Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

Facial Landmark Localization and Feature Extraction for Therapeutic Face Exercise Classification

Cornelia Lanz[1], Birant Sibel Olgay[1], Joachim Denzler[2], and Horst-Michael Gross[1]

[1] Neuroinformatics and Cognitive Robotics Lab, Ilmenau University of Technology, Ilmenau, Germany
[2] Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany

Abstract. In this work, we examine landmark localization and feature extraction approaches for the unexplored topic of therapeutic facial exercise recognition. Our goal is to automatically discriminate nine therapeutic exercises that have been determined in cooperation with speech therapists. We use colour, 2.5D and 3D image data that was recorded using Microsoft's Kinect. Our features comprise statistical descriptors of the face surface curvature as well as characteristic profiles that are derived from face landmarks. For the nine facial exercises, we yield an average recognition accuracy of about 91% in conjunction with manually labeled landmarks. Additionally, we introduce a combined method for automatic landmark localization and compare the results to landmark positions obtained from Active Appearance Model fitting as well as manual labeling. The combined localization method exhibits increased robustness in comparison to AAMs.

1 INTRODUCTION

Diseases like stroke or mechanical injury of the facial nerve can lead to a dysfunction of facial movements. These impairments of facial expressions and muscle control may have various consequences like eating difficulties and impaired face appearance, which can restrict daily life and can lead to social isolation. Similar to rehabilitation exercises that help to regain body functions, there are exercises for the recovery of facial expressions. Besides practising under supervision of a speech therapist, patients additionally have to conduct unattended exercises on their own. However, the incorrect conduction of exercises can impede the training success or even lead to further impairment. An accompanying training platform could enrich unsupervised training exercises by a feedback functionality [1].

The design and implementation of such a training platform is a challenging and complex task that comprises several subtasks. In this work, we will focus on two subtasks – the automated facial landmark localization and the evaluation of features. However, in order to enable a better understanding of the context of our work, we also give a brief overview of the remaining subtasks. Figure 1a presents five of the involved subtasks, which will be discussed in the following.

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

Facial movements cause changes of the face surface, which can be captured by depth image sensors like Microsoft's Kinect[3] or Time-of-flight Cameras[4],[5]. The extraction of *depth features* (see Fig. 1a) allows to examine the face surface, independently from skin colour and lighting conditions. Although there exist other systems that are capable of recording depth data with much higher depth resolution than the Kinect (e.g. [2]), we decided to use this sensor because of its moderate price. This makes our target application suitable for widespread use in low-cost training platforms. Furthermore, the Kinect allows to capture additional data channels such as intensity images in parallel to depth images. These might be helpful if depth information is not suitable to describe certain facial movements. For example, it can hardly be determined whether the eyes are closed by solely processing depth information. In a real-world scenario, where regions for feature extraction should be detected automatically, we additionally need a fully *automated facial landmark localization*.

The nine therapeutic face exercises that we focus on in this paper are rather static. The pace of the exercise conduction from neutral face to final state, e.g., both cheeks puffed, is not important. It is more relevant that the exercises final states are retained for a few seconds. Nevertheless, it is likely that additional information, obtained by examining the *dynamics* of an exercise instead of single *static* snapshots, may contain valuable information. Additionally, it is possible to reduce the amount of noise in the data by smoothing over time.

The *evaluation of the exercises*, which is essential for a feedback functionality, is a complex task. Besides the choice of appropriate technical tools, it is necessary to define in which cases an exercise is performed correctly and in which not. Additionally, it needs to be assessed how feedback should be communicated in order to be most beneficial for a patient.

Furthermore, it is necessary to collect a *database* of training and test images that contain the exercises performed by healthy people as well as the exercise conduction by people with dysfunction of facial expression abilities. In our experiments, nine therapeutic facial exercises are employed that had been defined in cooperation with speech therapists. In our studies, we only use training and test data recorded from exercises of healthy persons. 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 incorrect conduction of an exercise may resemble other exercises, as shown in Fig. 1b.

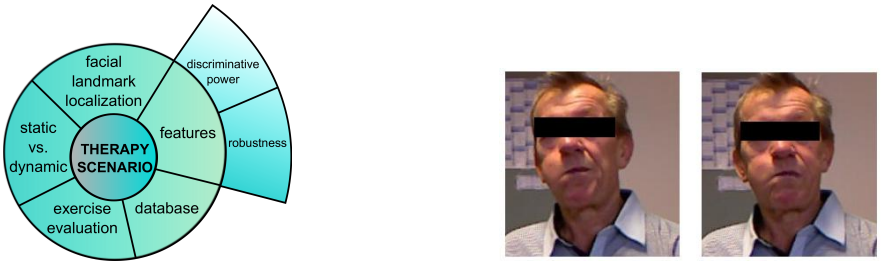
Since each of the above-mentioned subtasks covers diverse aspects, we focus on the landmark localization and the succeeding feature extraction for therapeutic exercise classification here. Our depth features are extracted from 2.5D images and 3D point clouds recorded by the Kinect Sensor. We refer to 2.5D images as 2D images that contain the object-to-camera distance instead of the object's intensity value. We analyse the facial surface by extraction of curvature information and surface profiles. Surface profiles comprise line profiles and

[3] http://www.xbox.com/en-US/kinect
[4] http://www.pmdtec.com/
[5] http://www.mesa-imaging.ch/

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

point signatures. Line profiles are based on paths that connect two landmark points, whereas point signatures are based on radial paths around single landmark points.

