Proc. of 6th German AAL Conference (AAL 2013), Berlin, Germany, pp. 186-195, VDE Verlag

Mimikdysfunktionen: Konzeption eines therapiebegleitenden Trainingssystems

Facial movement dysfunctions: Conceptual design of a therapy-accompanying training system

Cornelia Lanz, Ilmenau University of Technology, Neuroinformatics and Cognitive Robotics Lab, cornelia.lanz@tu-ilmenau.de

Joachim Denzler, Friedrich Schiller University Jena, Computer Vision Group, joachim.denzler@uni-jena.de

Horst-Michael Gross, Ilmenau University of Technology, Neuroinformatics and Cognitive Robotics Lab, horst-michael.gross@tu-ilmenau.de

Kurzfassung

In diesem Beitrag wird das Anwendungsszenario eines kamerabasierten, automatisierten Trainingssystems für Patienten/-innen mit Mimikdysfunktionen vorgestellt. Der Einsatz des Systems ist für eine häusliche Umgebung vorgesehen und soll therapiebegleitend erfolgen. Ziel der geplanten Anwendung ist es, die neben der Therapie notwendigen, nicht durch Logopäden/-innen überwachten, Trainingseinheiten zu begleiten und Fehler in der Übungsdurchführung zu verhindern. Ausgehend von einem Erfahrungsaustausch mit Logopäden/-innen und der Analyse existierender Anwendungen wird ein theoretisches Modell erstellt, welches sich als Basis für die Konzeption einer solchen Applikation eignet. Ein weiterer Schwerpunkt des Beitrags ist die Implementierung, mit besonderem Augenmerk auf der Automatisierung der Abläufe. Wir motivieren die Auswahl unserer Merkmale für die automatisierte Analyse von Gesichtsausdrücken und untersuchen sie sowohl hinsichtlich ihrer Unterscheidungsfähigkeit in Bezug auf die therapeutischen Gesichtsübungen als auch hinsichtlich ihrer Robustheit, die für eine Anwendung in der Praxis unverzichtbar ist.

Abstract

In this work, we present the scenario of a camera-based training system for patients with dysfunctions of facial muscles. The system is to be deployed accompanying to therapy in a home environment. The aim of the intended application is to support the unsupervised training sessions and to provide feedback. Based on conversations with speech-language therapists and the analysis of existing solutions, we derived a theoretic model that facilitates the conceptual design of such an application. Furthermore, the work is concerned with implementation details, with main focus on the automatisation of the face analysis. We motivate the selection of the features and examine their discriminative power and robustness for the automated recognition of therapeutic facial expressions in a real-world application.

1 Introduction

Facial expressions play an important role in interpersonal communication. Diseases like Parkinson, stroke, or mechanical injury of the facial nerve can lead to a dysfunction of facial muscle movements. The resulting problems are manifold. One consequence of this is that the structure of daily life needs to be adapted to the health impairments. For example, food intake affords more time, if eating and swallowing difficulties exist. Patients with impaired eyelid closure need to wear a bandage at night to protect their cornea and the loss of eyelid blink can contribute to drying of the eye. In the long term, this leads to damage of the cornea and may result in blindness. Furthermore, leisure activities like swimming have to be stopped because of the poor corneal protection [1].

Besides implications on daily life and physical abilities, facial muscle dysfunctions can also have negative effects on mental health. Lack of appropriate facial expressions may lead to misunderstandings in face-to-face communication. In combination with impaired appearance of the face caused by imbalance of the facial muscles, low self-confidence and social isolation may be the consequences.

In addition to medicinal treatment, the regular practice of therapeutic face exercises under supervision of a speech-language therapist is an important part of rehabilitation.

Due to the need for a high practicing frequency, patients need to conduct unsupervised exercises at home – accompanying to therapy. A view in the mirror supports the self-supervised training (Figure 1).

However, the incorrect execution of exercises can impede the training success or even lead to further impairment [2].

The development of technical assistance systems aims to overcome these problems. Such systems can be realized in various forms, e.g., as pure software applications running on a notebook, or as a multifunctional robotic assistance platform. The latter can additionally comprise reminder, communication and training functionalities. Training functionalities aim at improving the patients cognitive [3] and physical [4] state. A therapy-accompanying training system for facial exercises would complete the recent developments of such systems.

Against this background, we aim at the development of an automated, therapy-accompanying training system for patients with facial muscle dysfunctions.

In this publication, we give an overview of the status of our work with respect to design- and implementation-related tasks. We present a theoretic model, which supports the conceptual design of a training system that is suited to the needs of the target user group. The theoretic model is appropriate for the design of a variety of systems for cognitive and physical stimulation. However, in this paper, we concentrate on the topic of facial exercises.

Further emphasis is put on the automation of the training session monitoring. In this context, we motivate the application of the depth features, which we have selected for the specified task. To enable a better understanding for the practical side of this application scenario, we additionally present and examine the features' suitability for a real-world scenario by evaluating their discriminative power and their robustness. A more detailed description and evaluation of the features is given in [5].

The images that are necessary for the analysis of the training sessions are captured using the Kinect[1] from Microsoft. Although there are other methods that are suited for the recording of depth information with higher resolution, we decided for the Kinect because of its moderate price and widespread availability.

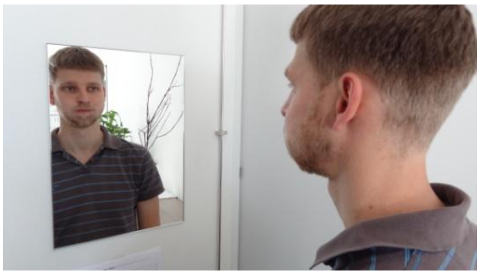

Figure 1 Patients regularly have to conduct unsupervised facial exercises at home in front of a mirror.

