Diagnostic use of facial image analysis software in endocrine and genetic disorders: review, current results and future perspectives

R P Kosilek, R Frohner, R P Würtz, C M Berr, J Schopohl, M Reincke and H J Schneider
Medizinische Klinik und Poliklinik IV, Ludwig-Maximilians-Universität München, Ziemssenstraße 1, 80336 Munich, Germany and 1Institute for Neural Computation, Ruhr-Universität Bochum, Bochum, Germany

Abstract

Cushing’s syndrome (CS) and acromegaly are endocrine diseases that are currently diagnosed with a delay of several years from disease onset. Novel diagnostic approaches and increased awareness among physicians are needed. Face classification technology has recently been introduced as a promising diagnostic tool for CS and acromegaly in pilot studies. It has also been used to classify various genetic syndromes using regular facial photographs. The authors provide a basic explanation of the technology, review available literature regarding its use in a medical setting, and discuss possible future developments. The method the authors have employed in previous studies uses standardized frontal and profile facial photographs for classification. Image analysis is based on applying mathematical functions evaluating geometry and image texture to a grid of nodes semi-automatically placed on relevant facial structures, yielding a binary classification result. Ongoing research focuses on improving diagnostic algorithms of this method and bringing it closer to clinical use. Regarding future perspectives, the authors propose an online interface that facilitates submission of patient data for analysis and retrieval of results as a possible model for clinical application.

Introduction

Cushing’s syndrome (CS) and acromegaly are endocrine diseases that manifest with metabolic complications and typical changes to the appearance, especially characteristic changes to facial features that are particular to either disease. These pathognomonic signs can be detected by face classification software and could hypothetically be used for early diagnosis. Recent studies have found that CS is currently diagnosed with a latency of about 2–6 years (1, 2). This may partly be due to an overlap of clinical features with the more common metabolic syndrome. Acromegaly is still diagnosed with a delay of about 6 years, most likely due to the slow
progression of symptoms (2, 3). To reduce the diagnostic delay and improve disease outcomes, new sensitive and simple screening methods and increased awareness for these diseases are needed. Based on previous studies regarding the use of face detection and classification technology in a medical setting, first applications of this technology to the diagnosis of acromegaly and CS in pilot studies have shown promising results. This review is intended to give an overview of available data on this topic and discuss possible future developments.

Basic description of the facial image analysis and classification method

There is a growing body of literature available regarding the medical application of face detection and classification software. This chapter aims to provide a basic explanation of the image analysis and classification method as it was used by the authors of this review. It has been extensively described in previous publications from our group from 2011 and 2013 (4, 5). The principles of image analysis and classification as described are comparable within the studies presented in this review.

Study subjects are photographed using a regular digital camera. Frontal and profile photographs of the face are taken in a standardized way. Patients and control subjects are matched for gender and age. The following process involves three steps: i) detection of landmarks on the photographs; ii) extraction and analysis of information from the photographs; and iii) categorization of the photographs using a classification algorithm. The authors have used the software facial image diagnostic aid (FIDA) (6) for analyses. In this software, analysis is based on applying mathematical functions to a grid of nodes semi-automatically placed on relevant facial structures (landmarks). These functions evaluate geometry by assessing node distances and image texture surrounding the nodes by using Gabor wavelet transformations (feature vectors). An unknown subject is classified using a maximum likelihood classifier by assessing similarity with either group (control subjects vs patients) within the training database. Overall classification accuracy is calculated using the leave-one-out cross-validation method. Figures 1 and 2 show representative illustrations of study subjects and node placement.

Application in acromegaly

Several reports address the detection of acromegaly using image processing methods. In 2006, Learned-Miller et al. (7) reported on using 3D models derived from facial photographs for the detection of acromegaly. In this study, a 3D morphable model of the subject’s head was derived from a frontal facial photograph and processed using support vector machines (SVMs). Forty-nine study...
subjects (24 patients, 25 controls) were included. The total classification accuracy was 85.7%, with 17 of 24 patients (71%) being correctly classified. The authors noted that the study sample was relatively small and also diverse in the patient group in terms of gender and race, while the control subjects were all white males. From our point of view, this can lead to a potential bias in classification analyses. In a follow-up publication from 2011, the accuracy of this computer system was compared to the diagnostic accuracy of ten generalist physicians, resulting in 86% vs 26% accuracy, respectively. Additionally, the authors sorted a database of 200 normal subjects according to the presence of features of acromegaly using the same method (8).