We examine the features' *discriminative power* with respect to the classification of nine therapeutic exercises and their *robustness* regarding varying feature extraction regions. In the targeted real-time scenario, regions and points for feature extraction need to be determined automatically. We expected that this step leads to variations from manually located face regions and landmarks. Therefore, it is necessary that the features are robust against these deviations. Two different approaches for automated landmark localization have been tested: Active Appearance Models [3] and a combined approach that consists of learned spatial relations of the facial landmarks and tree-structured parts models.



(a) (b)

Fig. 1: (a) Different subtasks of the design and implementation of an automated therapeutic exercise platform. (b) 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 buccinator on the left facial side. Right image: The exercise *left cheek puffed* is conducted incorrectly. The lack of contraction in the right buccinator leads to the bulge of the right cheek.

2 RELATED WORK

Automated recognition of therapeutic face excercises is a still relatively unexplored research field. In practice, there are already tools that support the patient with regard to exercising that is not supervised by a therapist. These tools comprise video tutorials (*LogoVid* [6]) or exercise diaries (*CoMuZu* [7]). However, at this moment there are no commercial solutions available that automatically recognize and evaluate a performed therapeutic exercise.

[6] http://www.comuzu.de
[7] http://www.logomedien.de/html/logovid7a.html

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

In [4] the benefit of facial exercises for the prevention of synkinesis after facial paresis is analyzed. Synkinesis is an involuntary associated facial movement such as eye closure during smiling. In order to determine the grade of synkinesis, [4] manually measure the eye opening width by using an image editing software. [5] present a system for the diagnosis support of patients with facial paresis using 2D colour images. Therefore, they analyse facial asymmetries in the eyes, nose and mouth regions.

At present, there are no publications known to us that focus on the automated recognition of therapeutic facial exercises using depth information. Nevertheless, we can utilize approaches from works on face detection, as well as person and emotion recognition. [6] use curvature of the surface of a 2.5D image to detect salient face features, like eyes and nose. A triplet consisting of a candidate nose and two candidate eyes is processed by a classifier that is trained to discriminate between faces and non-faces. Based on curvature information estimated on a 3D triangle mesh model, [7] classify 3D faces according to the emotional state that they represent.

Point signatures were developed by [8] as an approach for general 3D object recognition. Additionally, in [9] they present an enhanced point signature algorithm that is specialised on face recognition. [10] extract point signatures in 2.5D images and Gabor filter responses in gray-level images and employ their combination for face recognition.

In this work, we follow the method proposed in [7] to create histograms of curvature types. We utilize the face recognition algorithm from [9] for the classification of our nine therapeutic exercises and supplement it with a similar approach that employs line profiles instead of radial profiles. In contrast to [7], where manually placed landmarks are used, we additionally evaluate our results with automatically located landmark positions.

3 METHOD

In the following, we briefly summarize the determination of surface curvature (section 3.1) as far as it is necessary to understand the basic principles of our curvature feature types (section 3.2). For detailed information, we refer to [11]. In sections 3.3 and 3.4 the extraction of line profiles and point signatures is presented. In the last section, we focus on the automation of the feature extraction process.

3.1 Curvature Analysis

Our aim is the classification of faces according to the therapeutic exercises a patient performs. Facial movement leads to a change of the face surface. We analyse the surface by extracting curvature information from 2.5D range images and 3D point clouds. The parametric form of a surface in 3D is $\mathbf{s}(u, v) = [x(u, v)\ y(u, v)\ z(u, v)]^T$, with u and v denoting the axes of the parameter plane (Fig. 2a). Based on this function, we can determine the first and the second

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

fundamental forms, which uniquely characterize and quantify general smooth shapes. The elements of the first fundamental form $\mathbf{I}$ are:

$$\mathbf{I} = \begin{bmatrix} \mathbf{s_u} \cdot \mathbf{s_u} & \mathbf{s_u} \cdot \mathbf{s_v} \\ \mathbf{s_u} \cdot \mathbf{s_v} & \mathbf{s_v} \cdot \mathbf{s_v} \end{bmatrix} . \tag{1}$$

