Machine Learning Based Genetic Decision Making Methodology using Genotype-Phenotype Mapping

Kaya KURU[a], Yusuf TUNCA[b]

[a] IT Department, Gulhane Military Medical Academy (GATA), Turkey
[b] Department of Medical Genetics, Gulhane Military Medical Academy (GATA), Turkey

kkuru@gata.edu.tr

Abstract

Background: Specifying genotype-phenotype correlations correctly among many syndromes is labor intensive especially for very rare diseases. Faces of different persons share global characteristics while they are subject to considerable variabilities. Due to these variations, different sets of features which are most similar rather than the differences should be considered in evaluation to find out the phenotypes caused by similar deformation in genes in terms of comparing different persons throughout the similar anomaly classifications. **Objective:** In this study, we aim to present that accurate classification of dysmorphic faces through image processing techniques on two dimensional face images is feasible. **Methodology:** A methodology named as DSESPC (Dynamic Selection of Essential Similar Principal Components) is presented. This methodology evaluates the similarities while omitting the differences among features to accommodate for all possible similarities caused by genes. It has been tested on real data set collected from the dysmorphic facial images published in scholarly journals, thus accounting decent diagnostic information about the syndrome. **Results:** The methodology has been tested with 15 different syndromes that accommodate 5 examples per syndrome. A success rate of 79% which is achieved using all Principal Components (PC) in the previous study has been increased to 95% through dynamic evaluation of essential similar PCs, thus proving better image processing and machine learning based computer aided diagnostics for facial dysmorphology. **Conclusion:** It can be concluded that a great number of syndromes indicating a characteristic pattern of facial anomalies can be typically diagnosed by employing the approach we propose in this study.

Introduction

Dysmorphology is the aspect of clinical genetics concerned with syndrome diagnosis in patients who have a combination of congenital malformations and unusual facial features, often with delayed motor and cognitive development. Making a diagnosis for a dysmorphic patient requires a high degree of experience and expertise since many dysmorphic diseases are very rare.

There are specific properties, especially for facial dysmorphology caused by genetic syndromes and these properties are used by geneticists to pre-diagnose even before a clinical examination and genotyping are undertaken. For dysmorphic syndromes with known genetic causes, molecular and/or cytogenetic analysis is the appropriate route of investigation in order to confirm a diagnosis. However, applying right analysis method throughout many probable analyses is very much dependent to the accurate diagnosis considered before genotyping is undertaken.

Previously we had studied on automatic diagnosis of dysmorphologic syndromes. That study was dependent on comparing all PCs by which some satisfactory results were achieved, 79% (Tab.1, Fig.4). However, faces of different persons share global characteristics while they are subject to considerable variabilities. Due to these variations, a method measuring the similarities accurately rather than the differences between dysmorphic face images should therefore be implemented to accommodate for all these possible variabilities, especially for comparing the faces of different persons to diagnose the unknown cases by comparing them to those whose diagnoses are already known. We need to select the most similar and significant features in comparison while omitting some of the acquired dissimilar features since we aim to acquire similarities while comparing different persons. In this respect we developed a methodology named DSESPC (Dynamic Selection of Essential Similar Principal Components) to diagnose better in dysmorphology.

Methodology

This study proposes a novel and robust composite computer-assisted and cost-effective method by merging several methods in the characterization of the facial dysmorphic phenotype associated with genotype, in particular a method relying primary on face image capture

