Medical Image Analysis 18 (2014) 699–710

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ICA

Qian Zhao [a,*], Kazunori Okada [b], Kenneth Rosenbaum [c], Lindsay Kehoe [c], Dina J. Zand [c], Raymond Sze [a,d], Marshall Summar [c], Marius George Linguraru [a,e]

[a] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Medical Center, Washington, DC, United States
[b] Computer Science Department, San Francisco State University, San Francisco, CA, United States
[c] Division of Genetics and Metabolism, Children's National Medical Center, Washington, DC, United States
[d] Department of Radiology, Children's National Medical Center, Washington, DC, United States
[e] School of Medicine and Health Sciences, George Washington University, Washington, DC, United States

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form 25 March 2014
Accepted 2 April 2014
Available online 15 April 2014

Keywords:
Down syndrome
Computer-aided diagnosis
Hierarchical constrained local model
Independent component analysis
Machine learning

ABSTRACT

Down syndrome, the most common single cause of human birth defects, produces alterations in physical growth and mental retardation. If missed before birth, the early detection of Down syndrome is crucial for the management of patients and disease. However, the diagnostic accuracy for pediatricians prior to cytogenetic results is moderate and the access to specialists is limited in many social and low-economic areas. In this study, we propose a simple, non-invasive and automated framework for Down syndrome detection based on disease-specific facial patterns. Geometric and local texture features are extracted based on automatically detected anatomical landmarks to describe facial morphology and structure. To accurately locate the anatomical facial landmarks, a hierarchical constrained local model using independent component analysis (ICA) is proposed. We also introduce a data-driven ordering method for selecting dominant independent components in ICA. The hierarchical structure of the model increases the accuracy of landmark detection by fitting separate models to different groups. Then the most representative features are selected and we also demonstrate that they match clinical observations. Finally, a variety of classifiers are evaluated to discriminate between Down syndrome and healthy populations. The best performance achieved 0.967 accuracy and 0.956 F_1 score using combined features and linear discriminant analysis. The method was also validated on a dataset with mixed genetic syndromes and high performance (0.970 accuracy and 0.930 F_1 score) was also obtained. The promising results indicate that our method could assist in Down syndrome screening effectively in a simple, non-invasive way, and extensible to detection of other genetic syndromes.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Down syndrome is the most common chromosomal condition caused by trisomy of chromosome 21. The incidence of trisomy is influenced by maternal age and differs between populations (between 1 in 319 and 1 in 1000 live births are trisomic for Hsa21) (de Graaf et al., 2011). In the United States, there are one out of 691 infants born with Down syndrome and over 400,000 people living with it (Parker et al., 2010). Down syndrome causes lifelong mental retardation, developmental delays, heart defects, and respiratory and hearing problems. The early detection and intervention of Down syndrome is fundamental for managing the disease and providing patients with lifelong medical care that may involve physical and speech therapists, cardiologists, endocrinologists and neurologists.

Down syndrome may be diagnosed before or after birth. During pregnancy, screening tests including ultrasound and blood tests can be performed. The accuracy of screening tests is estimated to be as low as 79% (Benn, 2002). If a screening test indicates a high risk of Down syndrome, a more invasive diagnostic test may be used to confirm the diagnosis. However, diagnostic tests including amniocentesis, chorionic villus sampling and percutaneous umbilical blood sampling, carry a risk of miscarriage (Skotko, 2007). Recently, the non-invasive prenatal test (NIPT) has been introduced to screen Down syndrome by performing on maternal blood (Chiu et al., 2011), which is less invasive with high accuracy. After birth, the initial diagnosis of Down syndrome is often based on a

* Corresponding author. Tel.: +1 2024761285; fax: +1 2024761270.
E-mail address: kate.qzhao@gmail.com (Q. Zhao).

http://dx.doi.org/10.1016/j.media.2014.04.002
1361-8415/© 2014 Elsevier B.V. All rights reserved.

number of minor physical variations and malformations. Some common features can be expressed as both geometry (e.g. a flattened facial profile, upward slanting palpebral fissures, small ears, protruding tongue and extremity variations) and texture (e.g. appearance of nuchal skin and epicanthic fold) differences (Wiseman et al., 2009). 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.

The imprecise and non-standardized nomenclature places a major difficulty for the communication between dysmorphologists. For infants born in non-academic centers, rural settings and internationally, access to specialists is much more limited or not readily available. The accuracy of a clinical diagnosis of Down syndrome for pediatricians prior to cytogenetic results approximates 50–60% and is likely to be lower in many instances (Sivakumar and Larkins, 2004). The delayed diagnosis of Down syndrome may lead to lack of bonding and delayed initiation of medical interventions (Skotko et al., 2013). The development and implementation of automated remote computer-aided tools to detect Down syndrome and other genetic syndromes associated with facial dysmorphosis has the potential for dramatically improving the diagnostic rate and providing early guidance to families and involved professionals.

In this work, we propose a simple, automated and non-invasive framework for Down syndrome detection based on anatomical facial landmark analysis and machine learning techniques. The technique requires only non-standardized frontal facial photographs of patients taken by any basic camera. A hierarchical constrained local model based on independent component analysis (ICA) is introduced to locate the anatomical facial landmarks. Geometric features and local texture features based on local binary patterns (LBP) and Gabor wavelet transform are extracted to describe both the morphological and structural variation between Down syndrome and a healthy population. After feature fusion and selection, several classifiers including support vector machine (SVM), *k*-nearest neighbor (*k*-NN), random forest (RF) and linear discriminant analysis (LDA) are trained to identify Down syndrome cases. Leave-one-out cross-validation is performed to evaluate the method in terms of accuracy, precision and recall. The main contributions of our work lie in: (a) the proposal of a population and disease-specific hierarchical constrained local model to locate anatomical landmarks automatically; (b) the improvement of a data-driven independent component selection method for ICA-based shape models; (c) the development of hierarchical strategy of statistical models to refine the clinical relevant anatomical landmark detection; and (d) the combination of geometric and texture features to accurately describe the morphology and structure of genetic syndromes.

2. Related work

2.1. Genetic syndrome detection

Photogrammetry could serve as a readily available and powerful tool for automated assessment of facial dysmorphology. Recently, two-dimensional (2D) and three-dimensional (3D) facial image analysis methods have been used for syndrome detection. For 2D image analysis, the authors in (Loos et al., 2003) investigated the disease-specific facial patterns for ten syndromes, excluding Down syndrome, using Gabor wavelet features. Their latest study (Boehringer et al., 2011) classified 14 syndromes with 21% accuracy. However, the method did not discriminate the syndromes from a healthy population and it requires manually labeled landmarks. Saraydemir et al. (2012) applied the Gabor wavelet transform globally to manually cropped face images. Principal component analysis (PCA) and LDA were used for feature dimension reduction. The classification accuracy achieved by *k*-NN classifier and SVM were 96% and 97.3%, respectively. But the method also needs manual image standardization including rotation and cropping. Moreover, the dataset only consisted of 15 Down syndrome and 15 normal cases, a small database for robust results. In (Burçin and Vasif, 2011), the authors separated Down syndrome from a healthy group using LBP and template matching. Still manual cropping was required in pre-processing. Moreover, they applied LBP operator on non-overlapped blocks covering the whole face which may not capture the local texture features at clinical relevant facial landmarks.

Besides 2D image analysis, 3D facial morphology modeling has been investigated to assist the diagnosis in dysmorphology (Aldridge et al., 2005; Hammond et al., 2001, 2004; Wilamowska et al., 2009). Hammond et al. investigated the 3D facial morphology modeling for Williams, Smith-Magenis, 22q11 deletion and Noonan syndromes (Hammond et al., 2004). 3D facial images were captured with commercial photogrammetric devices (3dMD and Surfirm), and each image was manually annotated with 3D landmarks. Then dense surface models were built with PCA. For discrimination testing, closest mean, LDA and SVM were adopted. The discrimination performance for the ten pairwise control-syndrome and syndrome-syndrome comparison ranged from 87% to 98%. In (Wilamowska et al., 2009), Wilamowska et al. presented a method to classify 3D face shape in 22q11.2 deletion syndrome given a set of labeled 3D meshes acquired from stereo imaging of heads. The accuracy of the classification performed on feature vectors comprised of the PCA coefficients of these representations was 76%. All the above 3D studies did not investigate Down syndrome. In addition, they are not automatic; all require manual landmark placement. Most importantly, the use of 3D digital photogrammetric image acquisition systems is costly and may not be available in many clinical centers.

Our group has also investigated Down syndrome detection using machine learning techniques (Zhao et al., 2013a,b,c). In our prior work, we showed the effectiveness of geometric and local texture features. This work is a continuation and significant extension of our previous work (Zhao et al., 2013a) in the following aspects: (1) this paper completes the compact description in our conference paper by providing details of derivatives of our method; (2) we provided additional comparisons with the most popular method AAM in terms of point-to-point and point-to-curve distance; (3) Gabor wavelet features were also analyzed and compared with LBP in terms of accuracy and F_1 score; (4) the selected geometric and texture features were analyzed from clinical perspective; (5) the proposed method was validated on a larger Down syndrome dataset, as well as an additional mixed syndrome dataset; and (6) the advantages and shortcomings of the proposed method were discussed in detail.

2.2. Shape appearance models

Conventional statistical models such as active shape model (ASM) (Cootes et al., 1995), active appearance model (AAM) (Cootes et al., 1998) and other variants (Milborrow and Nicolls, 2008; Seshadri and Savvides, 2009; Zhou et al., 2010) have been widely applied in landmark detection and image segmentation (Heimann and Meinzer, 2009; Gerig et al., 2002).