Based on previous research, the authors of this review conducted a study regarding the detection of acromegaly and published the results in 2011 (4). In this study, a method consisting of Gabor wavelet transformations for feature extraction, geometric analysis, and a maximum likelihood classifier was used. The overall correct classification rate was 81% in a sample of 57 patients (29 females, 28 males) and 59 control subjects (29 females, 30 males) matched for gender and age, with 71.9% of patients (sensitivity) and 91.5% of control subjects (specificity) correctly classified. Additionally, the patients were categorized into subgroups by disease severity according to facial features of acromegaly. We also compared the accuracy of the software tool to the diagnostic accuracy of physicians assessing the images. Detailed results are shown in Table 1. It is of note that software classification accuracy exceeded the diagnostic performance of physicians assessing the same images, especially in the category of patients with mild features of acromegaly. In a further step the procedure was modified by reducing the number of nodes used for analyses and optimizing their placement. This approach was then validated within the original data set and a new database of subjects, showing a moderate improvement of classification rates (9).

Additionally, Gencturk et al. (10) achieved an overall classification accuracy of >90% using local binary patterns and a Manhattan classifier for the detection of acromegaly on facial photographs.

**Application in CS and current research**

The authors of this review also applied the method to the detection of CS in a proof-of-concept study and published the results in 2013 (5). For this, the experiments were repeated in the original setup with a cohort of 20 female patients suffering from endogenous or iatrogenic CS and 40 age- and gender-matched control subjects. The overall classification accuracy was 91.7%, with 85% of patients (sensitivity) and 95% of control subjects (specificity) correctly classified. The patient cohort was categorized by etiology of CS (Cushing's disease, n=8; adrenal CS, n=4; iatrogenic CS, n=8). Only patients with Cushing's disease were not correctly classified by the software; this is a finding of unclear significance. Since this was a proof-of-concept study, there were three major limitations (small, all-female study cohort, no matching of patients and controls for BMI) that need to be addressed in the future.

Our current studies aim to build on previous results described above by addressing limitations and implementing other measures. The recruitment targets for the next assessment of classification accuracy for CS are set at 50 patients and 100 control subjects matched for age, gender, and BMI. Additionally, the placement of nodes used for classification needs to be slightly modified to better match...

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Software&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Experts&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Correct classification rates (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>81.9</td>
<td>72.1</td>
</tr>
<tr>
<td>Acromegaly</td>
<td>71.9</td>
<td>63.2</td>
</tr>
<tr>
<td>Controls</td>
<td>91.5</td>
<td>80.8</td>
</tr>
<tr>
<td>By sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acromegalic women</td>
<td>62.1</td>
<td>48.3</td>
</tr>
<tr>
<td>Control women</td>
<td>93.1</td>
<td>82.8</td>
</tr>
<tr>
<td>Acromegalic men</td>
<td>82.1</td>
<td>78.6</td>
</tr>
<tr>
<td>Control men</td>
<td>90.0</td>
<td>76.7</td>
</tr>
<tr>
<td>By severity of facial features in acromegaly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>58.3</td>
<td>38.9</td>
</tr>
<tr>
<td>Intermediate</td>
<td>77.3</td>
<td>74.2</td>
</tr>
<tr>
<td>Severe</td>
<td>90.9</td>
<td>93.9</td>
</tr>
</tbody>
</table>

<sup>a</sup>Classification by software, using a combination of the Gabor jet-based function P (normed scalar product including phase) and geometry-based function L (edge difference length) on frontal and side views of patients.

<sup>b</sup>Means of visual classification by three medical experts with extensive experience in acromegaly.

<sup>c</sup>Means of visual classification by three general internists not particularly experienced with acromegaly.
facial features of CS. Clinical and biochemical parameters will be recorded alongside the software classification result with the goal of establishing a combined clinical prediction score.

The increasing prevalence of obesity and its sequelae (metabolic syndrome) continues to pose a diagnostic challenge in clinical practice. The problem has been extensively discussed in recent literature, such as the fact that the diagnostic accuracy of various biochemical tests for hypercortisolism drops significantly when applied to a sample of obese subjects (11, 12). In our experience so far, this problem also applies to face classification technology. Determining and improving diagnostic accuracy within a BMI-matched sample is a distinct goal of current research.

Application of face classification technology in genetic disorders

In a series of publications from 2003 to 2011, a group of authors from Germany reported on applying face classification technology to the detection of multiple genetic syndromes. In studies published in 2003 and 2006, Loos et al., (13) and Boehringer et al. (14) reported on performing face classification using only texture analysis via Gabor wavelet transformations and various classifiers. The data sets consisted initially of 55 patients suffering from five genetic syndromes, and later of 147 patients suffering from ten genetic syndromes, achieving an overall classification accuracy of 76 and 75.7%, respectively (13, 14). In a study published in 2008, Vollmar et al. (15) included side-view photographs and geometric information in the face classification analyses, resulting in improved accuracy in an even larger data set (14 genetic syndromes). In 2011, Boehringer et al. (16) applied the previously described method in clinical practice by using a classifier trained on an existing data set to classify a new set of patients with pictures taken in less standardized conditions. While classification accuracy was good in training conditions, this study showed an unfavorable result of 21% classification accuracy. The authors attributed this to large phenotypic differences in the groups of test subjects and suggested broadening the data set used for training to improve accuracy in clinical practice (16).