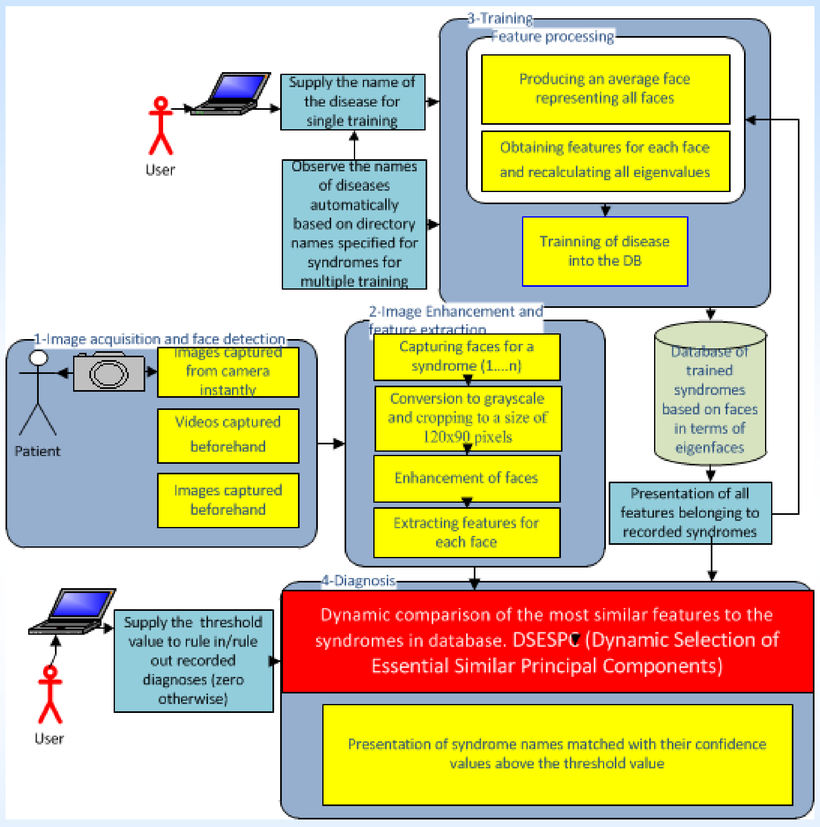

Fig. 1. Overall architecture of the methodology: the system consists of four main modules; face detection and image acquisition, image processing, training and diagnosis/recognition modules. These modules are divided into several sub modules that are delineated in the specified sections of the modules.

Face detection and image acquisition: patient images can be captured from several different environments. Image Processing: several image processing algorithms are employed to generate better frontal faces in order to observe better features. Training: the eigenfaces, eigenvalues and average image generated by PCA are recorded in DB with their labeled diagnosed names. Diagnosis/recognition: the trained classifiers are employed for prediction

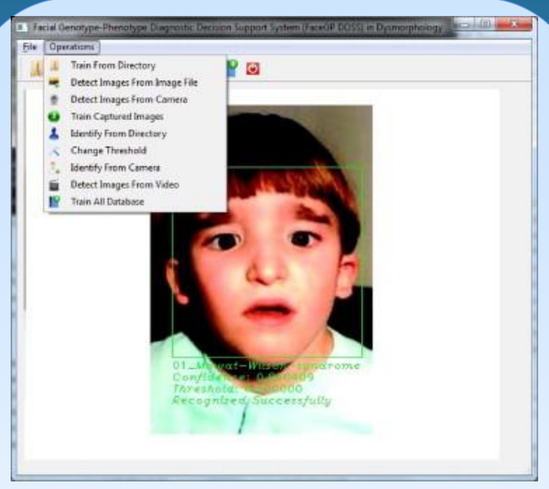

Fig. 2. Four messages are displayed; the name of the probable diagnosis (e.g. Treacher Collins syndrome), the degree of proximity to that diagnosis (e.g. 0.89), the threshold value entered by the user (e.g. 0.80), the message about whether a probable disease is found as *recognized successfully* or not as *unknown disease*. The messages change if other diseases are found above the threshold value to reveal them on the screen.

(acquisition from either camera, video or frontal face images) and manipulation to help medical professionals to diagnose syndromes efficiently (Fig.1, Fig.2). The methodology comprises several main modules, these main modules are divided into several sub modules as illustrated in Fig 1. The functions of main modules are summarized in Fig 1.

Our statistical methodology represents facial image data in terms of principal component analysis (PCA) and a leave-one-out evaluation scheme to train and quantify accuracy. The comparison is performed for each image trained in the database through Euclidean distance to find out all similar syndromes above the threshold value supplied by the user. The best matches above the threshold value are found for the syndromes that have minimum Euclidean distance. These syndromes are the probable diagnoses displayed to the user with confidence values. In this study different from our previous study, a new method named DSESPC is embedded into our overall methodology as depicted in Fig 1 highlighted in red. With the methodology, for each comparison, different sets of features (PC) which are most similar rather than the differences is considered in comparison to find out the phenotypes caused by similar deformation in genes.

Data set and design of the study: The application has been trained with selected 15 syndromes that accommodate 75 images as presented in Fig.3. The system could build a training set for these syndromes in less than a minute. Since the number of data samples available for training is low, a leave-one-out scheme in the training set has been applied to train and test the system for quantifying accuracy. In “leave one out cross validation” we hold one image out as a test image and train the system on all the remaining 74 images. We repeat this process 75 times, so that every data point gets a chance of being held out as a test sample. The application shows the confidence values in a table for a threshold value entered by the user. In the case study, we have aimed to find out the probable diagnoses with respect to rule-in I, II and III diagnoses respectively by adjusting the threshold value for each test image.