2 Existing practical solutions

In this section, we give an overview of therapy-accompanying solutions that are already employed for the rehabilitation of facial muscle dysfunctions. We discuss these solutions to identify the main functionalities, which are needed for the design of a comprehensive and automated training system. However, the use of media-technology is slowly evolving in this field. Conventionally, the therapist selects a set of exercises and hands out printed drawings or images as an instruction manual and reminder. The software *PhysioTools*[2] was developed in order to facilitate and streamline this process. It includes a database of various exercises for physical therapy and enables the therapist to compile a set of exercises for a training session. Furthermore, it arranges the images and their associated text in a printer friendly layout. Therapists do not need to search or create descriptive images and to write instructions on their own.

However, a video can even be more descriptive because it depicts the process of the exercise execution, instead of the final state only. The software *LogoVid*[3] comprises demonstrative videos of various exercises that are supplemented by oral instructions.

Both mentioned solutions mainly fulfill a tutorial function. The software *CoMuZu*[4] is supplemented by documentation and feedback functionalities. The target audience are teenagers. As a result, the whole user interface and the story is rather playful in order to give a motivating add-on. The therapist is able to unlock required exercises and in this way design an individual exercise schedule. Instructions for exercise execution are provided in videos. After each training session the teenager is advised to keep a diary about the training with respect to its success and difficulties. Afterwards, the diary can be reviewed by the therapist in order to get an impression of the training performance. However, it is rather impractical and questionable that the patient has to do the evaluation on his own. The three examples show, that current solutions lack an objective and sophisticated feedback function, because the patients have to perform the unsupervised exercises in front of a mirror and evaluate their correctness for themselves. This involves several difficulties. Experience and knowledge of the patient with respect to exercise evaluation may be insufficient, and especially children depend on the support of their parents.

In addition, the patient has to concentrate on the execution and evaluation of the exercises simultaneously, which can be very demanding. As a result the patient may lack attention with respect to important details of the exercise.

There are studies that indicate that incorrect execution of exercises may lead to an impairment of the facial muscle capabilities [6]. This impairment comprises synkineses that are caused by compensatory motions. Synkineses are involuntary facial movements that accompany voluntary facial movements. For example, a patient may tend to close the eyes to perform an exercise that is physically demanding, e.g., the stretching of the mouth. After a while, this leads to miswiring of nerves and both movements will be invo-

[1] http://www.xbox.com/en-US/kinect

[2] http://www.theorg.de

[3] http://www.logomedien.de/html/logovid7a.html

[4] http://www.comuzu.de/

luntarily connected. Further compensatory motions are the raise of the chin, if the patients have to touch the nose with their tongue.

A third difficulty is the lack of objective documentation. The evaluation that is made by the patient may be disproportionally optimistic or pessimistic, depending on the current mood. Every person is susceptible to “non-objectiveness”, even a therapist. However, patients, who are directly affected by success or failure, might even be more biased in their evaluation because of their mood.

More objective feedback is given by biofeedback approaches that employ electromyography to measure the electrical activity of the muscles during practice. This enables the detection of subtle muscle movements that are not visible to the eye. However, the method is more common in earlier states of facial muscle dysfunction, when no movements are visible, and has limited suitability for use in a home environment [1].

Besides the documentation of single practicing sessions it would be helpful to have a solution that enables long-term documentation. The therapist could browse through the exercising history and may identify processes of improvement or impairment, which developed slowly over a larger time span.

Other solutions focus on a more playful aspect. The game *Mimik Memo*[5] is designed for children between three and eight years. It can be played by two to six children. The game consists of cards that show drawings of animals, which perform facial exercises (e.g., tongue touches the tip of the nose). The task to mimic the exercises is embedded in a game scenario. Concerning the therapy of children with facial dysfunctions, a game scenario adds an important motivational component.

Summarizing the above yields four main functionalities which constitute an assistant and comprehensive training system. These functionalities refer to tutorial, feedback, documentation, and motivational aspects that are able to support and enrich exercising. In the next section, we will discuss these aspects in more detail. Furthermore, we will derive a schematic model that is suited to support future developments of such training systems.

3 A schematic model or: What is lacking in practical solutions?

The four aforementioned functionalities roughly coincide with the specifications that we have determined in discussions with speech-language therapists. In the following, we give a detailed description of each functionality and construct a schematic model as a basis for the design and implementation of an automated training system (Figure 2). The model is suited for various systems of cognitive and physical stimulation, however, we focus on facial exercises. The schematic model facilitates the conceptual work by enabling the identification of beneficial subfunctionalities (outer area of the illustration). The determination of the subfunctionalities is based on the analysis of the presented solutions and the discussions with speech–language therapists.

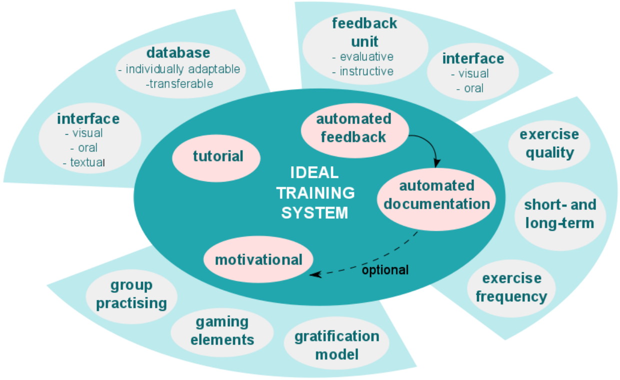

Figure 2 Schematic model for the conceptual design of an automated training system.