The subscripts denote partial differentation. The elements of the second fundamental form $\mathbf{J}$ are:

$$\mathbf{J} = \begin{bmatrix} \mathbf{s_{uu}} \cdot \mathbf{n} & \mathbf{s_{uv}} \cdot \mathbf{n} \\ \mathbf{s_{uv}} \cdot \mathbf{n} & \mathbf{s_{vv}} \cdot \mathbf{n} \end{bmatrix} , \tag{2}$$

with $\mathbf{n}$ being the unit normal vector of the tangent plane in the point with parameters (u, v). Although both fundamental forms are a unique representation of the surface, combinations of both are more common for surface characterization, because they allow for an intuitive interpretation. Using $\mathbf{I}$ and $\mathbf{J}$, the shape operator matrix $\mathbf{W}$ can be computed by:

$$\mathbf{W} = \mathbf{I}^{-1} \cdot \mathbf{J} . \tag{3}$$

The mean curvature H gives information about the direction of the curvature (convex, concave) and is determined by:

$$H = \frac{1}{2}\, tr\left[\mathbf{W}\right], \tag{4}$$

with $tr\left[\mathbf{W}\right]$ being the trace of the shape operator $\mathbf{W}$. The Gaussian curvature K contains the information whether curvatures that are orthogonal to each other point in the same or in different directions (Fig. 2b). It is computed as follows:

$$K = det\left[\mathbf{W}\right]. \tag{5}$$

Opposed to the general parametric representation, the parametrization of a 2.5D range image takes a very simple form $\mathbf{s}(u, v) = [u \;\; v \;\; z(u, v)]^T$. Because a 2.5D image is spanned by two axes that generate a discrete (pixel) grid, the derivation of $\mathbf{s}$ with respect to u and v is simplified and results in $\mathbf{s_u} = [1 \;\; 0 \;\; z_u]^T$ and $\mathbf{s_v} = [0 \;\; 1 \;\; z_v]^T$. Therefore, for the computation of H and K only the partial derivatives of z are relevant:

$$H = \frac{z_{uu} + z_{vv} + z_{uu} z_v^2 + z_{vv} z_u^2 - 2 z_u z_v z_{uv}}{(1 + z_u^2 + z_v^2)^{\frac{3}{2}}} , \tag{6}$$

$$K = \frac{z_{uu} z_{vv} - z_{uv}^2}{(1 + z_u^2 + z_v^2)^2} . \tag{7}$$

3.2 Extraction of Curvature Information

Prior to feature extraction, we smooth the face surface using an average filter. We extract the mean and Gaussian curvature for each pixel, respectively 3D-point, in order to obtain information about the facial surface. This results in

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013,
Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

around 2×8000 to 2×13000 values per face, depending on the face-to-camera distance. In order to reduce the dimensionality of the feature space, we accumulate the curvature values in a histogram [7]. To maintain spatial information, we define four regions (A-D) from which histograms are extracted (Fig. 2c). Each histogram is weighted with the number of pixels of the region described by it. The selected cheek regions are axially symmetric, due to the fact that some of the therapeutic exercises are asymmetric and each face side contains valuable information. Two additional regions, in which high facial surface variation among all exercises is visible, were included into the feature extraction process. Further refinement of the regions was omitted in order to to maintain a certain robustness in case of automatically determined regions.

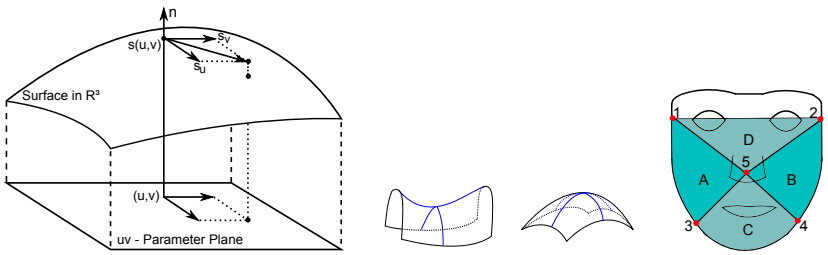


Fig. 2: (a) Surface in 3D with the corresponding parameter plane (image according to [11]). (b) Two surfaces with orthogonal maximum and minimum curvatures that point in different (left surface: hyperbolic convex) and in the same directions (right surface: elliptic convex). (c) Regions A-D are employed for curvature feature extraction. Region borders are derived from landmark points 1-5. The determination of the landmark points is explained in sections 3.5 and 4.1.

The curvature type histogram feature is obtained by extraction of mean curvature H and Gaussian curvature K for every 2.5D pixel, respectively 3D point according to (4)-(7). In the next step, both values are combined to eight discrete curvature types as shown in Table 1 [6]. Subsequently, the discrete curvature types of each region are summarized with histograms. The concatenation of these histograms forms the feature vectors that are subjected to the classification process. For each image, a 32 dimensional feature vector is extracted (8 curvature type histogram bins per each of the four regions).