The profile of a landmark in ASM is intensity values along a line perpendicular to the tangent of the landmark, which is not sufficient to describe the intensity information around the landmark. In addition, ASM does not consider the relationship between shape and appearance models. To overcome these limitations, AAM was proposed by (Cootes et al., 1998).

Similar to AAM, the Constrained Local Model (CLM) first proposed by Cristinacce and Cootes (2008) also uses a template appearance model, but with a more robust constrained search technique. The search is performed in a local region at each landmark by jointly optimizing the local appearance and shape constraints. It has demonstrated good performance in non-rigid object alignment/tracking. Since its emergence, few variants of CLM have been developed using different optimization methods (Saragih et al., 2009; Yang et al., 2008). CLM has several advantages over AAM. First, CLM is discriminative as well as generative to unseen appearance variations. Second, CLM offers greater invariance to global illumination variation and occlusion. Importantly, CLM models the object as an ensemble of independent individual experts. Finally, it does not require complicated piece-wise affine texture warp that may be prone to errors. Our method is different from the previous work (Saragih et al., 2009; Yang et al., 2008) because of its joint quadratic optimization of SVM local texture model and ICA shape model in landmark coordinate domain (shape parameter in previous works). Lucey et al. used a similar strategy for optimization in (Lucey et al., 2009). The main difference between their work and ours is that we used the reconstruction error of ICA as the shape constraints instead direct Gaussian assumption on the shape parameters.

In all the aforementioned models, point distribution models (PDM) use PCA to describe shape variations of the training data conventionally. However, such a PCA-based shape model assumes a Gaussian distribution of input data which is not often valid and may lead to an inaccurate statistical description of shapes and generation of implausible shapes. Furthermore, the principal components (PCs) of PCA tend to represent global shape variations: changing the parameter of a PC may deform the entire extent of shape (shown in Fig. 1). The accurate localization of anatomical landmarks is required in certain applications, such as medical image analysis and expression recognition. The PCA-based shape model fails to capture localized deformations in such situations. To address the above problems, localized component analysis, describing surface shape variation in an ensemble of objects using a linear subspace of spatially localized shape components, was investigated in statistical shape models in (Alcantara et al., 2009; Giessen et al., 2011). But the localized shape components anlaysis is parametric and the parameter setting is not trivial for modeling the shape variations. Independent Component Analysis (ICA) is considered as another alternative method to build a statistical shape model. To the best of our knowledge, previous work related to constructing statistical models using ICA is scarce. Different ICA algorithms and independent component (IC) selection methods were compared for statistical shape modeling (Uzümcü et al., 2003b). The experimental results showed that the different ICA algorithms yielded very similar results and the ordering according to the locality of the shape variations achieved the best selection and classification performance. However, their criterion of sorting ICs is only based on empirical local peaks. In (Uzümcü et al., 2003a), the authors compared the ICA and PCA in AAM for cardiac MR segmentation. The ICA-based AAM outperformed PCA-based model in terms of border localization accuracy. But this technique did not present how to select a relevant subset of ICs. The ICA-based models were also investigated in 3D and 4D cardiac MR images (Faghih Roohi and Aghaeizadeh Zoroofi, 2012; Lötjönen et al., 2004; Suinesiaputra et al., 2009; Suinesiaputra et al., 2011) adopting the ICA-based statistical models proposed in (Uzümcü et al., 2003b). Our proposed ICs selection criterion uses entropy and interquartile range to measure the sample variation locality and adds a data-driven selection of IC subsets according to the criterion (non-parametric sorting of ICs). Additionally, we use ICA in a hierarchical model to take advantage of the local modeling power of ICA.

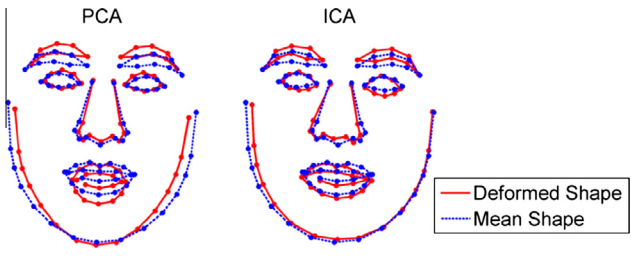

Fig. 1. Comparison of shape deformation when changing a component of PCA and ICA, respectively. The first shape parameter was set to $2\sigma_1$ and $|\omega|/2$ for PCA and ICA, respectively, and all other shape parameters were set to 0. The whole shape deforms when a principal component changed in the PCA-based model, while the ICA-based model allows for more local deformation (e.g. the shape of eyebrows and the tip of nose).

3. Methods

The framework of the proposed method, illustrated in Fig. 2, consists of landmark detection, feature extraction and selection, classification and evaluation.

3.1. Landmark detection with hierarchical constrained local model

3.1.1. Constrained local model

The procedure of locating landmarks using shape constrained local texture model (CLM) can be divided to model building and searching stages that will be described in detail.

3.1.2. Building CLM with ICA

A CLM consists of a statistical shape model and a local texture model. The shape model defines the plausible shape space and describes the shape variations of the training samples. The local texture model describes how the local region around each landmark appears. Combining these two models, both the shape morphology and local texture of a face are characterized.

3.1.2.1. ICA-based statistical shape model. A shape represented by n landmarks in two dimensions is denoted as $\mathbf{x} = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]^T$. To compensate for translation, rotation and scaling differences between the shape samples, all shapes are aligned using generalized Procrustes analysis by minimizing the squared distance between corresponding landmarks in the shape samples. The mean shape of the training set is calculated and subtracted from the aligned shapes represented by $\mathbf{x}$, where $\mathbf{x}$ now contains the new zero-mean coordinates of all shape samples resulting from alignment. Then ICA is applied to the aligned training shapes to obtain the modes of shape variation. The shape matrix X containing all centered shapes in the training set can be expressed as a mixture of independent components $X = A \cdot S$, where A is the matrix containing mixing parameters and S the independent components (ICs). It can also be written in a vector format $x = \sum_{i=1}^{n} \mathbf{a}_i s_i$, where $\mathbf{a}_i$ is the columns of mixing matrix A and s_i are the ith IC. After estimating the matrix A, the de-mixing matrix W ($W = A^{-1}$) and independent components can be computed by $S = W \cdot X$.

ICA is defined as a method that finds a linear transformation maximizing the non-Gaussianity of S. The de-mixing matrix can be computed by maximizing some measure of independency. In this study, we use the Joint Approximated Diagonalization of Eigenmatrices (JADE) method (Cardoso, 1999) for ICA. The JADE algorithm is based on the joint diagonalization of the cumulant matrices. All cumulants of order 2 and 4 are involved and a joint

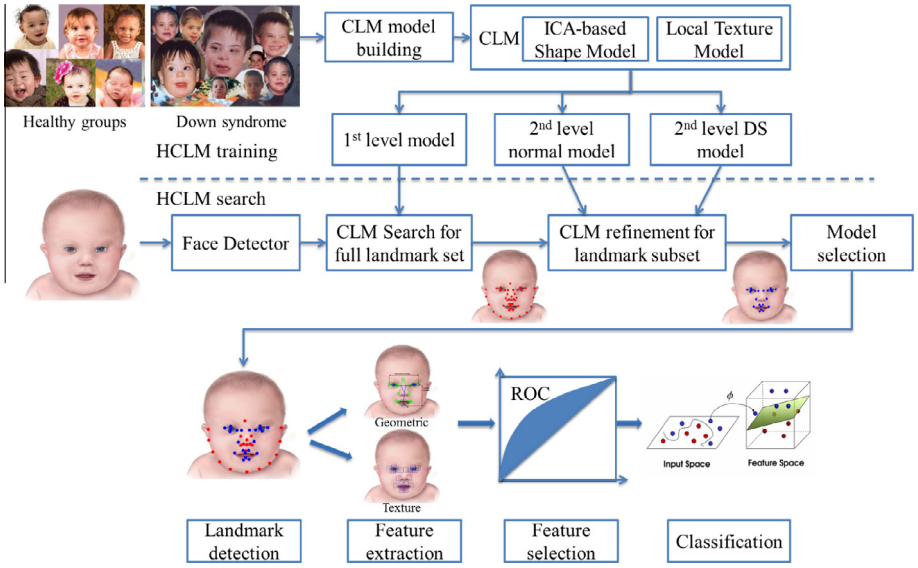

Fig. 2. The framework of the genetic syndrome detection.

diagonalization is performed with a Jacobi technique. The JADE method is non-parametric and robust (Skotko et al., 2013).