The design – or exclusion - of each subfunctionality depends on the needs of the target users and the intended price and complexity of the system. The schematic model represents an ideal system. ‘Ideal’ refers to the inclusion of a comprehensive range of subfunctionalities – a larger range than a real-world application in general may need.

The *tutorial functionality* consists of two elements: a database and an interface. The database provides a collection of therapeutic face exercises, which can be activated by the therapist for each of the patients individually. This allows for the creation of individual training schedules that can be adapted according to the success or failure of preceding training sessions. For each exercise, there is instruction material in form of videos including oral explanations. Important background knowledge can be documented in textual form as well.

The interface element of the tutorial functionality visualizes and verbalizes the instructions for the patient. It is important to keep the target users in mind, when designing this interface. While an adult patient may get along with a rather simple video and some textual instructions, a child needs more playful and vivid instructions to keep its attention. Furthermore, some patients may be impaired by additional disease patterns. As mentioned in the introductory section, possible causes of facial dysfunctions are brain lesions, generated by a stroke. Besides decreased physical abilities, brain lesions can also result in cognitive impairments. One example is the language ability impairment aphasia, which is characterized by difficulties with respect to reading, writing, speech production and speech processing. For persons with decreased speech processing abilities a high amount of visual instructions is essential in order to understand the correct exercise execution.

Similar to the tutorial functionality, the *feedback functionality* consists of two elements as well: a feedback unit and an interface. The feedback unit automatically generates information about mistakes and imprecisions in the exer-

[5] http://www.haba.de/

cise execution. Thereby, the mirror is replaced by a camera and the video of the patient doing the exercise is shown on the screen. As mentioned in the introductory section, exercise evaluation by the patient may be affected by the lack of experience, the emotional state, and the disability to concentrate on the execution and the evaluation at the same time. An automated feedback, however, guarantees results that are more objective and reproducible. The feedback unit provides two sorts of feedback: evaluative and instructive feedback.

Evaluative feedback gives a rating of the exercise execution. This rating can vary from a binary rating (good/bad) to a refined scale (0-100% similarity to the ideal exercise). Additionally, it is possible to realize such a rating for different areas of the face, e.g., mouth or cheeks, separately. The challenge is to find a suitable measure for the assessment of exercise quality. Instructive feedback comprises advice on inaccuracies during practice and gives concrete feedback for improvement ("Puff your left cheek stronger."). Therefore, it is more similar to a real therapist than the evaluative feedback.

The interface of the feedback functionality conveys the feedback information in an oral or visual form to the patient. A textual form is less feasible, because patients would have to watch the text and the video of their face simultaneously. Besides the output of evaluative and instructive information, the interface can be used to provide an avatar that synthesizes the face of the patient. The objective of this is twofold: first, an avatar would add a motivational aspect for children, e.g., by enabling children to slip into the role of their favorite comic character. Secondly, a neutral avatar helps patients who are emotionally affected by the impaired appearance of their face and who avoid looking in the mirror. This property of the feedback interface is closely related to the motivational functionality. Detailed information about the conceptual design of the feedback unit will be given in the following section.

The ideal training system additionally comprises a *documentation functionality*. This functionality is fully automated and focuses on the exercise quality and the exercise frequency. The exercise frequency can be logged to establish a schedule, in which every day of practice is registered, supplemented by the exercise duration. The exercise quality unit comprises the documentation of the exercise success or failure. Retrospectively, the therapist can see which exercises have been performed incorrectly or which have been less difficult for the patient. The automation of the documentation process allows the patient to fully concentrate on the exercise execution during practice. Additionally, no manipulation of the documentation with respect to exercise quality or frequency would be possible. The functionality can be used for short-term documentation, which may comprise information about one single training session or about long-term documentation, which would capture the process over several weeks or even months. The unit that documents the exercise quality needs input from the feedback functionality. Thus, to have a consistent documentation, it is important that the evaluation tool gives objective and reproducible results.

The *motivational functionality* contains elements that motivate the patients to do the practicing sessions with a regular frequency and with certain accuracy. The design of the motivational functionality depends on the target audience. Although one may assume that the inclusion of gaming elements is mainly beneficial for the motivation of children, studies showed a positive impact on the motivation of adults as well, when, e.g., using Wii sports[6] ([7], [8]). Furthermore, the integration of the documented training success, e.g. in form of high-score lists, may motivate the patient to practice with a higher frequency in order to exceed earlier performances. Additionally, some extra functionalities may be unlocked, if a patient achieves a further level, which may also enlarge the motivation. Group work may also be more motivating, e.g., as intended with the *Mimik Memo* game. However, in case of an application that is planned to be highly adaptable to the needs of an individual patient, practicing in groups may enlarge the complexity of system development.

Summarizing the above, we think that the feedback functionality plays an essential role because it contributes to the construction of a consistent documentation and the documented success, on the other hand, can be integrated into the motivational functionality. Therefore, in the following sections, we focus on the embedding of the feedback unit into the training system and discuss and evaluate the automation of the feedback process.

4 Conceptual design and details of the automatic training system

In this section, we focus on the conceptual and implementation-related aspects of the training system. First, we provide a schematic overview of the process steps comprised by the system. Additionally, we describe the collection of test images, the selection of features and the choice of the camera type. Finally, we present our preliminary results and status on the way to the solution of this extensive task.

4.1 Overview of the training system

In the following, we examine the embedding of the *feedback unit* in the process of automated feedback generation. As shown in Figure 3, the training system is divided into three layers: the *human actions*, the *interface* and the *algorithm*.