Apart from these reports, several studies addressed the detection of Down syndrome using facial photographs. A research group from Turkey achieved a maximum classification accuracy of 97.34% when analyzing a set of 30 facial images (15 Down syndrome, 15 healthy subjects) using Gabor wavelet transformations and a SVM classifier (17). A second study from Turkey published data on the use of an image-processing method based on local binary patterns for feature extraction and changed Manhattan distance for classification, with an overall correct classification rate of >90% in a sample of 107 facial images collected from the internet (51 subjects with Down syndrome, 56 healthy individuals) (18). Zhao et al. (19) reported that using Independent Component Analysis for landmark placement, local binary patterns for feature extraction and SVMs for classification, they achieved 96.7% accuracy in the classification of a data set consisting of 130 facial photographs (50 patients with Down syndrome, 80 healthy subjects).

Additionally, Douglas and Mutsvangwa (20) reviewed the use of various image analysis methods for the identification of facial features associated with fetal alcohol syndrome.

Conclusions and future perspectives

The use of face detection and classification technology in a medical setting appears to be a promising and emerging area of research. A growing number of publications regarding the use of this technology as a diagnostic tool have shown good results within certain boundaries, most importantly small study cohorts.

Independent of the particular method of image analysis employed, experimental clinical use of such systems requires a large training database to account for variations in the photographs taken by external physicians. This has also been noted by other authors as pointed out above. Until now, most methods require manual processing of data to a certain extent. This is another important limiting factor. We think that optimized and finally automatic selection of landmarks will also facilitate clinical use.

We estimate that a database consisting of 500–1000 subjects should in theory be sufficient for experimental clinical application as a screening tool. Possible targets for screening for acromegaly are patient populations with a higher pretest probability, such as patients presenting with dysgnathia at specialized centers. Regarding CS, similar applications are possible once recruitment targets have been reached. For both diseases, integrating results from software face classification with clinical results could further increase diagnostic accuracy.

Regarding future perspectives, we do not expect that this technology will ever replace biochemical testing in the diagnosis of acromegaly or Cushing’s syndrome. Rather, it can be helpful as a screening tool in selecting
patients that need additional biochemical testing. Hypothetically, it could also be used for follow-up exams by comparing an individual patient’s photographs taken at various points in the course of the disease either directly with each other or with a database that does not include any of the images in question, with the results being an indicator of disease severity and potential progression or relapse. However, there are no studies on this particular application to date. Additionally, we are currently developing a fully functional online interface as a model for possible clinical applications. This interface is designed to provide participating physicians with the option of submitting patient photographs along with additional information such as clinical and laboratory findings for analysis and retrieving results. Beyond this, we are also developing a fully automatic live classification system.

Independent of developments in diagnostic methods, increased awareness, especially among physicians in frequent contact with possible CS or acromegaly patients, is essential for reducing the diagnostic delay mentioned in the introduction. Information programs might be helpful in achieving this goal.

Declaration of interest
R P Kosilek, R Frohner, R P Wurtz, and C M Berr report no conflict of interest. J Schopohl received research grants, speaker fees, and travel grants from Ely Lilly, Ipsen, Novartis, and Pfizer. M Reincke received research grants and speaker fees from Novartis and speaker fees from Pfizer and Ipsen. H J Schneider received research grants, speaker fees, and travel grants from Pfizer and speaker fees and travel grants from Novartis.

Funding
The research has been partially funded by a German Research Foundation (DFG) grant WU 314/6-2 to R P Wurtz. It has also been partially funded by a research grant from Pfizer to H J Schneider.

Acknowledgements
We thank Sandra Rutz for help in organization and patient recruitment. We also thank Gudrun Hackenberg and Kathrin H Popp for help in patient recruitment.

This paper forms part of a special issue of European Journal of Endocrinology on Cushing’s syndrome. This article is adapted from work presented at the IMPROCUISH-1: Improving Outcome of Cushing’s Syndrome symposium, 12–14 October 2014. The meeting was supported by the European Science Foundation, Deutsche Forschungsgemeinschaft, Carl Friedrich von Siemens Stiftung, European Neuroendocrine Association and the Deutsche Gesellschaft für Endokrinologie. The opinions or views expressed in this special issue are those of the authors, and do not necessarily reflect the opinions or recommendations of the European Science Foundation, Deutsche Forschungsgemeinschaft, Carl Friedrich von Siemens Stiftung, European Neuroendocrine Association and the Deutsche Gesellschaft für Endokrinologie.

References