For PCA, the eigenvectors are sorted naturally according to their corresponding variances (eigenvalues), while for ICA, the variances and the order of ICs are not determined naturally based on the variance of the ICA samples. We propose to sort the columns of the mixing matrix A based on the non-parametric estimate of the sample variance and the basis locality. First the shape matrix X are projected onto each $\mathbf{a}_i$. A normalized cumulative histogram (CH) is computed from these projections. The interquartile range (IQR) (Upton and Cook, 1996), a measure of statistical dispersion, is defined as $\mathrm{IQR} = Q_3 - Q_1$, where $Q_3 = \mathrm{CH}^{-1}(0.75)$ and $Q_1 = \mathrm{CH}^{-1}(0.25)$. The statistical dispersion of the projections is determined by the range approximating $\pm 3\sigma$ range of normal distribution. The values out of this range, considered as outliers, are filtered out to obtain robust ICs selection. The dispersion measure can be regarded as a robust non-parametric estimate of sample variance along ICs. The shape deformation along $\mathbf{a}_i$ to the limit $\pm|\omega_i|/2$ is given by the norm of inner product of $\mathbf{a}_i$ and $\omega_i, \mathbf{v}_i = |\mathbf{a}_i \cdot \omega_i|$. The maximum value of $\mathbf{v}_i$ determines the significance of the peak; the entropy $H(\mathbf{v}_i)$ characterizes the locality of ICs and the dispersion range ω_i describes the shape variance. With these three factors, the criterion C to order the ICs and corresponding columns of the mixing matrix is then determined by

$$C = \omega_i \cdot v_{\max}/H(v_i), \tag{1}$$

where v_{max} is the maximum value of $\mathbf{v}_i$ and $H(\mathbf{v}_i)$ is the entropy of $\mathbf{v}_i$. For modes that describe relevant independent directions in the training set, the variations are localized and have large peaks, therefore have large values. While for noisy modes, the variations are relatively small and not localized, thus have small values. After sorting the ICs with this criterion, we select dominant ICs by removing noisy ICs with very small C values. All kept ICs have reconstruction error smaller than a threshold (0.001 in this study). Fig. 3 compares the first three principal modes of PCA and ICA. Please note that the PCA modes depict only global variations, while the ICA modes highlight local variations. In particular, the first two modes of PCA present the head poses (up–down, left–right), while the ICA modes highlight the local variations around major facial features (e.g. mouth and nose).

3.1.2.2. Local texture model. The local texture model describes the appearance of the local patch around each landmark. It is built using a patch descriptor and a support vector machine (SVM) (Cortes and Vapnik, 1995) with a linear kernel due to its computational advantages. For each landmark per training image, we extract m ($m = 25$) square patch samples throughout the training set as training data for SVM, including 12 negative and 9 positive samples. The 9 positive samples are extracted centered at each landmark with error tolerance up to 3 pixels, and the 12 negative samples include 8 samples randomly sampled in the region shifted away from the ground truth by ten pixels and 4 samples with constant intensity values (intensity ∈ $\{0, 128, 255\}$) to make the learning and search robust. All m patches have the same size (for 512×512 pixels images, the size of patch is chosen empirically as 40×40 pixels that contain sufficient information of a landmark). The descriptor for each patch is chosen as the oriented histogram of gradient (HoG) (Dalal and Triggs, 2005) with the same dimensions N ($N = 576$ for 40×40 patches). For HoG descriptor, the parameters including number of orientation bins, the cell size and the block size were set to 9, 8 and 2, respectively. Thus for 40×40 patches, there are 16 blocks; and for each block, the histogram dimension is 36 ($N = 16 \times 36 = 576$). Thus we obtain m training vectors $\{\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \ldots, \mathbf{g}^{(m)}\}$ and each training sample $\mathbf{g}^{(i)} = \left[g_1^{(i)}, g_2^{(i)}, \ldots, g_N^{(i)}\right]^T$, $i = 1, 2, \ldots, m$. Each training vector is assigned an output value for SVM, $y^{(i)} = \{-1, 1\}$, $i = 1, 2, \ldots, m$. A SVM is trained for each landmark. The decision function of linear SVM can be represented by an inner product of training data and the support vectors

$$f(\mathbf{g}^{(i)}) = \sum_{j=1}^{N_S} \alpha_j \langle \mathbf{g}_j, \mathbf{g}^{(i)} \rangle, \tag{2}$$

where $\mathbf{g}_j$ is a support vector, α_j the weight of the support vector and N_S the number of support vectors. α_j is solved by standard quadratic programing technique (Cortes and Vapnik, 1995). For CLM, the decision function of linear SVM can also be written as a linear combination of the input vector

$$f(\mathbf{g}^{(i)}) = \varphi^T \cdot \mathbf{g}^{(i)}, \tag{3}$$

where $\boldsymbol{\varphi}^T = [\varphi_1, \varphi_2, \ldots, \varphi_N]$ represents the weight for each input pixel calculated by $\boldsymbol{\varphi} = \sum_{j=1}^{N_S} \alpha_j \mathbf{g}_j$. The reason of choosing linear

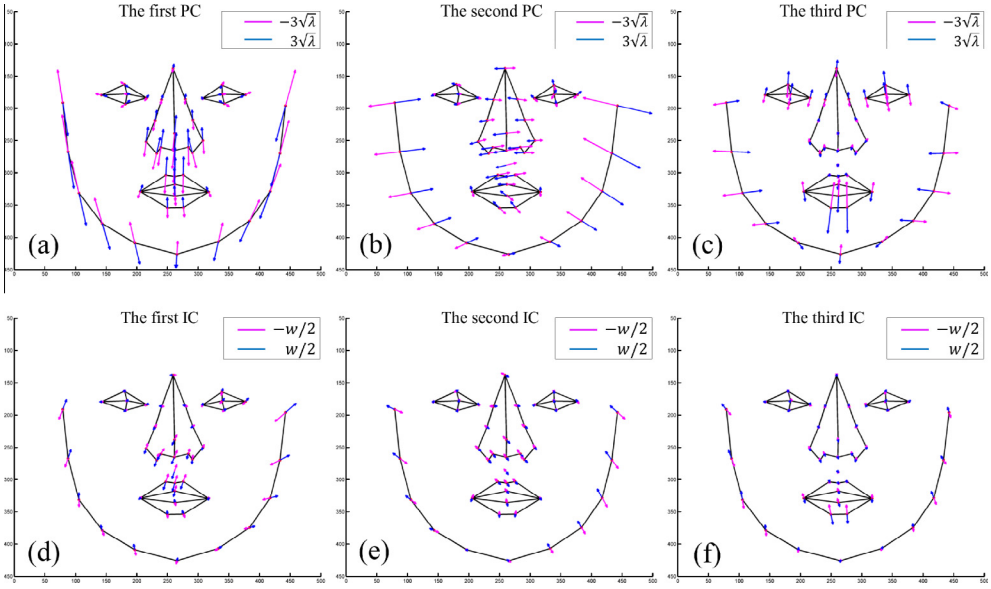

Fig. 3. The first three principal modes for PCA (a–c) and ICA (d–f) trained on both the Down syndrome and healthy groups. Note the global vs. local variations expressed by PCA and ICA, respectively.

SVM is that $\boldsymbol{\varphi}^T \cdot \mathbf{g}^{(i)}$ can be pre-computed rather than evaluated separately for each example. Therefore, it reduces the computational complexity and time dramatically by avoiding sliding window processing.

To obtain the response image of the patch, an approximate probabilistic output of SVM is computed by fitting a logistic regression function (Bishop, 2007) to the binary output of SVM

$$\Pr(y^{(i)} = 1 | f(\mathbf{g}^{(i)})) = \frac{1}{1 + \exp(uf(\mathbf{g}^{(i)}) + v)}, \tag{4}$$

where u and v are learned through a cross-validation.

3.1.3. Searching landmarks with CLM

Using CLM to search the optimal landmark locations is equivalent to estimating the landmark coordinates by maximizing the probabilities outputted by SVM with shape constraints. First an initial estimation of the landmark locations is made via a similarity transformation of the mean shape. Then each landmark is searched in the local region of its current position using the local texture model. We denote the SVM response image with $R(x, y)$ that is fitted with a quadratic function

$$r(x, y) = a(x - x_0)^2 + b(y - y_0)^2 + c, \tag{5}$$

where $\boldsymbol{\gamma} = (a, b, c)$ are the quadratic function parameters and (x_0, y_0) is the center point of the quadratic function. It can also be written in matrix format

$$r(x, y) = \mathbf{z}\Phi\mathbf{z}^T - 2\Gamma\mathbf{z}^T + ax_0^2 + by_0^2 + c, \tag{6}$$

where $\mathbf{z} = [x, y]$, $\Phi = [a, 0; 0, b]$ and $\Gamma = [ax_0, by_0]$. The parameters are solved by minimizing the sum-of-the-squares objective function

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}=(a,b,c)} \sum_{x,y} [R(x, y) - r(x, y)]^2. \tag{7}$$

Finally, the landmark positions are estimated by jointly optimizing the quadratic function and the shape reconstruction error using ICA. The joint objective function is given by

$$\begin{aligned} \mathbf{x}^* = &\arg\max_{\mathbf{x}} \mathbf{x}^T\boldsymbol{\Phi}\mathbf{x} - 2\boldsymbol{\Gamma}\mathbf{x} - \beta(\mathbf{x} - AW\mathbf{x})^T(\mathbf{x} - AW\mathbf{x}), \\ &\text{subject to} - \boldsymbol{\omega}/2 < W \cdot \mathbf{x} < \boldsymbol{\omega}/2, \mathbf{x} \in [l, u], \end{aligned} \tag{8}$$

where $\boldsymbol{\Phi} = \text{diag}(\Phi_1, \ldots, \Phi_n), \boldsymbol{\Gamma} = [\Gamma_1, \ldots, \Gamma_n], \boldsymbol{\omega}$ is the dispersion range of the projections and l, u are the lower and upper boundary of $\mathbf{x}$ determined by the search region. The objective function is solved by a standard quadratic optimization method. In a PCA-based shape model, the shape parameters are usually limited by

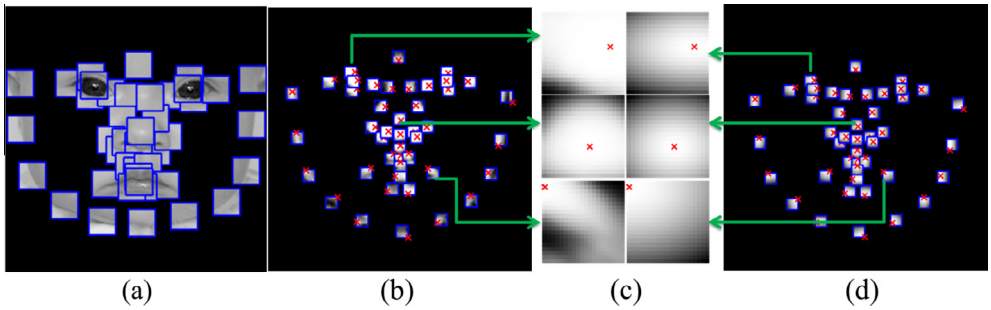

Fig. 4. Examples of local patch response map: (a) shows the local patches around landmarks; (b) shows the local search response map obtained by linear SVMs; the left column of (c) are the raw response map $R(x, y)$ from (b) and the right column of (c) are the fitted quadratic maps $r(x, y)$ from (d); (d) shows the estimated response map by fitting the quadratic function to (b). The red crosses in (b), (c) and (d) illustrate the ground truth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

three times square root of eigenvalues. To make a fair comparison, the range of ICA shape parameters $[-|\omega_i|/2, |\omega_i|/2]$ covers the same amount of sample variance as that by 3σ range with PCA given a normal distribution by recalculating $\boldsymbol{\omega}$ to cover 99.7% sample variance along ICs. Fig. 4 shows an example of local texture model in the search process.