The layer on the top comprises the actions of the patient. Via an input interface, such as a camera, the algorithmic layer receives an image or a video of these actions.

The algorithmic layer is the basis of the automated feedback and consists of two units: the *automated face analysis unit* and the *feedback unit*. The task of these units is to analyze the appearance of the face in order to derive informa-

[6] http://www.nintendo.co.uk/

tion about the training performance. The properties of the face are captured by the extraction of descriptive features from facial regions. As a result, in each image of the data stream, the face has to be localized and distinctive facial points (e.g., the nose tip) and regions (e.g., the cheeks) have to be detected. The extracted features are analyzed automatically in order to generate evaluative and instructive feedback. The feedback is forwarded to the output interface, e.g., a display or a speech synthesis (or both). The instructive feedback comprises information about necessary changes in exercise execution and, therefore, directly affects the actions of the patient. Evaluative feedback only comprises an assessment of the exercising quality. However, we assume that a negative evaluation of the training will also affect the actions of the patient.

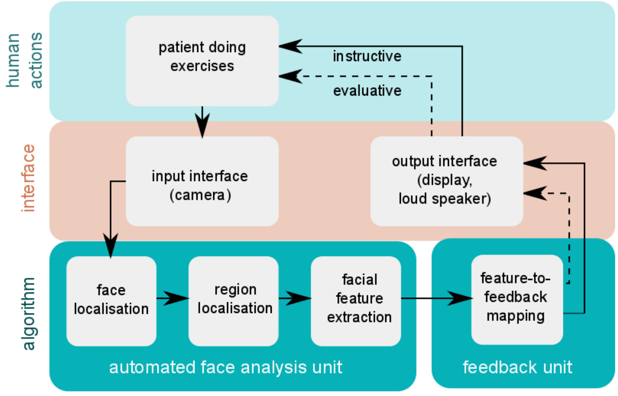

Figure 3 Embedding of the feedback unit in the process of automated feedback generation.

To be more precise, we can say, that the described scenario does not involve a face-appearance-to-feedback mapping but rather a feature-to-feedback mapping. However, features only describe a part of the face properties. Thus, an important question for the selection of the features is, whether they are suited to represent the properties of the face and the quality of the exercise. If the feedback that is given by the training system does not correlate to the feedback of a therapist, then there are two main possibilities: the mapping of the describing features to the feedback is incorrect or the features are not suited to represent the appearance of the face. In order to reduce the probability for the latter, the features need to be examined more closely (left image of Figure 4). The first question is, whether the features are suited to separate the different exercises. If the features are not able to capture the characteristics that distinguish the different exercises then it is unlikely that the features are able to describe the more detailed differences that are necessary to characterize a correct or an incorrect exercise execution. The second question is, in how much detail the features are able to describe different states of an exercise: How do the feature values change if a face expression changes from a neutral state to the final state of the exercise? The third question refers to the robustness of the features. In a real-world application it is not feasible to localize the position of the points and regions for feature extraction manually. As a result, they have to be detected automatically. However, automated labeling is less accurate than manual positioning. Thus, we need to evaluate the robustness of the features with respect to varying regions of feature extraction. The performance of the features – extracted from manually labeled points and regions – must be compared to the performance of features extracted from automatically detected areas. The right image of Figure 4 shows the 58 manually labeled landmarks used in our approach. For the automated labeling of these landmarks, we train an Active Appearance Model (AAM) ([9], [10]). AAMs have various applications in the area of object detection. Commonly, they are employed for the classification of facial expressions. In this work, however, we apply them for finding and placing the landmarks. The nose tip is detected robustly by a threshold-based localization algorithm using curvature analysis [11]. This approach is more accurate for the nose tip detection than the solution found by the AAMs, however, it is not suited for landmarks that lie in areas with less characteristic and changing surface shape, as for example the corners of the mouth or points on the cheek.

In the following, we motivate the selection of depth features as robust descriptors of the landmarks and evaluate their discriminative power with respect to the distinction between different exercises. Furthermore, we compare the results for manually and automatically labeled regions. Prior to that, we will have a closer look on the exercises to be included into the training system that is currently developed. These exercises are the basis to define, which regions of the face need to be localized and analyzed more closely.

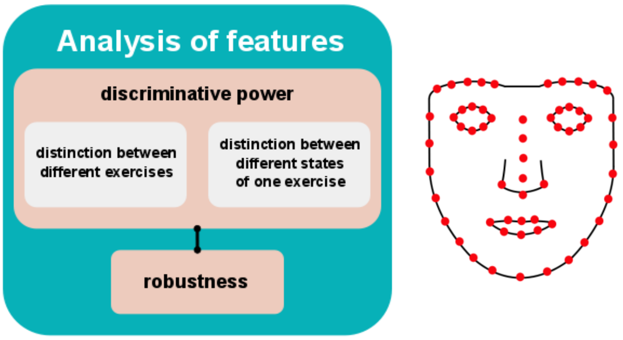
Analysis of features
discriminative power
distinction between different exercises
distinction between different states of one exercise
robustness

Figure 4 Left image: Analysis of features comprises the evaluation of the discriminative power and the robustness. The discriminative power consists of the features abilities to discriminate between the different exercises and the different states of one exercise. Right image: 58 manually labeled landmarks.

4.2 Therapeutic face exercises

In cooperation with speech therapists, we selected a set of nine therapeutic face exercises by certain criteria (Figure 5).