Evaluation of the Methodology

Fig. 3. Frontal faces for fifteen syndromes: 5 examples per syndrome, from upper left to the right and down respectively, Mowat Wilson, Goldenhar Treacher Collins, Williams Beuren, X-linked Mental retardation, Cardiofaciocutaneous, Cohen, Angelman, Craniofrontonasal, Crisponi, Laron, PMSE, Fragile X, Pitt Hopkins, Potocki-Lupski.

The total time required to search through 15 classes trained with syndromes to find the nearest syndromes for diagnoses with respect to the threshold value is 3 seconds. The user can adjust the threshold value to rule in/out diseases during the identification process. The greater the threshold value, the less number of probable diagnoses are proposed and vice versa; the less the threshold value, more probable diagnoses are revealed to the user together with their confidence values. Thus, the success rates of these rule-in observations were obtained. The diagnoses of these syndromes were made by applying appropriate genetic tests and were confirmed by the authors scientifically in their publications.

Rule-in I diagnosis corresponds to the proposition of the most probable diagnosis, rule-in II is the second most probable diagnosis in addition to the first one proposed by rule-in I, whereas rule-in III is the third most probable disease in addition to the other two proposed by rule-in I and II. DSESPC methodology has been tested on the images that weren`t diagnosed correctly using all PCs (Tab.1) to evaluate its success better.

Results

	Mowat Wilson					Goldenhar					Treacher Collins					Williams Beuren					X-linked Mental				
	1a	1b	1c	1d	1e	2a	2b	2c	2d	2e	3a	3b	3c	3d	3e	4a	4b	4c	4d	4e	5a	5b	5c	5d	5e
Rule in I	1e	12e	5e	6b	1a	2e	2c	2d	2c	2e	6b	3e	3b	3e	2d	4e	4e	4b	8b	4a	5b	5a	13e	14a	10a
II	15e	4a	1d	13e	15e	11b	12b	14e	14e	14e	9e	10a	5e	5e	9e	3b	3b	3e	13e	3b	9b	14e	1e	10e	3d
III	9e	13e	6b	1e	9e	10e	10a	11b	11b	10a	1d	5e	15a	10a	6b	1b	14e	14e	15d	6e	6b	6b	2e	2d	2d
	Cardiofaciocutaneous					Cohen					Angelman					Craniofrontonasal					Crisponi				
	6a	6b	6c	6d	6e	7a	7b	7c	7d	7e	8a	8b	8c	8d	8e	9a	9b	9c	9d	9e	10a	10b	10c	10d	10e
Rule in I	5a	6d	14a	6b	12d	10d	7d	10e	7b	15e	12e	8e	8b	10e	9b	9e	9e	6b	9b	8b	10e	10d	10d	15e	10a
II	9b	9e	6d	5a	5e	9e	8d	2e	8d	10b	8d	10e	1a	15b	15e	15e	5b	9b	6b	10e	2d	15e	8b	10b	2d
III	1d	5b	5b	9e	3e	8b	13a	11e	13a	7e	1a	4d	2b	9e	2d	10b	6b	15e	5b	7a	14a	9e	15e	2a	14a
	Laron					PMSE					Fragile X					Pitt Hopkins					Potocki Lupski				
	11a	11b	11c	11d	11e	12a	12b	12c	12d	12e	13a	13b	13c	13d	13e	14a	14b	14c	14d	14e	15a	15b	15c	15d	15e
Rule in I	11b	11c	11b	11b	11b	12e	12d	15a	12b	1b	13e	13e	13d	13e	5e	5d	5b	5b	14e	15a	15b	14a	10d	15e	11b
II	15e	15e	15e	2d	15e	13d	10a	5d	5e	8a	15d	12e	4d	12a	1d	10e	14a	2d	5b	4b	14a	15a	15e	1e	9e
III	10e	2d	9e	14e	2d	8a	2b	12d	6e	12a	6e	1b	8b	1d	6b	15b	6b	14a	2e	14a	5e	10e	1e	9e	15e

Tab. 1. Previous results using all PCA. Ruling in I, II and III diagnoses regarding the greatest values in comparison: the grey cells correspond to the right diagnosis.

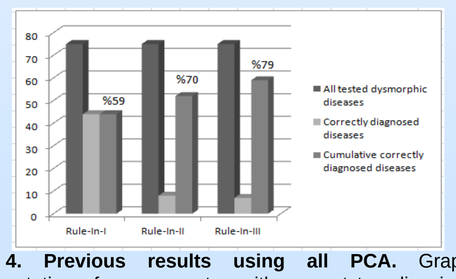

Fig. 4. Previous results using all PCA. Graphical representation of success rates with respect to ruling in one, two and three diagnoses concerning the values in Table 10.