To make the search more efficient and robust, a three-level multi-resolution search is performed. Before the search begins, we build an image pyramid (subsampled by a factor of two) and repeat the CLM search at each level, from coarse resolution (128×128 pixels) to fine resolution (512×512 pixels). The initial estimation for the first search is the shape generated from the Viola-Jones face detector (Viola and Jones, 2001). The initial estimation at each subsequent level is the best shape found by the search at the level below.

3.2. Hierarchical constrained local model

The differences between Down syndrome and a healthy population mainly lie in the inner face features around the mouth, nose and eyes. Fig. 5 shows the mean shape comparison of the normal model trained with the healthy group and the Down syndrome model trained with Down syndrome cases. From the point-wise mean shape difference we can see that the nose and mouth regions show bigger differences. So it is desirable to build separate models for the Down syndrome and healthy groups to refine the locations of inner face anatomical landmarks.

We propose a hierarchical constrained local model (HCLM) consisting of two levels. The first level CLM is trained using the full landmark set and all images including all groups (normal and Down syndrome). It roughly locates all of the facial landmarks. For the second level, separate CLMs are trained using inner facial anatomical landmarks (landmark subset on the mouth, nose and eyes) for respective groups.

The first level model is trained using the total M training samples. Its shape and local texture model are built with the full landmark set $\mathbf{x}^F$ containing n_F landmarks. Suppose there are K groups and for each group there are $M^{(i)}$ training samples $i = 1, \ldots, K$. Thus the second level consists of K CLMs that are trained with landmark subset $\mathbf{x}^S (\mathbf{x}^S \subset \mathbf{x}^F)$ containing n_S landmarks. Using this hierarchical model, the first level helps to locate landmarks using the power of a larger and general dataset (Down syndrome + healthy). Then the second level refines the landmarks using best fitted models of anatomy and disease (Down syndrome vs. healthy).

The landmarks are first estimated using the first level CLM. The results of the landmark subset from the first level serve as the initialization of the second level search. Then the shape parameters are estimated using the landmark subset $\mathbf{x}^S$ resulting from the second level search. Finally, the full landmark set $\mathbf{x}^F$ is re-estimated using the updated shape parameters. The above searching process is repeated for all K CLMs in the second level. The model selection is based on the assumption that the sample distribution of separate groups should be different. To validate the assumption, we performed paired Hotelling's T-square test (Hotelling, 1992) on Down syndrome data and normal data, and the difference of these two samples were significant ($p = 0.047$). The best fitted model is selected as the one whose result is closer to its own mean shape and holds smaller changes compared with the first level search results

$$CLM^* = \underset{CLM, i=1,\ldots,K}{\arg\min} \left| f(CLM^{(i)}) - \bar{\mathbf{x}}^{(i)} \right|, \quad \text{subject to } \left| f\left(CLM^{(i)}\right) - \mathbf{x}^0 \right| < \delta, \text{ for } i = 1, \ldots, K. \tag{9}$$

3.3. Feature extraction, selection and classification

Patients with Down syndrome present special facial morphology that relates both to shape or geometry (e.g. upward slanting eyes, small nose and wide-opened mouth) and texture (e.g. flattened philtrum and prominent epicanthic folds) (Sivakumar and Larkins, 2004). To describe these two types of information, geometric and texture features are extracted on the registered patient image which is aligned to a reference image using Procrustes analysis (Gower, 1975) to remove the translation, in-plane rotation and scaling. Geometric features are defined via interrelationships among anatomical landmarks to incorporate clinical criteria for Down syndrome diagnosis. Geometric features include Euclidean distances between the landmarks and corner angles spanned by the landmarks, shown in Fig. 6(a). The Euclidean distances can be divided to horizontal and vertical lines according to their directions. The horizontal and vertical lines are both normalized by their baselines, respectively. The horizontal baseline is the distance between the left corner of left eye and the right corner of right eye (the width of the face), and the vertical baseline is the vertical distance between the eyes and the lower lip (the height of the face). The normalized geometric features are invariant to scale, translation and rotation. The angles are analyzed via linear statistics despite the non-linear nature as all of them are acute angles

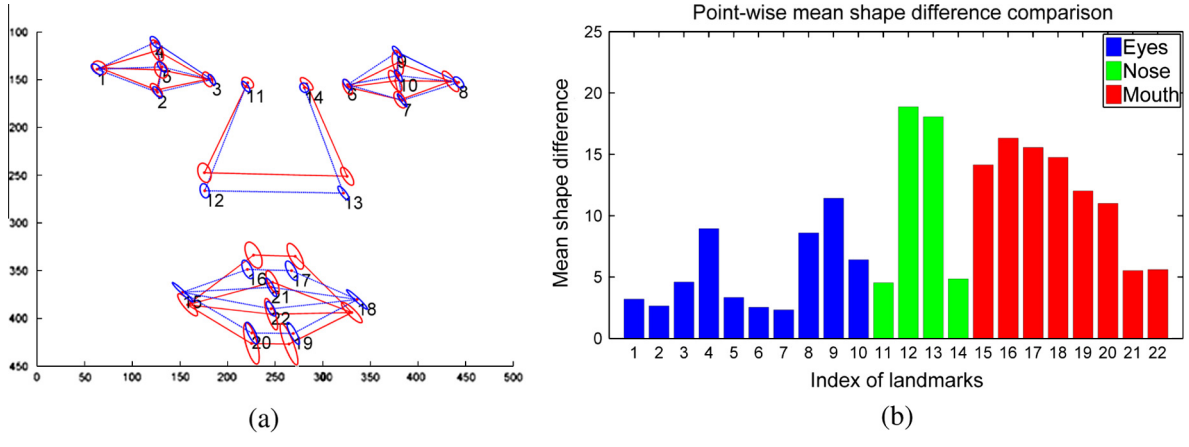

Fig. 5. The mean shape comparison on inner face landmarks between the Down syndrome and healthy groups: (a) shows the definition of inner face landmarks and compares the mean shapes qualitatively, where the ellipses represent the point distribution (covariance) for each landmark; (b) shows the mean shape differences around different facial regions. Nose and mouth regions present bigger differences.

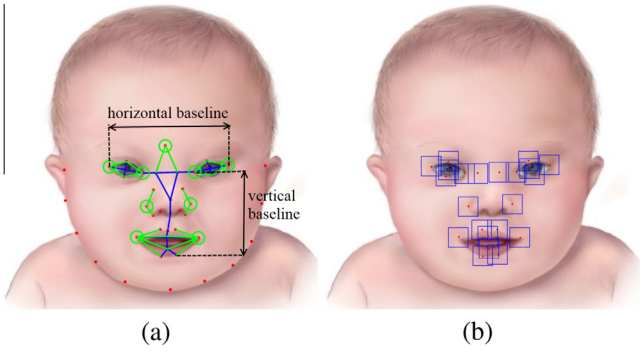

Fig. 6. Feature extraction: (a) The graphic definition of geometric features; the blue lines are normalized by the vertical or horizontal baseline and the green circles illustrate the location of corners; (b) shows the region of interests (blue squares) of inner face landmarks for local texture feature extraction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(less than 90 degree) according to their definitions; therefore, the angle features are monotonic in our application. There are a total of 27 geometric features.

Two types of local texture features are extracted to delineate facial features. One is based on LBP histogram (Ojala et al., 1996, 2002), and the other is based on the Gabor wavelet transform (Ma and Manjunath, 1995). For LBP, we use the uniform LBP (Ojala et al., 2002) in which the number of neighboring sample points P on a circle of radius R is not limited. In this study, we set $P = 8$ and $R = 1$ We also tested other settings of P and R (e.g. $P = 16, R = 2$), however the performance of the technique was inferior. Thus in the following experiments we use $P = 8$ and $R = 1$ setting for LBP-based texture features.

To extract local texture features, a LBP histogram is first computed from a square region of interest (ROI) around each of the 22 inner facial landmarks (on mouth, nose and eyes, Fig. 6). The LBP histogram contains information about the distribution of the local micro-patterns, such as edges, spots and flat regions. Then six first-order statistical measurements of the ROI LBP image are computed from the histogram, which are the mean, variance, skewness, kurtosis, energy and entropy. Finally, the feature vectors in all ROIs are concatenated to form the LBP-based local texture features for the image. Thus, the local texture features also contain the spatial information of the texture. There are a total of 132 LBP-based texture features.