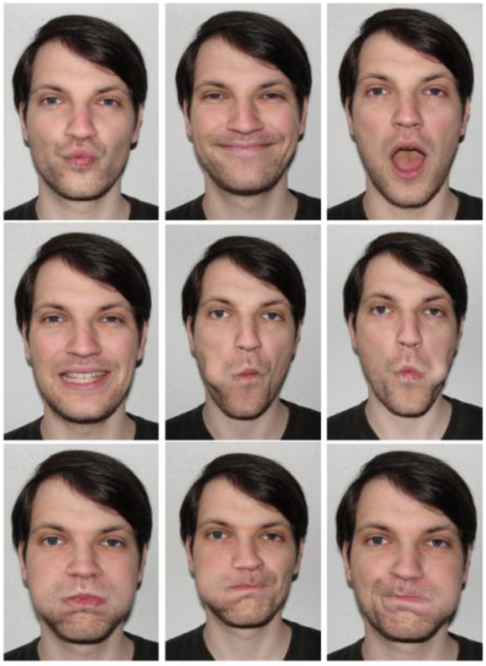

Figure 5 Exercises that have been selected in cooperation with speech therapists (from left to right and top to bottom): *pursed lips*, *taut lips*, *A-shape*, *I-shape*, *cheek poking* (right/left side), *cheeks puffed* (both/right/left side(s)). Exercises are performed by a person without facial movement dysfunctions for better visualization.

The first criterion was the ability to transfer the exercises to various disease patterns because speech-language therapy is geared towards people with various facial movement dysfunctions. Facial palsy for example comprises a reduced ability as well as the total inability to move facial muscles [12]. It can be caused by brain lesions or mechanical injury of the facial nerve. Another result of brain lesions can be dysarthria, which results in speech disorders and articulation problems. A further disease pattern is the myofunctional disorder, which is caused by an imbalance of facial muscle strengths, and often affects children [12]. Typical symptoms are a constantly opened mouth and an incorrect swallowing pattern. The exercises that we selected are beneficial for each of these disease patterns. Additionally the exercises should train several face regions: the lips, the cheeks and the tongue.

Each exercise has to be retained for around two or three seconds. The speed of the performance is not important.

Therapeutic tools like spoons and spatulas, as well as movements of the head, e.g., moving the chin to the chest, should be avoided in order to prevent occlusions. Occlusions lead to missing information, which would necessitate more cameras for observing the patient. However, we constrain the number of cameras to a frontal one to reduce hardware costs, which is important in order to guarantee widespread use of such a system. Additionally, the complexity of camera calibration is reduced.

The selected exercises are easy to practice and build a set of sub-exercises that can be combined to more complex and dynamic series of exercises: As an example, the alteration between pursed and taut lips or pursed lips and a neutral face are possible.

Due to the lack of a public database that comprises facial exercises, we collected our own dataset which will be made available as soon as possible. It contains eleven persons conducting the nine exercises. For each exercise, there are around seven images showing different states throughout exercise execution. This amounts in a total size of 696 images in the dataset.

For the following tests, we only employ image data that show healthy persons doing the exercises. Because our main focus in this paper is the selection and evaluation of the features, we want to eliminate other sources of error. Thus, we omit data recorded from persons with dysfunction of facial expressions, as we expect their ground-truth to be ill-defined. This is due to the circumstance, that an incorrect execution of an exercise may resemble other exercises (Figure 6).

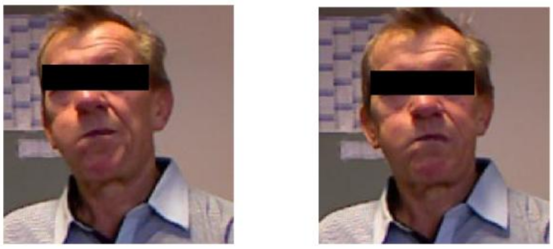

Figure 6 Patient with facial paresis on his right side. Left image: The exercise *right cheek puffed* is conducted correctly, because the bulge of the cheek is a passive process as reaction of a higher air pressure inside the mouth and a contraction of the buccinators on the left facial side. Right image: The exercise *left cheek puffed* is conducted incorrectly. The lack of contraction in the right buccinators leads to the bulge of the right cheek.

4.3 Choice of suitable features

Looking at the example images of Figure 5 reveals that the execution of the exercises has strong and manifold impact on the facial surface. Whereas the exercises *pursed lips* and *A-shape* lead to a rather concave cheek surface, the other exercises produce a convex curvature. But even the convex surfaces are manifold. The *cheek boxing* exercise results in a rather steep and local bulge, whereas the *cheeks puffed* exercise causes a more global and smooth bulge. Exercises with a wide mouth, like *taut lips* and *I-shape*, produce small wrinkles. However, the magnitude of the surface bulge differs between individuals because it depends on the face type (e.g., full versus slim). Nevertheless, the shape of the face surface is a reasonable property to separate the different appearances of the face as shown in earlier works of [13], and [14]. In total, we use three depth feature types that analyze the shape of the surface: curvature type histograms, point signatures and line profiles. They will be discussed more detailed in the following sections.

To capture depth data, we use a Kinect camera. The camera outputs 2.5D depth images. These are two-dimensional images – similar to a gray-value image – that contain object-to-camera distance information in each pixel instead of intensity information. In addition to the depth image, the Kinect simultaneously captures a color image. Via camera calibration, the intrinsic and extrinsic camera parameters can be determined [15]. Using the information of the 2.5D depth image, these can be employed to generate a 3D point cloud. Figure 7 shows a 2.5D depth image, the corresponding color image and a 3D point cloud.