3.3 Extraction of Line Profiles

Although curvatures are extracted from each pixel, their combination in a histogram blots out some of the local information. Line profiles, in comparison, contain highly localized information by describing paths along the face surface. Instead of using 2.5D images, line profiles are extracted from a point cloud in

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

Table 1: Curvature type definition using mean and Gaussian curvature (H, K)

	$K < 0$	$K = 0$	$K > 0$
$H < 0$	hyperbolic concave	cylindric concave	elliptic concave
$H = 0$	hyperbolic symmetric	planar	impossible
$H > 0$	hyperbolic convex	cylindric convex	elliptic convex

3D. Each of the three dimensions is expressed in meter. For a 2.5D image, two dimensions are given in pixel units. However, the real world distance that is described by the difference of one pixel depends on the person-to-camera distance. The smaller the distance of an object to the camera is, the more pixels does this object cover on a 2.5D image. As a result, comparison of different line profiles is more difficult, when using 2.5D images.

In total, we extract nine line profiles from the 3D point cloud of the face. Every line profile connects two defined landmark points. Figure 3a shows the paths of the profile lines. Seven profiles start at the nose tip, connecting it in radial direction to silhouette points. Two line profiles are horizontally located and link two silhouette points.

The paths over the face consist of N equidistant points $p_n(x, y, z)$, with $n = 1...N$. Nearest-neighbour interpolation is employed in order to calculate missing points. The L2-norm of the position vectors of every 3D point p_n already creates a distinctive curve as can be seen in Fig. 3b. However, in order to achieve invariance with respect to the viewpoint (i.e., translation and rotation operations of the facial point cloud), relatively coded central differences between the 3D points are calculated (left image of Fig. 3c).

The images show, that the curves consist of 70 samples. This value may vary because the size of the head (subject-specific) or the length of the curve (exercise-specific) may change. To get an identical size of the curve for every subject and every exercise and to reduce the amount of feature dimensions, we conduct a discrete cosine transform [12] on the curves and build our feature vector using the first 12 dct-coefficients. The right image of Fig. 3c shows, that the inverse discrete cosine transform with 12 coefficients yields a reasonable reconstruction of the original curve. We derived the line profiles from the point signature approach presented in the following section.

3.4 Extraction of Point Signatures

Similar to line profiles, point signatures are paths on a surface [8]. Instead of connecting two landmark points, the curve runs radially around a distinctive point p_0 of a 3D point cloud. As can be seen in Fig. 4a, in our approach the point p_0 is located on the tip of the nose. In order to obtain the point signature, a sphere is centered into the point p_0 of the 3D point cloud. The intersection of the sphere with the facial points forms a curve Q in the three-dimensional space (Fig. 4b). The depth information of these intersection points, combined with the value of the sphere radius, contains characteristic and unique information about

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

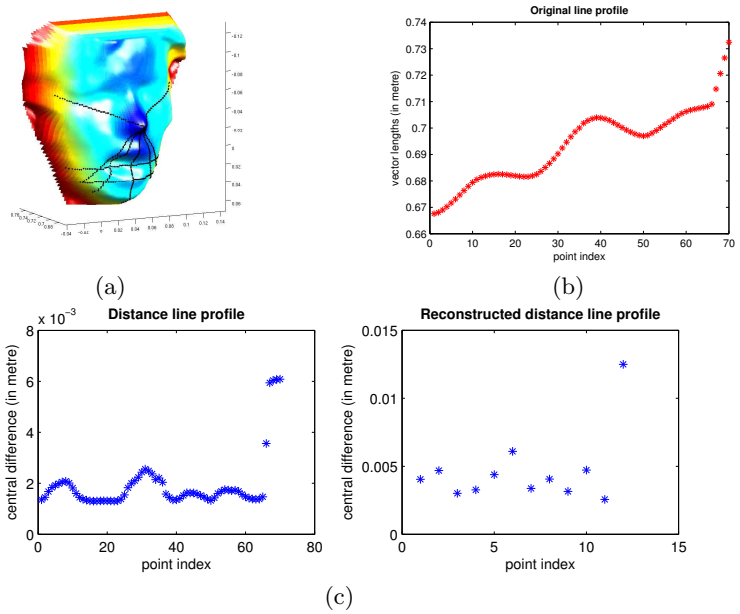


Fig. 3: (a) 3D face with marked paths of the nine line profiles. (b) Line profile from nose tip to the point of the chin for the exercise *A-shape*. The curve shows the length (in metre) of the position vector of each point p_n. The opening of the mouth, resulting in higher values, in the middle of the curve and the chin shape on the right are visible. (c) Left: Distance line profile. Right: The reconstructed line profile using the first 12 dct-coefficients.

