Netherlands," *Journal of Intellectual Disability Research,* vol. 55, pp. 462-473, 2011.

[2] F. K. Wiseman, *et al.*, "Down syndrome—recent progress and future prospects," *Human Molecular Genetics,* vol. 18, pp. R75-R83, 2009.

[3] K. Fried, "A score based on eight signs in the diagnosis of Down syndrome in the newborn," *J Ment Defic Res,* vol. 24, pp. 181-5, 1980.

[4] S. Sivakumar and S. Larkins, "Accuracy of clinical diagnosis in Down's syndrome," *Archives of Disease in Childhood,* vol. 89, pp. 691-691, 2004.

[5] H. S. Loos, *et al.*, "Computer-based recognition of dysmorphic faces," *European Journal of Human Genetics,* vol. 11, pp. 555-560, 2003.

[6] S. Boehringer, *et al.*, "Automated syndrome detection in a set of clinical facial photographs," *American Journal of Medical Genetics Part A,* vol. 155, pp. 2161-2169, 2011.

[7] Ş. Saraydemir, *et al.*, "Down Syndrome Diagnosis Based on Gabor Wavelet Transform," *Journal of Medical Systems,* vol. 36, pp. 3205-3213, 2012.

[8] K. Burçin and N. V. Vasif, "Down syndrome recognition using local binary patterns and statistical evaluation of the system," *Expert Systems with Applications,* vol. 38, pp. 8690-8695, 2011.

[9] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*: John Wiley & Sons, 1998.

[10] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," *Computer Vision – ECCV 2008,* vol. 5305, pp. 504-513, 2008.

[11] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition,* vol. 41, pp. 3054-3067, 2008.

[12] W. Yang, *et al.*, "Enforcing convexity for improved alignment with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, 2008, pp. 1-8.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518 vol.1.

[14] T. Ojala, *et al.*, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 971-987, 2002.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[16] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 25, pp. 804-813, 1995.

[17] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.