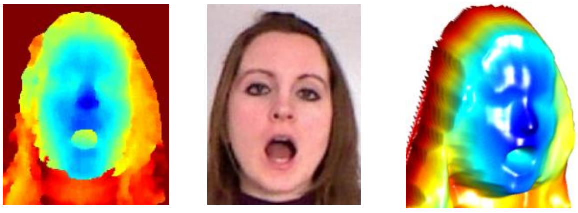

Figure 7 2.5D image, its corresponding color image and the generated 3D point cloud. For better visualization, the point cloud is shaded using Gouraud's method. Depth information in the 2.5D image is visualized by colors. The scale reaches from dark blue (close) to dark red (far).

4.3.1 Curvature type histograms

We determine the curvature type for each pixel of the face ([11], [16]). The curvature type contains information about the surface that is surrounding the pixel. This information comprises the direction of the surface curvature (convex, concave) and its shape (hyperbolic, cylindric, and elliptic). There are eight different types of curvature. Figure 8 shows four examples.

In an image, the face is represented by 8.000 to 13.000 pixels. If – for each pixel and its neighborhood – the curvature type is determined this results in a feature vector with a length similar to the number of pixels. In order to reduce the dimension of the feature vector, we summarize the curvature values with a histogram. To maintain spatial information, we define several facial regions from which separate histograms are extracted. Here, our approach follows the work of [13], who focus on the classification of six facial expressions. They divide the face into seven regions (e.g., chin, lower cheek, upper cheek) and summarize the curvature types with histograms. In their dataset that is used for testing purposes regions for feature extraction were localized manually by humans. In contrast, we detect the regions automatically, which is less accurate than manual labeling. As a result, we have reduced the number of regions from seven to four in order to increase the size of each region (Figure 9). This decreases the influence of small variations of the region border locations, but also decreases the accuracy. The borders of each of the four regions are determined by connecting fiducial points of the face. To enable a stable detection of the regions it is important that the fiducial points can be localized easily. Suitable positions lie in distinctive areas of the face that are only slightly influenced by changes of the face surface. This enables a good detection of the same point in different images.

In Figure 10, we show examples for the distribution of curvature types in the left cheek region for two different facial expressions. The curvature types are represented by different colors.

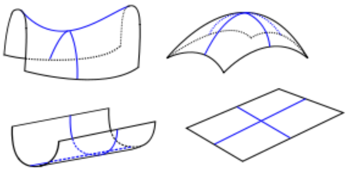

Figure 8 Examples of curvature types. Top left: hyperbolic convex, top right: elliptic convex, bottom left: cylindric concave, bottom right: planar.

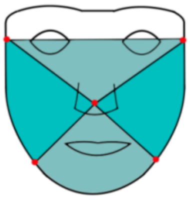

Figure 9 Four regions that are used for feature extraction. Borders of the regions are determined by fiducial points (red).

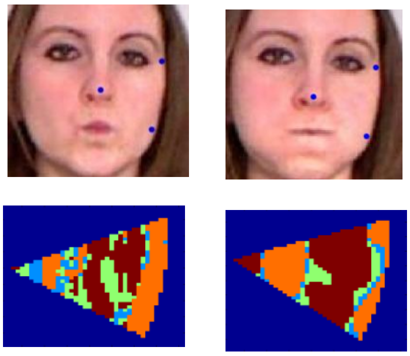

Figure 10 Top: Person performing the exercises *pursed lips* and *both cheeks puffed*. The blue points mark the corners of the left cheek region of the person. Bottom: Detail view for the left cheek area, showing the curvature types represented by colors (brown: elliptic convex, orange: eliptic concave, green: hyperbolic convex, blue: hyperbolic concave). As expected, the cheek has a large amount of elliptic convex area.

4.3.2 Point Signatures

In [14] point signatures are employed for the recognition of faces. We adapt this approach for our task of therapeutic face exercise classification. Similar to curvatures, the idea of point signatures is to describe the properties of the surface shape. Point signatures capture the slope of a path that runs around a distinctive point in the face to describe the neighborhood of this point. We selected the nose tip as centre point because it can be detected more robustly.

The point signature is calculated as follows: A sphere is centered into the nose tip. The intersection of the sphere with the facial surface creates the path (left image of Figure 11). To capture the slope of the path, the distance information of the points that lie on the path needs to be sampled. Depending on the position of the person to the camera, the absolute distance values vary, although the face may be identical. To obtain a distance measure relative to the position of the person, we fit a plane into the intersection points and displace this plane along its position vector until it goes through the tip of the nose (right image of Figure 11).

The distance of the curve to the displaced plane is now sampled in regular steps of 15 degrees. The slope of the path can be visualized by a coordinate system with the axes ‘degree’ and ‘depth distance’ (Figure 12).

The size of the radius is determined by multiplying the distance between the eyes with a factor *f*. We use the following values for *f*: 0.4, 0.5, 0.7, 0.8 and 1.0. As a result, we get five point signatures with a length of 24 samples each. To reduce the length of the resulting feature vector, we apply a discrete cosine transform (DCT) on each point signature and retain the first twelve coefficients [17].

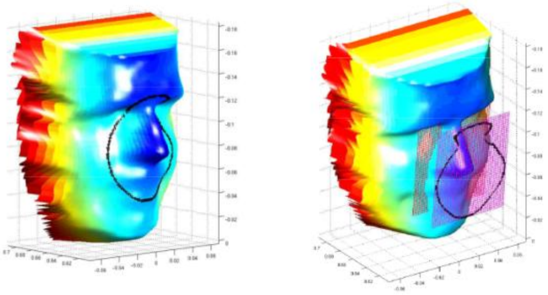

Figure 11 Left image: Intersection path of the 3D face point cloud and a sphere. Right image: Plane fitted in the intersection points (red) and the displaced plane (magenta) that is used for distance calculation. The black curve on the plane is the projection from the intersection curve. The distance between these two curves is sampled in an interval of 15 degrees.