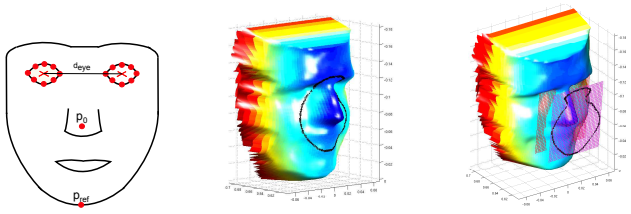
the depth value distribution in the surrounding area of the point p_0. However, taking the absolute depth values of this intersection points is not reasonable (as already discussed for the line profiles in section 3.3) because they are not independent with respect to translation and rotation of the head. As a result, we create a reference curve Q' that can be employed to calculate relative depth information. To obtain this curve, we fit a plane P through the set of intersection points. The plane is determined with regression analysis by a singular value decomposition that gives the surface normal of the plane. The plane is now shifted along its normal vector into the point p_0. This results in a new plane called P' (Fig. 4c).

In the next step, the curve Q is projected onto P' building a new curve Q'. Now the curve Q' is sampled around the approximated surface normal at p_0 with a rotation angle of 15 degrees. For each sampled point in Q' the distance to its corresponding point in Q is collected. The starting position for the distance sampling needs to be equal between the different images to obtain curves that are comparable. Therefore, we define a starting position, which is determined by a reference point p_{ref}. The reference point is located on the chin as marked

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

in Fig. 4a. The sphere radius length has to be determined such that the arising path does not protrude beyond the surface of the face and no background points are sampled. The length of the radius is computed from the eye distance d_{eye}, multiplied by a factor f. The eye distance is estimated from the distance between the mean positions of each eye that are obtained by the landmark positions of each eye (Fig. 4a). We use the following values for the empirically determined factor f to extract five different point signatures that cover varying areas of the face: 0.4, 0.5, 0.7, 0.8 and 1.0.

Sampling of the radial curve with a fixed interval of 15 degrees generates 24 values per point signature. The more point signatures are extracted, the more precisely the surface of the face can be described. However, a high amount of point signatures leads to a high-dimensional feature space. Again, we reduce the dimension of the feature vector to twelve values by applying discrete cosine transform on each point signature as shown in section 3.3.



(a) (b) (c)

Fig. 4: (a) Landmark points and line segments that are employed for the extraction of point signatures. (b) Intersection curve Q of the sphere with the 3D point cloud. (c) The planes P (red) and P' (magenta). The projected curve Q' is marked on P'.

3.5 Automation of the Feature Extraction Process

The features presented above have in common that distinct facial areas need to be determined for extraction. Manual determination of these landmarks and regions is not feasible in a real-world application. Thus, they have to be detected automatically, which may lead to less accurate localizations. In this work, we compare two different approaches for landmark localization: Active Appearance Models and a tree-structured parts model algorithm that is combined with a 3D spatial relations model.

AAMs are mainly applied in the field of facial expression recognition on 2D gray-value images ([3], [13]). On the basis of several training images a combined mean texture and shape model is derived. The fitting of this mean model to a new and unknown face is improved by determination of a coarse initialization position using the Viola and Jones face detector [14]. In the next step the AAM adapts

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

itself to the new face by minimizing the error between the model intensities and the image intensities. The parameters that describe the fitted model are usually subjected to classification of facial expressions. In constrast to this, the AAM can be used for the mere detection of landmarks without further consideration of the model parameters [15]. In this paper, we focus on the application of AAMs for the detection of the 58 landmarks only (Fig. 5b).

Tree-structured parts models are an approach for face detection, pose estimation and landmark localization [16]. In total 68 landmarks are located in this approach. The number of landmarks on the face silhouette is similar to the number of silhouette landmarks detected by the AAM approach. However, tests showed that the placement of landmarks in the center of the face, e.g., in the nose or eye region, is too imprecise for the targeted scenario. Therefore, only the information of the silhouette landmark positions is kept.

A spatial relations model and surface curvature are computed in a parallel process in order to localize the landmarks in the upper, rigid face half (Fig. 5a). The spatial relations model comprises a smaller subset of landmarks, which was derived from the landmarks and regions that are necessary for feature extraction. The idea of the spatial relations model is based on the fact that distances and angles between the landmarks of a face lie in a constrained range. The model is computed from training data and centered in the nose tip of a face (Fig. 5c). In total, 14 position vectors show the direction and absolute value to 14 landmarks (Fig. 5d). Additionally, for each landmark the maximum deviation of the training data from the mean position is computed. As a result, a spherical search space can be constructed around each position vector tip by using the maximum deviation distance as radius. In order to be able to fit the model to a new image with unknown landmark localizations, the nose tip and the nose ridge vector must be detected (Fig. 5e). This can be done via curvature analysis and Support Vector Machine (SVM) classification because of the distinctive surface of the nose. The 3D mean model is then translated and rotated so that the model reference vector and the nose ridge vector are congruent. Possible landmark candidates lie in each of the 14 spherical search spaces that are centered at the tip of a vector. Now, the previously computed curvature information can additionally be used as input for 14 single SVMs in order to further reduce the landmark candidate number. For each of the 14 landmarks a separate SVM is trained. In the end, for each landmark a centroid of the remaining candidates is computed and defined as the new landmark position.