Besides LBP-based texture features, we also present texture information with the Gabor wavelet because the Gabor features are robust against local distortions caused by variance of expression, illumination and pose (Saraydemir et al., 2012). Gabor filters provide the optimal localization of spatial and frequency information and have been widely applied to face recognition (Jadhav and Holambe, 2009; Srinivasan and Ravichandran, 2013). In this study, five scales and eight orientations are used to extract features, which lead to 40 Gabor wavelets. The 880-dimensional Gabor jet is generated by concatenating the magnitudes of 40 complex-valued Gabor features on each of the 22 inner facial landmarks. More details on the Gabor wavelet transform can be found in (Ma and Manjunath, 1995; Okada et al., 1998).

The geometric features are concatenated with the LBP-based texture features to 159 combined features (Geo + LBP), and with the Gabor jet to 907 combined features (Geo + Gabor). Feature selection is performed using the method in (Cai et al., 2010). The method is based on manifold learning and $\mathscr{L}1$ regularized models for subset selection. Specifically, the method selects the features such that the multi-cluster structure of the data is well preserved. By using spectral analysis techniques, the feature selection method measures the correlations among different features without label information. The corresponding optimization problem only involves a sparse eigen-problem and a $\mathscr{L}1$-regularized least square problem, thus can be solved easily. The optimal dimension for feature space is found based on maximizing the area under the receiver operating characteristic (AUROC) curves by empirical exhaustive search.

After feature extraction and selection, we compare various classifiers for Down syndrome detection, including linear, non-linear and tree-structure classifiers. For linear classifiers, we use linear SVM (Cortes and Vapnik, 1995) and LDA (Mika et al., 1999). For non-linear ones, SVM with radial-basis function (RBF) kernel and k-NN (Denoeux, 1995) are adopted. We also utilize a state-of-the-art tree-structure classifier, the random forests (RF) (Breiman, 2001). The performance of these classifiers is tight to the distribution of data. Since the high dimensional features cannot be simply estimated, the performances of these classifiers may provide us some information of the distribution of features.

SVM is a robust and powerful classifier that deals well with high-dimensional data. It uses a kernel function to map the data into a high-dimensional feature space. When the kernel function is a first-order polynomial, the SVM behaves as a linear classifier, while when the kernel function is RBF, the SVM becomes a non-linear classifier. LDA is another commonly used linear classifier that maximizes the ratio of between-class variance to within-class variance in the data thereby guaranteeing maximal separability. LDA can easily handle cases where the within-class frequencies are unequal. Alternatively, a k-NN classifier is a simple, stable and efficient non-linear classifier that is based on non-parametric density estimation. It assigns a label to a sample from the predominant label among its k nearest neighbors. Differently, the random forest is an ensemble classifier that consists of multiple decision trees. Each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest. Random forest is one of the most accurate learning algorithms and robust with respect to noise. The parameters for the SVM (C, σ), k-NN (k) and RF (the number of trees) are found optimally by grid search (Bergstra and Bengio, 2012).

4. Experiments

4.1. Data and evaluation metrics

The performances of AAM, PCA-based CLM, ICA-based CLM and ICA-based HCLM were compared on the Down syndrome dataset containing 130 frontal facial images from 130 different subjects (one image per subject), 50 of who are Down syndrome pediatric patients and 80 are healthy individuals. Photographic data acquisition and processing with a variety of cameras and under variable illumination, expression and poses was approved by the Institutional Review Board (IRB) and Children's National Medical Center. The subjects are from multiple ethnicities including 98 Caucasian, 20 African American and 12 Asian and both genders (86 males and 44 females). The age of patients varies from 0 to 3 years. Leave-one-subject-out experiments were conducted on the DS dataset as a single procedure throughout the framework including SVM texture model training, group CLM learning, feature selection and classification. As there is one image per subject in our dataset, leave-one-subject-out, same as leave-one-out cross-validation, leaves one subject out each time for testing and the rest of subjects are used for training. This makes sure that the test face is not known by the model.

We also validated the method on a mixed syndromes dataset including 24 images from 24 different subjects with 14 dysmorphic syndromes including Klinefelter syndrome, Charge syndrome, Fetal

alcohol spectrum disorder, Facio-Auriculo-Vertebral dysplasia, Stickler syndrome, Treacher Collins syndrome, Noonan syndrome, Williams syndrome, Velo-Cardio-Facial syndrome, Trisomy 18, Beckwith Wiedemann syndrome, Rubinstein-Taybi syndrome, Waardenburg syndrome and Cerebral Gigantism (Sotos syndrome) (Swarts, 2009). There were 16 Caucasian, 6 African American and 2 Asian patients aged from 0 to 5 years. As control group, we use the same 80 healthy subjects described before.

For the first level CLM, we define 44 anatomical landmarks to cover the entire face including eyes (10 landmarks), nose (13 landmarks), mouth (10 landmarks) and face contour (11 landmarks). For the second level, 22 landmarks are retained to cover the inner facial features (shown in Fig. 5(a)).

For landmark detection evaluation, two error measures are computed. A distance is computed between points using automated methods and manually labeled ground truth and then normalized by the distance between the two pupils. This metric is used as the point-to-point error measure

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N_I}\sum_{x\in I} d(\tilde{x},x)/d_n\right), \tag{10}$$

where I is the set of inner face landmarks (around the eyes, nose and mouth), $d(\tilde{x},x)$ is the Euclidean distance between inner face landmarks located by the automatic search and the corresponding ground-truth landmarks placed manually by experts, d_n is the distance between the two pupils as normalizing factor, and N is the number of images.

Another error measure, the point-to-curve error, is calculated as the distance from the landmarks obtained by the automated method perpendicularly to the piece-wise linear curve defined by the ground truth, which is also normalized by the distance between the pupils. The curve is estimated by a linear chordal approximation (Horst and Beichel, 1997).

For Down syndrome and mixed syndrome detection, the conventional accuracy and F_1 score (the harmonic mean of precision and recall) are used for evaluation.

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}, \tag{11}$$

$$F_1 = \frac{2\cdot tp}{2tp+fp+fn}, \tag{12}$$

where tp, tn, fp, and fn correspond to true positive, true negative, false positive, and false negative, respectively.

4.2. Landmark detection

In all experiments, we first detected the face, eyes and tip of the nose in the image using the Viola-Jones face detector (Viola and Jones, 2001) to initialize the first level HCLM. The accuracy of face detection was 0.987. The images in the training data have various resolutions. In pre-processing, all images were cropped to 512×512 pixels keeping the detected faces in the center of the image and leaving adequate margins (1/6 of the image size). The local patch sizes were set to 40 and 32 pixels for the first and second level HCLM, respectively. The linear SVM parameter C in local texture model was selected as 0.005. The shape regularization parameter β was empirically set to 0.01.

We compared four statistical models: AAM, PCA-based CLM, ICA-based CLM and ICA-based HCLM. To learn the appearance model used for this comparison, AAMs were built to cover 99.7% shape variance in the training samples. In this study, 47 ICs were selected for Down syndrome dataset and 23 ICs for mixed syndrome dataset. Table 1 shows the point-to-point and point-to-curve landmark detection errors. The significance was computed using the Wilcoxon rank-sum test (Hollander and Wolfe, 1999) that is a non-parametric test for two populations when samples are independent. It can be seen that the ICA-based HCLM achieved the best performance in terms of both point-to-point and point-to curve errors. The performance of all CLM methods is significantly better than AAM ($p < 0.001$). A significant improvement was also recorded by using ICA-based HCLM vs. PCA-based CLM ($p = 0.001$ for point-to-point error and $p < 0.001$ for point-to-curve error), and by using ICA-based HCLM vs. CLM ($p = 0.002$ for point-to-point error and $p = 0.003$ for point-to-curve error). An improvement (not significant, $p = 0.89$ for point-to-point error and $p = 0.74$ for point-to-curve error) was also noted when ICA was used instead of PCA with CLM.

Table 1
Inner facial landmark detection error normalized to the distance between pupils. Bold values are the maximum values of the specific metric.

Model	Point-to-point	Point-to-curve
AAM	0.085 ± 0.071	0.065 ± 0.064
PCA-based CLM	0.036 ± 0.037	0.026 ± 0.035
ICA-based CLM	0.034 ± 0.030	0.024 ± 0.025
ICA-based HCLM	$\mathbf{0.031 \pm 0.030}$	$\mathbf{0.021 \pm 0.025}$

Fig. 7 shows the normalized errors per region and per group. ICA-based HCLM achieved the best performance for both 22 inner facial landmarks and all 44 facial landmarks. AAM performed poorly on the Down syndrome group that has large shape variations. As seen in Fig. 7(b), the advantage of our method is particularly obvious on diseased populations where there is large variability in the facial model. The cumulative distributions of error measures are shown in Fig. 8.

4.3. Genetic syndrome detection

The results below are obtained using ICA-based HCLM, the method that performing best on landmark detection. For feature selection, we use selected features for linear SVM as examples. Table 2 shows the automatically selected and ranked geometric features for Down syndrome and mixed genetic syndromes detection. The selected features are consistent with the clinical symptoms of Down syndrome, for example, the top three ranked features of Down syndrome correspond to the clinical findings of Down syndrome: upward slanting eyes, small nose and narrow palpebral fissure.