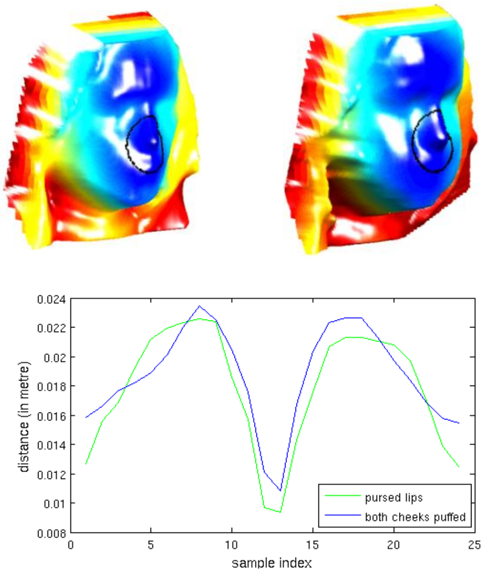

Figure 12 Top: Person doing the exercises pursed lips and both cheeks puffed. Point signature paths (for *f*=0.5) are marked on both faces. Bottom: Curves showing the slopes of the paths. Each curve consists of 24 samples (360°:15°= 24). The sample index multiplied with the sampling interval (15 degrees) results in the size of the angle, starting from the point on the curve that is intersected by an imaginary connection of the nose tip to a point on the center of the chin. The middle of the curve represents the root of the nose, which has the smallest distance to the displaced plane. At the beginning and the end of the curves it can be seen that the distance of the lips to the displaced plane is smaller for the exercise *pursed lips* (green curve) than for the exercise *both cheeks puffed* (blue curve).

4.3.3 Line Profiles

We developed the line profiles on the basis of the point signatures. Whereas a point signature consists of a path that runs radially around a point, a line profile connects two landmark points. We selected line profile paths that comprise the cheeks and the mouth because these regions show characteristic changes if a face performs facial exercises. The paths can be seen in Figure 13. Seven paths run from the tip of the nose to silhouette landmark points. The two remaining paths connect silhouette points. A path consists of *N* equidistant points in a three-dimensional space. To obtain a representation of the path, which is invariant with respect to translation and rotation operations of the face, we need to extract relative distance values. Therefore, we extract the Euclidean distances between the points. The number of points per path depends on the size of the face and the executed exercise. To get a constant length and to reduce the length of the feature vector, again we apply a discrete cosine transform and retain the first twelve DCT coefficients.

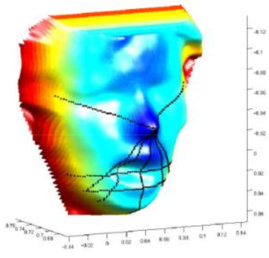

Figure 13 3D face with the marked paths of the 9 line profiles curves.

4.4 Feature evaluation: Results

In the preceding sections, we introduced three feature types that comprise information about the surface of a face:

- curvature type histograms
- point signatures
- line profiles

In the following, the features are examined with respect to their ability to discriminate between the executed nine therapeutic exercises. We assess the quality of this ability by the average recognition accuracy, which describes the ratio of the correctly detected exercises to the total number of exercises. According to the number of images in our dataset the total number of exercises is 696.

Feature types are evaluated individually and in combination. As mentioned in section 4.1, several steps are necessary for the evaluation of the features' suitability for the planned scenario. We concentrate on two of these aspects. First, we evaluate the features that were extracted from manually labeled regions in order to exclude other influences like deviating region borders. Second, we evaluate the features that were extracted from automatically determined regions using an AAM and a curvature-based nose tip detection.

Training and classification is performed by applying linear Support Vector Machines [18]. The dataset was split up into training and test set using the leave-one-out cross-validation. Additionally, all images of a person that is present in the test image are excluded from the training set. This approach is consistent with the mentioned application scenario in which the images of the test person will not be part of the training data. The number of feature dimensions was reduced from 232 to 8 by using a Linear Discriminant Analysis [19]. If features are extracted from manually labeled regions, line profiles perform better than the other features. However, the best result is obtained by the combination of the three types. This results in an average recognition accuracy of 91.2%. The performance of the curvature type histograms is rather low compared to the other two feature types. However, curvature type histograms outperform point signatures and line profiles if automatically detected regions are used. This is due to the fact that curvature features extract information from larger regions than point signatures and line profiles. As a result, the curvature type histogram is more robust against small variations of the region borders. Again, the combination of the three features leads to the best performance and results in an average recognition rate of 75.1%. This result confirms the suitability of the features for the classification of the presented therapeutic facial exercises, even in an automated scenario.

The deviations of the landmarks, determined by the AAM, compared to the position of the manually labeled landmarks were -1.9 pixels (mean value) in x-direction with a standard deviation of 4.7 pixels. In y-direction the mean value of the deviation was 6.0 pixels with a standard deviation of 15.9 pixels. Considering the distances of the persons to the camera, six pixels correspond to about 0.95 centimeters on the face.