In contrast to the rigid upper face half, the lower one has a more dynamic surface appearance. As a result, mean and Gaussian curvature are not appropriate for landmark localization in this area. In the last step, upper face half landmarks from the spatial relations model and lower face half landmarks from the tree-structured parts model are fused to one landmark set. Thus, at present, both processes are parallel and independent from each other. Our future goal is to combine the results of both approaches for complementary verification and error minimization.

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

The AAM and the tree-structured parts model are fitted on the 2D intensity images. Subsequently, we need to transform these landmark positions to positions in depth images. Therefore, intrinsic and extrinsic camera parameters were determined by camera calibration [17]. They can be employed to align the 2.5D images with their corresponding intensity images. As a result, corresponding points have the same position in the images of both channels, and the labeled landmark positions can be accordingly transferred. Additionally, these camera parameters can be used to transform the points of the 2.5D image to a discrete 3D point cloud [17].

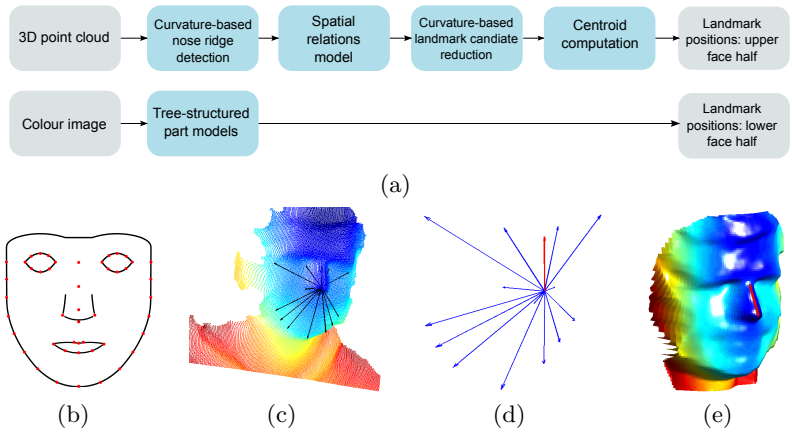


Fig. 5: (a) Process of landmark localization. (b) The 58 landmark positions of the Active Appearance Model. (c) Spatial relations model fitted on a face. (d) 3D spatial relations model. (e) Face with nose ridge vector (red).

4 EXPERIMENTS

In the first section of the experimental part, we focus on the dataset and the exercises that are used for our experiments. The evaluation of the features discriminative power with respect to the classification of therapeutic exercises is presented in section 4.2. Results from experiments that test the robustness of the features related to variations of region borders are given in the last section.

4.1 Exercises and Dataset

In cooperation with speech therapists, we selected a set of nine therapeutic face exercises by certain criterions. The exercises should train the lips, the cheeks, and the tongue and should be beneficial for various types of facial muscle dysfunctions, e.g. paresis of muscles or muscle imbalance. Furthermore, the selected

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

exercises should be easy to practice and should build a set of sub-exercises that can be combined to more complex dynamic exercise units, e.g. by alternating between them. The exercises have to be performed in an exaggerated manner, to enable a maximum training effect, and have to be retained for around two or three seconds. The speed of the performance is not important. Although some of these are vocal exercises, it is not necessary to vocalize a continuous sound while performing the shape. Images that visualize the exercise conduction are shown in Fig. 6.

Fig. 6: Exercises that have been selected in cooperation with speech therapists (l. to r.): pursed lips, taut lips, A-shape, I-shape, cheek poking (right/ left side), cheeks puffed (both/ right/ left side(s)). For better visualization colour images are shown.

Due to the lack of a public database that shows the performance of therapeutic exercises, we recorded a dataset, which contains eleven persons, who conducted the nine exercises. For each exercise, there are around seven images, showing different states of exercise conduction. This amounts to a total size of 696 images in the dataset. Some parts of the scene, which was captured by the Kinect may be shadowed, if they are seen by the depth camera but are not illuminated by the infrared projector. This leads to invalid values in the 2.5D image [18]. These values were removed by replacing them with the mean depth values of adjacent valid neighbour pixels. For every depth image, there exists a corresponding colour image that has been recorded with maximum time difference of 16 milliseconds. The colour images have been labeled manually with 58 landmark points that were used for the training of the AAM (Fig. 5b), or for the feature extraction from depth data. The transferability of landmark positions between the 2.5D image and the colour image was already explained in section 3.5.