Leave-one-subject-out cross-validation was performed for Down syndrome and mixed syndromes detection, respectively. As there is only one photo for each subject in the dataset, the experiments are also leave-one-subject-out. Tables 3 and 4 show Down syndrome detection results. The highest accuracy of 0.967 was achieved by the combined features with geometric and Gabor jet (Geo + Gabor) using either SVM with RBF or LDA. The highest F_1 score of 0.956 was achieved by the combined features with geometric and LBP-based texture features (Geo + LBP) using LDA. The sensitivity and specificity of the Geo + Gabor with SVM-RBF method to detect Down syndrome (see Table 3) were 0.977 and 0.962, respectively, at the ROC point of highest accuracy. Similarly, Geo + Gabor with LDA achieved sensitivity and specificity of 1.000 and 0.949, respectively. All the metrics improved when combining geometric and texture features, but not significantly using Fishers exact test ($p = 0.058$ for combined features vs. geometric features) (Fisher, 1922). However, all classifiers achieved competitive performance. The ROC curve of different features with LDA is shown in Fig. 9(a). The AUROC was 0.976, 0.972, 0.990, 0.990 and 0.991 for geometric, LBP-based texture, combined features with geometric and LBP-based texture, Gabor jet and combined features with geometric and Gabor jet, respectively.

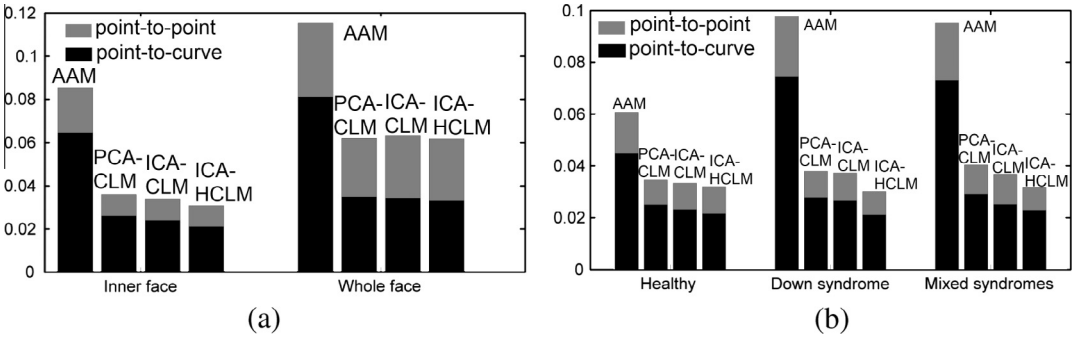

Fig. 7. Landmark detection errors: (a) The average normalized errors for 22 inner facial landmarks and 44 all landmarks, respectively; (b) the average normalized errors for the healthy group, Down syndrome and mixed syndromes, respectively. The ICA-based HCLM outperformed AAM, PCA- and ICA-based CLM in terms of both point-to-point and point-to-curve error.

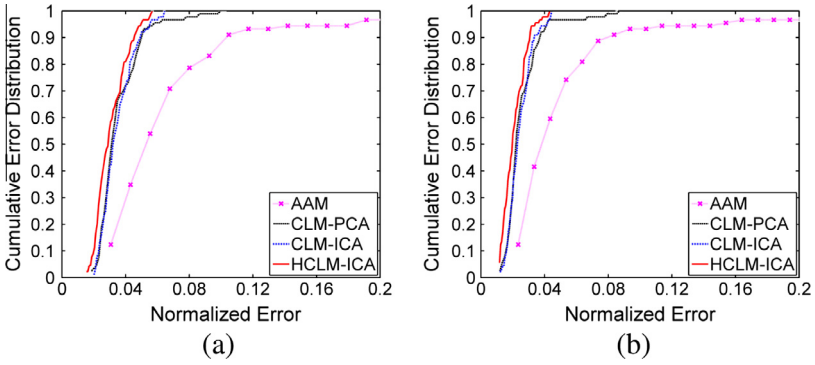

Fig. 8. The cumulative error distribution for AAM, PCA-, ICA-based CLM and ICA-based HCLM: (a) point-to-point error (b) point-to-curve error.

Table 2
The selected top ranked geometric features and their clinical relevance.[a]

Geometric features	Feature selection and ranking	
	Down syndrome	Mixed syndromes
Width of left eye	17	–
Inner eye distance	8	5
Width of right eye	19	4
Width of mouth	12	10
Left palpebral fissure	14	1
Right palpebral fissure	3	–
Length of upper nose	2	–
Length of lower nose	4	2
Length of philtrum	7	–
Thickness of upper lip	5	–
Openness of mouth	18	–
Thickness of lower lip	16	9
Orientation of left eye	6	–
Orientation of right eye	1	–
Left corner angle of right eye	10	–
Right corner angle of right eye	–	3
Upper corner angle of nose	13	7
Left corner angle of nose	11	–
Right corner angle of nose	–	6
Outer left corner angle of mouth	9	–
Inner left corner angle of mouth	15	–
Inner right corner angle of mouth	–	8

[a] The first column presents the anatomical description of the selected geometrics features. The second column denotes the ranks of the 19 selected features for Down syndrome detection. The third column denotes the ranks of the 10 selected features for the detection of mixed genetic syndromes.

Tables 5 and 6 show the experimental results for detecting mixed syndromes. The highest accuracy of 0.970 was achieved by the combined features with geometric and LBP-based texture features (Geo + LBP) using linear SVM. The highest F_1 score 0.930 was obtained by the combined features with geometric and LBP-based texture (Geo + LBP) using linear SVM. The sensitivity and specificity of the Geo + LBP with linear SVM method to detect mixed genetics syndromes (see Table 5) were 0.909 and 0.987, respectively, at the ROC point of highest accuracy. As with the Down syndrome data, all the metrics improved when combining geometric and texture features. Significant improvements, computed with linear SVM using Fishers exact test (Fisher, 1922), were recorded by using LBP-based texture vs. geometric features ($p = 0.04$), and by using combined (Geo + LBP) vs. geometric features ($p = 0.007$). Performances with other features were not significant.

Table 3
Accuracy of Down syndrome detection using different features and classifiers. Bold values are the maximum values of the specific metric.

Accuracy	Geometric	LBP	Geo+LBP	Gabor	Geo+Gabor
SVM-RBF	0.934	0.926	0.959	0.934	**0.967**
Linear SVM	0.950	0.934	0.959	0.909	0.950
k-NN	0.901	0.934	0.901	0.917	0.926
RF	0.868	0.884	0.909	0.909	0.934
LDA	0.901	0.917	0.959	0.950	**0.967**

Table 4
F_1 score of Down syndrome detection using different features and classifiers. Bold values are the maximum values of the specific metric.

F_1 score	Geometric	LBP	Geo+LBP	Gabor	Geo+Gabor
SVM-RBF	0.905	0.894	0.943	0.905	0.955
Linear SVM	0.929	0.902	0.943	0.874	0.930
k-NN	0.857	0.900	0.860	0.881	0.892
RF	0.784	0.829	0.867	0.864	0.905
LDA	0.867	0.889	0.945	0.929	**0.956**

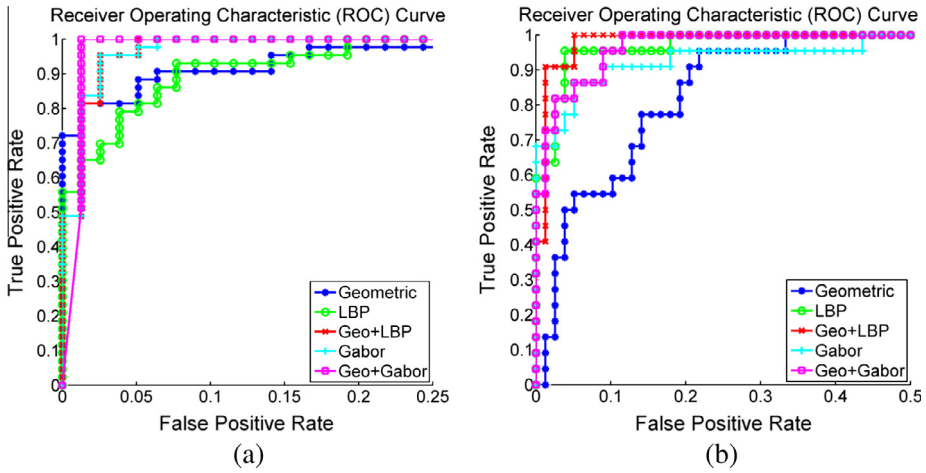

Fig. 9. The ROC curves of different features (a) using LDA for Down syndrome and (b) using linear SVM for mixed syndrome detection, respectively.

Table 5
Accuracy of mixed syndromes detection using different features and classifiers. Bold values are the maximum values of the specific metric.

Accuracy	Geometric	LBP	Geo+LBP	Gabor	Geo+Gabor
SVM-RBF	0.840	0.930	0.940	0.930	0.920
Linear SVM	0.850	0.930	**0.970**	0.920	0.940
k-NN	0.830	0.940	0.960	0.880	0.920
RF	0.890	0.920	0.910	0.880	0.900
LDA	0.810	0.960	0.960	0.830	0.890

Table 6
F_1 of mixed syndromes detection using different features and classifiers.

F_1 score	Geometric	LBP	Geo+LBP	Gabor	Geo+Gabor
SVM-RBF	0.600	0.837	0.870	0.821	0.810
Linear SVM	0.595	0.829	**0.930**	0.789	0.857
k-NN	0.541	0.850	0.905	0.667	0.778
RF	0.686	0.789	0.757	0.667	0.722
LDA	0.655	0.905	0.909	0.667	0.744

Fig. 9(b) shows the ROC for the mixed syndromes detection using linear SVM. The AUROC of geometric, LBP-based texture, combined features with geometric and LBP-based texture, Gabor jet and combined features with geometric and Gabor jet were 0.904, 0.981, 0.989, 0.960 and 0.980, respectively. The performance of geometric features for mixed syndromes detection degraded compared to Down syndrome, probably due to the fact that multiple syndromes with different symptoms were included in the mixed syndrome dataset.

