Silver standards obtained from Fourier-based texture synthesis to evaluate segmentation procedures

Thomas M. Lehmann, Jörg Bredno, and Klaus Spitzer

Institute of Medical Informatics

Aachen University of Technology (RWTH), Aachen, Germany

ABSTRACT

Segmentation is fundamental for automated analysis of medical images. However, a unified approach for evaluation does not yet exist. Gold standards are often unapplicable because they require invasive preparations or tissue extraction. Empirical evaluations only reflect the conformity of segmentation with the subjective visual expectance of users, which is underlying inter- as well as intra-observer variabilities.

This paper presents a consistent approach to create synthetic but realistic images with a-priori known object boundaries (silver standards), which are suitable for optimization and evaluation of various segmentation algorithms. Rectangular example patches are collected for each tissue (interior, exterior, and a contour zone). Fourier amplitude and phase images are stored together with the mean gray value. For silver standard generation, a reference contour is either manually given or automatically extracted from real data applying the algorithm under evaluation. For each class of tissue, the amplitude of one patch is randomly combined with the perturbed phase of another. A randomly chosen mean from the same class is superimposed to the inverse Fourier transform.

Numerous silver standards are obtained from only a few texture patches of each tissue. Based on microscopy, CT, and functional MRI data, the applicability of silver standards is proven in two, three, and four dimensions. They are analysed with respect to systematic deviations. Minor deviations occur for two dimenional images while those for three or four dimensions are larger but still acceptable.

Keywords: Gold standard, Silver standard, Evaluation, Quality assessment, Segmentation, Texture synthesis

1. INTRODUCTION

Automatic segmentation is a fundamental step for computer-based analysis of medical images but until now, there exist no unified strategies for their evaluation.[1,2] On the one hand, the exact location of tissue or object boundaries in real data is unknown. Therefore, empirical evaluation only reflects the conformity of segmentation with subjective visual expectances of users, which are underlying both, inter- and intra-observer variability. On the other hand, evaluations that are based on synthetic data can rather seldom be generalized.[3] Furthermore, synthetic images often apply the same restrictive assumption on image properties that are used by the segmentation itself. Therefore, those evaluations are invalid.[4] Since comparable validations are missing, it is nearly impossible to choose the best algorithm for a specific segmentation task as well as its optimal set of parameters.[5] Therefore, automated segmentations of medical images still are rarely used in clinical routine.

Most methods that have been proposed for evaluation of image segmentations can be classified into three groups, the analytical method, the empirical goodness, and the empirical discrepancy.[1] The first class of methods addresses the algorithm only by theory and does not require its implementation. Hence, analytical analysis is not suitable for detailed evaluations as required in medical applications. Approaches from the second class are based on some desirable properties ("goodness") of segmented images, often established according to human intuition. However, a reference image is not applied. The third class of evaluation methods relies on distance measures to a specific reference, which is usually called the *gold standard*.[1]

To reflect the high variability of medical images, a steady evaluation of segmentation procedures must rely on large sets of gold standard images. So far, a comprehensive collection of realistic image data together with confirmed

Send correspondence to Thomas Lehmann, Institute of Medical Informatics, RWTH Aachen, D–52057 Aachen, Germany, E–mail: lehmann@computer.org

Medical Imaging 2001: Image Processing, Milan Sonka, Kenneth M. Hanson, Editors,
Proceedings of SPIE Vol. 4322 (2001) © 2001 SPIE · 1605-7422/01/$15.00

segmentations is missing.[2] Efforts to create such a collection by means of the visible human data set are unrealistic as the quality of these images is not reached in clinical routine. Furthermore, it is doubtful whether a manual segmentation of cryosections[6] indeed is a gold standard because the main disadvantage of real images has not changed: The ground truth is unknown.[7] Therefore, in his keynote speech at SPIE's Symposium on Medical Imaging 2000, ROBERT M. HARALICK termed such kind of *gold* standards to be from *plastic*.