4.2 Evaluation of the Discriminative Power

The following section gives an overview of the classification results. Since we wanted to evaluate the basic suitability of the described features for the task of classifying therapeutic exercises, we extracted the features from regions obtained via manually labeled landmarks, thus excluding other influences like deviating region borders. We evaluated each feature group individually and in combination. Training and classification was performed by applying SVMs of the LIBSVM package [19]. We tested linear SVM and a Radial Basis Function kernel. Optimal

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

values for the penalty parameter C and the kernel parameter γ were obtained by a grid search on the training set [20]. In order to avoid overfitting to the training set, we employed a 5-fold cross-validation during parameter optimization. In combination with the amount of data (696 images, 232 feature dimensions), the linear SVM led to the best results because it avoided overfitting. The dataset was split up into training and test set using the leave-one-out cross-validation. Additionally, all images of the person present in the test images were 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. Linear discriminant analysis (LDA) was used prior to the linear SVM classification in order to reduce the feature dimensions from 232 to 8. LDA is a linear transformation of the feature space that maximizes the between-class separability and minimizes the within-class variability [21]. We obtained an average recognition accuracy over the nine classes of 90.89 %. Detailed results for the single features are given in Fig. 8.

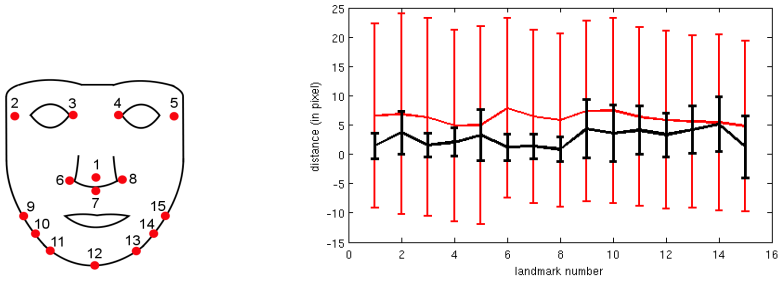
4.3 Evaluation of the Automated Landmark Localization

As mentioned before, in a real-world scenario regions and landmark points for feature extraction have to be detected automatically. Therefore it is crucial, to employ a robust landmark localization. Although AAMs usually comprise 58 landmarks, in this section we constrain our evaluation to the landmarks that are relevant for our succeeding feature extraction (Fig. 7a). Figure 7b shows the mean pixel distances and standard deviations between the manually labeled landmarks and the two automated localization approaches. The AAMs are visualized in red and the combined parts and spatial relations model approach is visualized in black. The localization using the combined approach led to smaller deviations than using AAMs. A deviation of six pixels corresponds to about 0.95 cm. Additionally, it can be seen that the landmarks in the upper rigid half of the face were more robustly detected than the landmarks in the lower face half. Better localization resulted from the more distinctive and invariant surface shape in these landmark areas. Furthermore, images were labeled manually on 2D colour images. The landmarks with the smallest deviations are landmarks that are easier to label in the colour image because of distinctive visual properties, e.g., the darker inner eye corners or the edge between cheek and nose wing.

4.4 Evaluation of Feature Extraction from Automatically Determined Regions

In this section, we evaluate the robustness of our different features types with respect to varying region borders and landmark positions. Figure 8 shows the results for each of the three feature types for manually and automatically localized landmarks. For manual determination of the landmark positions, curvature analysis is weaker than point signatures and line profiles with respect to the discrimination of nine therapeutic exercises. This result occured because curvature

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014



(a) (b)

Fig. 7: (a) The 15 landmarks that are relevant for feature extraction. (b) The plot shows the mean values and standard deviations for the distances (in pixel units) between the manually labeled landmark positions and the positions localized by AAMs (red) and by the combined approach (black). Six pixel correspond to about 0.95 cm.

information for several pixels is combined into histograms, and therefore, averaged over larger regions. However, curvature analysis achieved better results for automatically detected landmarks than the line profiles because it covers a larger region. Thus, small deviations of the silhouette landmarks have less influence on the regions used for feature extraction, especially if a landmark is incorrectly localized outside the face region.

As shown in section 4.3, the combined approach led to more robust landmark localization than the AAMs. As expected, this resulted in better average recognition rates for each of the feature types. By concatenating the different feature types a rate of 90.89 % correct exercise classification was obtained if manual labeling is used and 69.14 % if the combined localization approach is used.

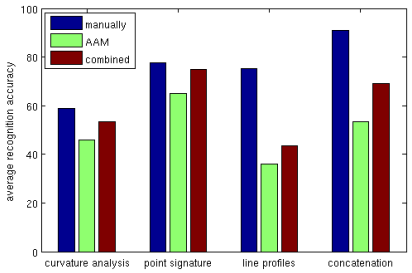


Fig. 8: The bar plot shows the average recognition rates (in %) for each of the three feature groups. As expected, feature extraction from manually determined regions and landmarks led to better results than the extraction from automatically determined areas using AAMs and combined models.