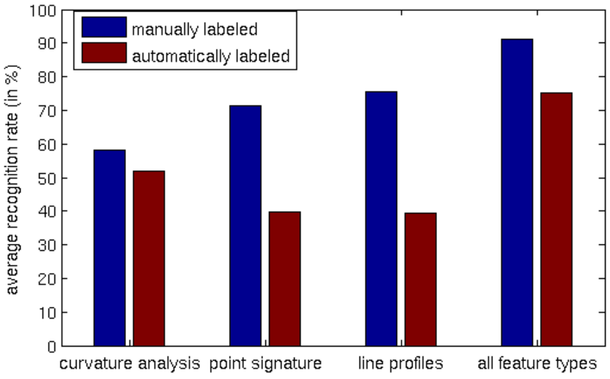

Figure 14 Results for the single feature types and their combination. The blue bars represent the results for the features that are extracted from manually labeled regions. The red bars show the results for the features extracted from automatically detected regions.

5 Conclusion and future work

In this publication, we presented our state-of-work for the development of an automated, therapy-accompanying training system. On the basis of existing therapy solutions and conversations with speech-language therapists, we derived a theoretical model that supports the conceptual design and implementation of such a system. Furthermore, we presented nine facial exercises, which were – in cooperation with therapists – determined as beneficial for the therapy of facial dysfunctions. On the basis of the selected exercises, we collected and manually labeled a dataset that comprises 696 depth images with their corresponding color images. This dataset was used to evaluate features that are the fundament for the implementation of an automated face analysis unit. The features were examined in two respects. First, we evaluated their discriminative power concerning the classification of different exercises. Second, we tested their robustness regarding varying locations of feature extraction. The latter is relevant to determine, whether these features are suitable for a real-world application. Future work will be focused on the evaluation of the features' suitability for the separation of different states of an exercise. Furthermore, we will examine the mapping of the feature values to an evaluative and instructive feedback scale.

Acknowledgements

We would like to thank the *m&i Fachklinik Bad Liebenstein* (in particular Prof. Dr. med. Gustav Pfeiffer, Eva Schillikowski) and *Logopädische Praxis Irina Stangenberger*, who supported our work by giving valuable insights into rehabilitation and speech-language therapy requirements and praxis.
This work is partially funded by the TMBWK ProExzellenz initiative, Graduate School on Image Processing and Image Interpretation.

Literature

[1] Brach, J. and VanSwearingen, J.M. (1999). Physical therapy for facial paralysis: a tailored treatment approach. *Physical Therapy: Journal of the American Physical Therapy Association*, pages 397-404.

[2] Wolowski, A. (2005). Fehlregenerationen des Nervus facialis - ein vernachlässigtes Krankheitsbild. *Dissertation*. Universität Münster.

[3] Gross, H.-M., Schroeter, C., Mueller, S., Volkhardt, M., Einhorn, E., Bley, A., Langner, T., Merten, M., Huijnen, C., van den Heuvel, H., van Berlo, A. (2012). Further progress towards a home robot companion for people with mild cognitive impairment. Proc. *IEEE Int. Conf. on Systems, Man, and Cybernetic,* Korea, Seoul, pages 637-644.

[4] Geue, P.-O., Scheidig, A., Kessler, J., and Gross, H.-M. (2012). Entwicklung eines robotischen Bewegungsassistenten für den Langzeiteinsatz zur physischen Aktivierung von Senioren. *Ambient Assisted Living Kongress*, Deutschland, Berlin, 5 pages.

[5] Lanz, C, Denzler, J., Gross, H.-M. (to appear). Automated classification of therapeutic face exercises using the kinect. *Int. Conf. on Computer Vision Theory and Application (VISAPP 2013)*, Spain, Barcelona.

[6] Shiau, J., Segal, B., Danys, I., Freedman, R. and Scott, S. (1995). Long-term effects in neuromuscular rehabilitation of chronic facial paralysis. *The Journal of Otolaryngology*, 24(4):217-220.

[7] Halton, J. (2008). Virtual rehabilitation with video games: a new frontier for occupational *therapy. Occupational Therapy Now*, pages 12-14.

[8] John, M., Häusler, B., Frenzel, M., Klose, S. and Ernst, T. (2009). Rehabilitation im häuslichen Umfeld mit der Wii Fit – Eine empirische Studie. *Ambient Assisted Living*.

[9] Cootes, T., Edwards, G., Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681-685.

[10] Stricker, R., Martin, Ch., Gross, H.-M. (2009). Increasing the robustness of 2D active appearance models for real-world application. *Proc. IEEE Int. Conf. on Computer Vision Systems*, Belgium, Liege, pages 364-373.

[11] Colombo, A., Cusano, C. and Schettini, R. (2006). 3d face detection using curvature analysis. *Pattern Recognition*, 39(3):444-455.

[12] Gordon-Brannan, M.E. (2007). *Clinical management of articulatory and phonologic disorders.* Lippnicot Williams & Wilkins.

[13] Wang, J., Yin, L., Wei, X., and Sun, Y. (2006). 3d facial expression recognition based on primitive surface feature distribution. *Int. Conf. on Computer Vision and Pattern Recognition*, pages 1399-1406.

[14] Wang, Y., Chua, C.-S., and Ho, Y.-K. (2002). Facial feature detection and face recognition from 2d and 3d images. Pattern Recognition Letters, 23:1191-1202.

[15] Hartley, R. and Zisserman, A. (2004). *Multiple view geometry in computer vision*. Cambridge University Press.

[16] Besl, P. and Jain, R. (1986). Invariant surface characteristics for 3d object recognition in range images. *Computer vision, Graphics, and Image Processing*, 33(1):33-80.

[17] Salomon, D. (2011). *Data compression: the complete reference.* Springer-Verlag New York Inc.

[18] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, pages 1-27.

[19] Webb, A., Copsey, K., & Cawley, G. (2011). *Statistical pattern recognition*.Wiley.