5. Discussion

A framework for automated detection of genetic syndromes, in particular Down syndrome, has been proposed. The supervised scheme includes facial landmark detection, feature extraction, feature selection and classification. The approach can easily be adapted to the detection of other genetic syndromes associated with facial dysmorphosis.

Conventional statistical models such as ASMs, AAMs and their variants have been widely applied to face alignment. However, they suffer from some limitations. ASMs only consider the object boundary, but not the sufficient appearance information of the object. Additionally, they assume independence of the shape and appearance models. On the other hand, AAMs need complicated piece-wise affine texture warp operations that may be prone to errors. Moreover, they often perform poorly on unseen faces. Traditionally, both ASMs and AAMs use PCA to build statistical shape models and may not always accurately capture local shape variations (Ballester et al., 2005; Reyes et al., 2009). Compared to ASMs and AAMs, CLMs have several advantages: (a) better discriminative and generative on unseen appearance; (b) more invariant to global illumination variation and occlusion; and (c) modeling objects as ensembles of low dimensional independent local texture patches. From Table II, it can be seen that CLMs outperformed AAM significantly ($p < 0.001$) in every group and region (Fig. 9), especially for the Down syndrome and mixed genetic syndrome datasets, which have large shape variations. The superiority of CLMs is also supported by the cumulative error distribution in Fig. 8 and the fact that the area under curve of CLMs is much larger than that of AAM.

To address the inaccuracies in shape description by PCA, an ICA-based model was proposed in this paper. Unlike using PCA, the ICA-based model does not assume Gaussian distribution on data and can capture local shape variations. To select dominant ICs in terms of sample variance and locality, a novel non-parametric data-driven ordering method was proposed based on entropy. The entropy served as a good metric for locality of ICs. The ordering criterion considered data variance, data magnitude and locality of ICs. We also validated the ordering method on PCA. The ordering of PCs by entropy was identical to that by eigenvalues (variance). The principal modes of ICA shown in Fig. 3 depict local variations instead of global changes as PCA does. From Table 1, ICA-based CLM outperformed PCA-based CLM, but not significantly. The reason may be the trading off β between the shape constraints and data-driven likelihood in optimization (8). In this paper, β was empirically set to 0.01 which puts more reliability onto the data-driven term than shape constraints. Besides, ICA is a statistical description of the data that needs large datasets to train. As the ICs of ICA are not orthonormal, ICA needs high dimensions to describe shapes. Our current dataset containing hundreds of images may still be limited to represent the shape space, especially for the diseased populations of mixed genetic syndromes. But data collection is on-going and a more accurate shape space description can be expected using more training samples.

The different facial morphology between healthy subjects and patients with genetic syndromes lead to two clusters in the shape space indicated by the big difference of mean shapes between healthy groups and Down syndrome shown in Fig. 5(a). To delineate the shapes more accurately in different groups (i.e., healthy, Down syndrome and mixed genetic syndromes), a two-level hierarchical CLM was proposed. The first level locates landmarks using the power of a larger and general dataset. Then the second level

refines the landmarks using best fitted models of anatomy and disease. The ICA-based HCLM outperformed CLMs significantly ($p < 0.01$), as shown in Table 1 and Fig. 7. The cumulative error distribution curves in Fig. 8 show that the ICA-based HCLM has the largest area under curve for the detection of genetic syndromes.

The landmark detection performance depends on image quality and shape variations. Large image noise (i.e., non-discriminative features on landmarks) or large shape variation can cause poor search results. However, at the second level of HCLM, the accuracy of landmark detection improved substantially. Conceptually, our method targets classes of data in a hierarchical fashion that can be generalized to applications that have well defined priors, as well as to the segmentation and registration of images.

There are general technical shortcomings that apply to the ICA-based HCLM as well. First, the limitations of the classifiers used to model local texture may lead to false positives. Subsequently, the optimization can converge to false locations or local minima. This usually happens on landmarks that lack prominent texture features, such as landmarks between the eyebrows and the nose tip in our application. Secondly, HoG is an intensity-based descriptor and its performance is influenced by local illumination changes. Finally, the quadratic fitting can be treated as a form of smoothing to the response map that may induce poor approximations. For Down syndrome detection using machine learning techniques, the combined features of geometric and Gabor jet (though very similar to the combined geometric and LBP-features) achieved the highest accuracy of 0.967 and F_1 score of 0.956 using LDA. The overall accuracy of the automatic Down syndrome detection system was 0.954, given by factoring the best classification accuracy of 0.967 with the best face detection accuracy of 0.987. The face detection can be improved in the future by using a controlled protocol for the acquisition of photographs. Our method based on 2D facial image analysis outperformed the methods based on 3D facial modeling whose highest performance of control-syndrome discrimination was 87% (Hammond et al., 2004). Similar conclusions could be made for mixed genetic syndromes detection. An overall accuracy of 0.957 was obtained by correcting the mixed syndromes detection by the factor of face detection accuracy. Although the current mixed syndrome dataset is small, promising results of the experiments proved the feasibility of the framework extensible to more general genetic disorders.

The geometric and texture features had similar performances for our syndrome classification tasks. The geometric features are associated with specific genetic syndromes, such as the orientation of eyes, the asymmetry of the face, the size of nose normalized to the size of the face and palpebral fissure. They are suitable to describe the facial morphology of most genetic syndromes, as indicated by their good performance on the mixed genetic syndromes. Different studies found different preferences to geometric features. In this study, the feature selection leads to "explainable geometric features" that correspond to clinical evidence for DS diagnosis, such as upward slanting eyes, small nose, flat profile and protruding tongue. Such clinical related feature selection not only confirms the clinical findings of DS, but also brings new knowledge and opportunities to the genetic screening field. The texture features generally had a slightly better performance than geometric features, probably due to the fact that anatomical information is already embedded in the computation of the texture features since they were extracted around each detected landmark. LBP-based texture features and Gabor jet achieved comparable performances. All metrics improved when combining the geometric and texture features. Moreover, all classifiers obtained competitive performance, but the linear classifiers (linear SVM and LDA) outperformed slightly the non-linear ones (k-NN and SVM with RBF). It may indicate that the data is linearly separable.

The proposed framework can be applied to a range of other medical image analysis tasks. The current method applied to 2D images can be extended to 3D facial imaging which may provide more critical features and higher accuracy. The hierarchical statistical model could be similarly applied to image segmentation and registration. In this hierarchical fashion, the first level model helps locate the landmarks (i.e., rough segmentation or alignment) using the power of a larger and general dataset; and the second level refines the application using the best fitted model of anatomy or disease. The framework for genetic syndromes detection follows a standard computer-aided detection scheme that can be extended to other disorders associated with dysmorphosis/malformation. Moreover, disease severity can be analyzed by using regression techniques based on disease-specific features. Finally, models of personalized disease development could be analyzed using HCLM.

6. Conclusions

We proposed a framework for automated Down syndrome detection using non-standard frontal facial photographs of patients. Facial landmarks were located using an ICA-based hierarchical constrained local model. To sort the ICs with respect to data variance in ICA, we also proposed a non-parametric data-driven ordering method based on entropy. The ICs sorted by the proposed ordering method highlighted local shape variations. The two-level structure of HCLM significantly improved the accuracy of landmark detection over CLM. Based on the detected facial landmarks, disease specific geometric features and local texture features were extracted and selected to describe the facial morphology and appearance. Finally, several classifiers were employed to discriminate between the Down syndrome and a healthy population. The highest accuracy of 0.967 and F_1 score of 0.956 were achieved by the combined features with geometric and Gabor jet using LDA. The framework was also validated on a more challenging dataset with mixed genetic syndromes. In spite of the large diversity of the mixed syndrome dataset, the combined features with geometric and LBP-based texture features achieved 0.970 accuracy and 0.930 F_1 score. The automated computer-aided screening of genetic disorders from simple home photography of variable quality can bring genetic expertise in areas without access to specialized clinics and create affordable, instant and accurate solutions for doctors worldwide.

Acknowledgements

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.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.media.2014.04.002.

References

Alcantara, D., Carmichael, O., Harcourt-Smith, W., Sterner, K., Frost, S., Dutton, R., Thompson, P., Delson, E., Amenta, N., 2009. Exploration of shape variation using localized components analysis 31, 1510–1516.

Aldridge, K., Boyadjiev, S.A., Capone, G.T., DeLeon, V.B., Richtsmeier, J.T., 2005. Precision and error of three-dimensional phenotypic measures acquired from 3dmd photogrammetric images. Am. J. Med. Genet. 138A, 247–253.

Gonzalez Ballester, M., Linguraru, M.G., Reyes Aguirre, M., Ayache, N., 2005. On the adequacy of principal factor analysis for the study of shape variability. In: Proc. SPIE, pp. 1392–1399.

Benn, P.A., 2002. Advances in prenatal screening for down syndrome: I. general principles and second trimester testing. Clin. Chim. Acta 323, 1–16.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Bishop, C., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

Boehringer, S., Guenther, M., Sinigerova, S., Wurtz, R.P., Horsthemke, B., Wieczorek, D., 2011. Automated syndrome detection in a set of clinical facial photographs. Am. J. Med. Genet. 155, 2161–2169.

Breiman, L., 2001. Random forests. Machine Learn. 45, 5–32.

Burçin, K., Vasif, N.V., 2011. Down syndrome recognition using local binary patterns and statistical evaluation of the system. Expert Syst. Appl. 38, 8690–8695.

Cai, D., Zhang, C., He, X., 2010. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In: Unsupervised Feature Selection for Multi-cluster Data. ACM, New York, NY, USA, pp. 333–342.

Cardoso, J., 1999. High-order contrasts for independent component analysis. Neural Comput. 11, 157–192.