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

5 SUMMARY AND DISCUSSION

In this paper, we presented several aspects that are necessary for the design and implementation of an automated training platform for patients with facial muscle dysfuctions. We introduced nine therapeutic exercises, which - in cooperation with speech language therapists - were determined as beneficial for the planned application scenario. Additionally, the automated classification of these exercises was evaluated. The presented approach employs 2.5D depth images and 3D point clouds and is based on three different feature types: curvature analysis, point signatures, and line profiles. The features were evaluated with respect to their discriminative power for exercise classification. Additionally, we examined their robustness regarding varying locations of feature extraction. This is relevant for all applications, planned for practical use, where a manual detection of landmarks is not feasible.

Curvature analysis, in the form we have implemented it, is rather global compared to point signatures and line profiles and showed a relatively robust performance. However, with suitable landmark localizations point signatures and line profiles outperform curvature analyis. We used two approaches for automated landmark detection: Active Appearance Models and tree-structured parts models. The latter lead to the best results. Line profiles showed only weak contribution to the classification process, if the landmark positions are detected automatically. Nevertheless, the results based on manually defined regions are promising.

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.

References

1. Lanz, C., Denzler, J., Gross, H.M.: Facial movement dysfunctions: Conceptual design of a therapy-accompanying training system. In: Ambient Assisted Living - Advanced Technologies And Societal Change, Springer Heidelberg (2013)
2. Grosse, M., Schaffer, M., Harendt, B., Kowarschik, R.: Fast data acquisition for three-dimensional shape measurement using fixed-pattern projection and temporal coding. Optical Engineering **50** (2011) 100503
3. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6) (2001) 681–685
4. Nakamura, K., Toda, N., Sakamaki, K., Kashima, K., Takeda, N.: Biofeedback rehabilitation for prevention of synkinesis after facial palsy. Otolaryngology–Head and Neck Surgery **128**(4) (2003) 539–543

Computer Vision, Imaging and Computer Graphics - Theory and Applications: Revised Selected Papers of VISIGRAPP 2013, Communications in Computer and Information Science (CCIS), vol. 458, pp. 179-194, Springer 2014

5. Gebhard, A., Paulus, D., Suchy, B., Wolf, S.: A system for diagnosis support of patients with facialis paresis. Kl **3/2000** (2000) 40–42
6. Colombo, A., Cusano, C., Schettini, R.: 3d face detection using curvature analysis. Pattern Recognition **39**(3) (2006) 444–455
7. Wang, J., Yin, L., Wei, X., Sun, Y.: 3d facial expression recognition based on primitive surface feature distribution. Int. Conf. on Computer Vision and Pattern Recognition **2** (2006) 1399–1406
8. Chua, C.S., Jarvis, R.: Point signature: a new representation for 3d object recognition. In: Int. Journal of Computer Vision. Volume 25. (1997) 63–85
9. Chua, C.S., Han, F., Ho, Y.K.: 3d human face recognition using point signature. In: Proceedings of the 4th Int. Automatic Face and Gesture Recognition Conf. (2000) 233–238
10. Wang, Y., Chua, C.S., Ho, Y.K.: Facial feature detection and face recognition from 2d and 3d images. In: Pattern Recognition Letters. Volume 23. (2002) 1191–1202
11. Besl, P., Jain, R.: Invariant surface characteristics for 3d object recognition in range images. Computer Vision, Graphics, and Image Processing **33**(1) (1986) 33–80
12. Salomon, D.: Data compression: the complete reference. Springer-Verlag New York Inc (2004)
13. Martin, C., Werner, U., Gross, H.M.: A real-time facial expression recognition system based on active appearance models using gray images and edge images. In: Int. Conf. on Automatic Face and Gesture Recognition. (2008)
14. Viola, P., Jones, M.: Robust real-time face detection. Int. Journal of Computer Vision **57**(2) (2004) 137–154
15. Haase, D., Denzler, J.: Anatomical landmark tracking for the analysis of animal locomotion in x-ray videos using active appearance models. In: Image Analysis. Volume 6688 of Lecture Notes in Computer Science. (2011) 604–615
16. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: Int. Conf. for Computer Vision and Pattern Recognition. (2012) 2879–2886
17. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2000)
18. Khoshelham, K.: Accuracy analysis of kinect depth data. In: ISPRS Workshop Laser Scanning. Volume 38. (2011)
19. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at urlhttp://www.csie.ntu.edu.tw/ cjlin/libsvm.
20. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification. TR available at http://www. csie. ntu. edu. tw/ cjlin/papers/guide/guide. pdf (2009)
21. Webb, A., Copsey, K., Cawley, G.: Statistical pattern recognition. Wiley (2011)