	1b	3a	3e	4d	5c	5d	5e	6a	6e	7a	7c	8d	8e	9e	13e	14a
Rule in I	1a	3d	3a	12e	11d	4d	4c	13a	5e	13e	14d	7b	7a	13b	7b	7a
II	3a	15c	13c	7c	7a	1b	7c	4d	1a	11c	7e	13d	4e	9c	10d	9c
III	5e	12c	6e	4b	15d	5b	3d	15e	5c	7e	3a	8a	8c	10e	13c	14d

Tab. 2. Results applying DSESPC on the images not diagnosed correctly in Tab. 1 . Ruling in I, II and III diagnoses regarding the greatest values in comparison: the grey cells correspond to the right diagnosis. Te images of 5c, 5e, 6a and 6e are not diagnosed correctly.

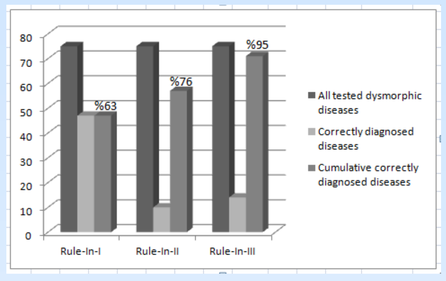

Fig. 5. Adding results obtained using DSESPC to Tab. 1 and Fig 4. Graphical representation of success rates with respect to ruling in one, two and three diagnoses concerning the values in Table 2.

The methodology has become succesful to diagnose 10 more cases among 16 patients (Tab.2, Fig. 5).

Discussion

The results show that our methodology is able to make a biometric identification among syndromes successfully and efficiently based on the features of the patients’ frontal faces, even though, the methodology has been tested by a limited number of 15 syndromes. Diagnosing syndromes correctly among many syndromes can be eased by the methodology provided that it is trained with those syndromes. The achieved success rate, 95%, is greater than our expectations. Thus, the methodology should be validated on other more data set to quantify the results.

Conclusion

The methodology may contribute to the medical professionals in several aspects. Some of these are:

- To support medical professionals who do not have expertise in the particular domain of dysmorphology such as general practitioners or pediatricians in rural areas,
- To support geneticists throughout thousands dysmorphic diseases, most of which are very rare and difficult to memorize and remember,
- To guide geneticists to employ correct cyto- and/or molecular genetic analysis that is the appropriate route of investigation in order to confirm a diagnosis with - known genetic causes by ruling in probable syndromes,
- No preprocessing of data manually that may cause the users to avoid the utilization of any system is required and all preprocessing is managed by the methodology.

Limits of the study

The results would be more reliable if better images were utilized and if the patients were in similar age group and sexes in the study. The more faces bearing the characteristics of any syndrome included in the training, the better the recognition of that syndrome will be.

Future work

The overall methodology should be tested on more cases and the methodology should be improved in terms of new findings

References

[1] Kuru, K., Tunca, Y. and Niranjan, M., 2013. A Novel Approach to Improve the Diagnostic Success of Computers in Dysmorphology: The DSESPC methodology and its Applications.

[2] Kuru K, Tunca Y, Niranjan M. Establishment of diagnostic decision support system (DDSS) in clinical diagnosis of genetic diseases: the facegp DDSS methodology and its applications. European Journal of Human Genetics. 2012 Jun 1;20(1):70..

[3] Kuru K, Tunca Y. Diagnostic Decision Support System in Dysmorphology. In Decision Support Systems 2012 Oct 17. IntechOpen.

[4] Kuru K, Girgin S, Arda K, Bozlar U, Akgün V. Developing diagnostic dsss based on a novel data collection methodology. In Knowledge Science, Engineering and Management: Third International Conference, KSEM 2009, Vienna, Austria, November 25-27, 2009. Proceedings 3 2009 (pp. 110-121). Springer Berlin Heidelberg.

[5] K. Kuru, M. Niranjan and Y. Tunca, "Establishment of a Diagnostic Decision Support System in Genetic Dysmorphology," 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 2012, pp. 164-169, doi: 10.1109/ICMLA.2012.234.

[6] Kuru, K., Girgin, S., Arda, K. and Bozlar, U., 2013. A novel report generation approach for medical applications: the SISDS methodology and its applications. International journal of medical informatics, 82(5), pp.435-447.