Chiu, R.W.K., Akolekar, R., Zheng, Y.W.L., Leung, T.Y., Sun, H., Chan, K.C.A., Lun, F.M.F., Go, A.T.J.I., Lau, E.T., To, W.W.K., Leung, W.C., Tang, R.Y.K., Au-Yeung, S.K.C., Lam, H., Kung, Y.Y., Zhang, X., van Vugt, J.M.G., Minekawa, R., Tang, M.H.Y., Wang, J., Oudejans, C.B.M., Lau, T.K., Nicolaides, K.H., Lo, Y.M.D., 2011. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma dna sequencing: large scale validity study. BMJ, 342.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Understand. 61, 38–59.

Cootes, T., Edwards, G., Taylor, C., 1998. Active appearance models. In: ECCV98, pp. 484–498.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learn. 20, 273–297.

Cristinacce, D., Cootes, T., 2008. Automatic feature localisation with constrained local models. Pattern Recogn. 41, 3054–3067.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: CVPR 2005, vol. 1, pp. 886–893.

de Graaf, G., Haveman, M., Hochstenbach, R., Engelen, J., Gerssen-Schoorl, K., Poddighe, P., Smeets, D., van Hove, G., 2011. Changes in yearly birth prevalence rates of children with down syndrome in the period 1986–2007 in the Netherlands. J. Intellect. Disabil. Res. 55, 462–473.

Denoeux, T., 1995. A k-nearest neighbor classification rule based on dempster-shafer theory. IEEE Trans. Syst. Man Cybernet. 25, 804–813.

Faghih Roohi, S., Aghaeizadeh Zoroofi, R., 2012. 4D statistical shape modeling of the left ventricle in cardiac mr images. Int. J. Comput. Assist. Radiol. Surg., 1–17.

Fisher, R.A., 1922. On the interpretation of χ^2 from contingency tables, and the calculation of p. J. Roy. Stat. Soc. 85, 87–94.

Gerig, G., Styner, M., Szekely, G., 2002. Statistical shape models for segmentation and structural analysis. In: ISBI 2002, pp. 18–21.

Giessen, M., Raedt, S., Stilling, M., Hansen, T., Maas, M., Streekstra, G., Vliet, L., Vos, F., 2011. Localized component analysis for arthritis detection in the trapeziometacarpal joint. In: MICCAI 2011, vol. 6892, pp. 360–367.

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40, 33–51.

Hammond, P., Hutton, Tim, J., Patton, Michael, A., Allanson, Judith, E., 2001. Delineation and visualisation of congenital abnormality using 3d facial images. In: IDAMAP at MedInfo 2001, pp. 26–29.

Hammond, P., Hutton, T.J., Allanson, J.E., Campbell, L.E., Hennekam, R.C.M., Holden, S., Patton, M.A., Shaw, A., Temple, I.K., Trotter, M., Murphy, K.C., Winter, R.M., 2004. 3D analysis of facial morphology. Am. J. Med. Genet. Part A 126A, 339–348.

Heimann, T., Meinzer, H.P., 2009. Statistical shape models for 3d medical image segmentation: a review. Med. Image Anal. 13, 543–563.

Hollander, M., Wolfe, D.A., 1999. Non-parametric Statistical Methods, second ed. Wiley-Interscience.

Horst, J.A., Beichel, I., 1997. A simple algorithm for efficient piecewise linear approximation of space curves. In: Proceedings of International Conference on Image Processing, vol. 2, pp. 744–747.

Hotelling, H., 1992. The Generalization of Student's Ratio. Springer Series in Statistics. Springer, New York, Chapter 4, pp. 54–65.

Jadhav, D.V., Holambe, R.S., 2009. Feature extraction using radon and wavelet transforms with application to face recognition. Neurocomputing 72, 1951–1959.

Loos, H.S., Wieczorek, D., Wurtz, R.P., Malsburg, C.v.d., Horsthemke, B., 2003. Computer-based recognition of dysmorphic faces. Eur. J. Hum. Genet. 11, 555–560.

Lötjönen, J., Kivistö, S., Koikkalainen, J., Smutek, D., Lauerma, K., 2004. Statistical shape model of atria, ventricles and epicardium from short- and long-axis mr images. Med. Image Anal. 8, 371–386.

Lucey, S., Wang, Y., Cox, M., Sridharan, S., Cohn, J.F., 2009. Efficient constrained local model fitting for non-rigid face alignment. Image Vis. Comput. 27, 1804–1813.

Ma, W.Y., Manjunath, B.S., 1995. A comparison of wavelet transform features for texture image annotation. In: Proceedings.of International Conference on Image Processing, vol. 2, pp. 256–259.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R., 1999. Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, pp. 41–48.

Milborrow, S., Nicolls, F., 2008. Locating facial features with an extended active shape model. In: Computer Vision – ECCV 2008, pp. 504–513.

Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. 29, 51–59.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns 24, 971–987.

Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., Malsburg, C., 1998. The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test. NATO ASI Series, vol. 163. Springer, Berlin, Heidelberg, Chapter 10, pp. 186–205.

Parker, S.E., Mai, C.T., Canfield, M.A., Rickard, R., Wang, Y., Meyer, R.E., Anderson, P., Mason, C.A., Collins, J.S., Kirby, R.S., Correa, A., 2010. Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. Birth Defects Res. Part A: Clin. Molec. Teratol. 88, 1008–1016.

Reyes, M., Gonzalez Ballester, M.A., Li, Z., Kozic, N., See, C., Summers, R.M., Linguraru, M.G., 2009. Anatomical variability of organs via principal factor analysis from the construction of an abdominal probabilistic atlas. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09, pp. 682–685.

Saragih, J.M., Lucey, S., Cohn, J.F., 2009. Face alignment through subspace constrained mean-shifts. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1034–1041.

Saraydemir, c., Taşpinar, N., Eroğul, O., Kayserili, H., Dinçkan, N., 2012. Down syndrome diagnosis based on Gabor wavelet transform. J. Med. Syst. 36, 3205–3213.

Seshadri, K., Savvides, M., 2009. Robust modified active shape model for automatic facial landmark annotation of frontal faces. In: IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS '09, pp. 1–8.

Sivakumar, S., Larkins, S., 2004. Accuracy of clinical diagnosis in down syndrome. Arch. Disease Childhood 89, 691–691.

Skotko, B.G., 2007. First- and second-trimester evaluation of risk for down syndrome. Obstetr. Gynecol., 110.

Skotko, B.G., Davidson, E.J., Weintraub, G.S., 2013. Contributions of a specialty clinic for children and adolescents with down syndrome. Am. J. Med. Genet. Part A 161, 430–437.

Srinivasan, M., Ravichandran, N., 2013. A new technique for face recognition using 2D-gabor wavelet transform with 2D-hidden markov model approach. In: 2013 International Conference on Signal Processing Image Processing & Pattern Recognition (ICSIPR), pp. 151–156.

Suinesiaputra, A., Frangi, A., Kaandorp, T., Lamb, H., Bax, J., Reiber, J., Lelieveldt, B., 2009. Automated detection of regional wall motion abnormalities based on a statistical model applied to multislice short-axis cardiac mr images. IEEE Trans. Med. Imag. 28, 595–607.

Suinesiaputra, A., Frangi, A.F., Kaandorp, T.A., Lamb, H.J., Bax, J.J., Reiber, J.H., Lelieveldt, B.P., 2011. Automated regional wall motion abnormality detection by combining rest and stress cardiac mri: correlation with contrast-enhanced mri. J. Magnet. Reson. Imag. 34, 270–278.

Swarts, K., 2009. Genetic Disorders. Greenhaven Press.

Upton, G., Cook, I., 1996. Understanding Statistics. Oxford University Press.

Uzümcü, M., Frangi, A., Sonka, M., Reiber, J., Lelieveldt, B., 2003a. ICA vs. PCA active appearance models: application to cardiac mr segmentation. In: Proc. MICCAI 2003, LNCS 2878.

Uzümcü, M., Frangi, R.F., Reiber, J.H.C., Lelieveldt, B.P.F., 2003b. Independent component analysis in statistical shape models. In: Proc. SPIE Medical Imaging, pp. 375–383.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, vol. 1, pp. I-511–I-518.

Wilamowska, K., Shapiro, L., Heike, C., 2009. Classification of 3d face shape in 22q11.2 deletion syndrome. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09, pp. 534–537.

Wiseman, F.K., Alford, K.A., Tybulewicz, V.L., Fisher, E.M., 2009. Down syndrome recent progress and future prospects. Human Molec. Genet. 18, R75–R83.

Yang, W., Lucey, S., Cohn, J.F., 2008. Enforcing convexity for improved alignment with constrained local models. In: CVPR 2008, pp. 1–8.

Zhao, Q., Okada, K., Rosenbaum, K., Zand, D., Sze, R., Summar, M., Linguraru, M., 2013a. Hierarchical constrained local model using ica and its application to down syndrome detection. In: MICCAI 2013, Lecture Notes in Computer Science, vol. 8150, pp. 222–229.

Zhao, Q., Rosenbaum, K., Okada, K., Zand, D., Sze, R., Summar, M., Linguraru, M.G., 2013b. Automated down syndrome detection using facial photographs. In: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Zhao, Q., Rosenbaum, K., Sze, R., Zand, D., Summar, M., Linguraru, M.G., 2013c. Down syndrome detection from facial photographs using machine learning techniques. In: SPIE Medical Imaging, pp. 867003–867003.

Zhou, M., Liang, L., Sun, J., Wang, Y., 2010. Aam based face tracking with temporal matching and face segmentation. In: CVPR2010, pp. 701–708.