The importance of a ground truth for the evaluation of image processing procedures was also emphasized by NGUYEN & ZIOU.[8] Evaluation without a ground truth is termed parameter-free.[9] It is most questionable whether parameter-free evaluation sufficiently reflects the complex process of object delineation in medical images. Another category of evaluation techniques distinguishes contextual and non-contextual approaches.[8] Non-contextual methods evaluate a segmention algorithm by tests involving images with adjustable properties or systematic changes of parameters.[3] However, the evaluation is directly related to the segmentation, while contextual methods evaluate the suitabilty of segmentation methods only for a certain application by means of the segmentation-based parameters subsequently determined by the application.

The aim of this paper is to give a method and means for contextual as well as non-contextual optimization and evaluation of segmentation algorithms using images with a-priori known ground truth. Based on Fourier texture properties, we present a consistent approach to synthezise realistic images, which are suitable as *silver standards* for the evaluation of various segmentation procedures. Their usefulness is proven for a balloon-based segmentation of two-, three-, and four-dimensional images from various modalities.

2. METHODS

Silver standard images are created as a stochastic combination of realistic contours and textures representing the appearance of tissue imaged by any modality. Segmentation is considered as a mean to localize and delineate a certain contour in an image at which two types of tissue are in contact. Assuming closed objects, the silver standard images must produce a realistic interior and exterior, each with a texture that is characteristic for the tissue type. It depends on the imaged tissues, the selected modality, and the applied segmentation method whether the contour zone in between is described by a third texture.

2.1. Collection of Texture Samples

A simple graphical user interface (GUI) is used to extract rectangular samples of tissue. All texture samples are represented by a Fourier description. Since the fast Fourier transform is applied, the GUI is designed to ensure that the side length of all patches is a power of two. For a single type of tissue, characteristic examples are collected from different images. All kinds of artefacts, which are usually connected to this tissue when captured by this modality, are included. Note that this appropriate inclusion of artefacts is of major importance for reliable evaluations and required before a method is applicable to clinical routine.[7]

2.1.1. Interior and Exterior

For both interior and exterior textures, each exemplary patch is extracted from the image, mirrored, and duplicated in all directions giving the $C^{(0)}$-continuous two dimensional discrete signal $t(x,y)$. Then, $t(x,y)$ is normalized to mean zero by subtracting its mean μ_t and the resulting signal $t_0(x,y)$ is Fourier-transformed

$$t(x,y) - \mu_t = t_0(x,y) \;\circ\!\!-\!\!\bullet\; T_0(u,v) = r_t(u,v) \cdot \exp\big(-j\varphi_t(u,v)\big) \tag{1}$$

For each sample t, the amplidute and phase, $r_t(u,v)$ and $\varphi_t(u,v)$, respectively, as well as its mean μ_t are stored in a database together with the filename of the original image and the position and size of the extracted example patch (Fig. 1).

2.1.2. Contour Zone

The extraction of exemplary textures along the contour is based on either a manual sketch that was drawn by means of the GUI or an automatic segmentation that was accepted by subjective visual inspection. In a polygonal representation, the contour consist of I vertices v_i, which are connected by straight edges e_i that join v_i and v_{i+1}. Here, we assume a closed contour with $v_{i+I} \equiv v_i$.

Texture samples of the contour zone must results in a rectangular band of width and length being again a power of two. For each edge e_i, this requires the transform of the trapezoid $\square ABCD$ into a rectangle (Fig. 2), which is

part of the linearized contour zone $z(x,y)$. The coordinates of the trapezoid corners A, B, C, and D are determined to lie on the intersection of parallels to the edges e_i. These parallels run in distance h_{in} and h_{out} on the inside and outside of the contour, respectively. If $\vec{n}_i$ denotes the unity normal vector of e_i, the position of A is given by

$$\vec{A} = \vec{v}_i + \frac{h_{\text{out}}}{1 + \vec{n}_i \cdot \vec{n}_{i-1}} \left(\vec{n}_i + \vec{n}_{i-1}\right) \tag{2}$$

Expressions are similar for the corner positions $\vec{B}$, $\vec{C}$, and $\vec{D}$. The extracted image values are successively copied from the image $g(x,y)$ into the linearized contour zone $z'(x,y)$ by transforming all egdes

$$z'(x + \Delta x, y) = g\left(\frac{y}{h_{\text{in}} + h_{\text{out}}}\left[\vec{D} + \frac{x}{|\vec{v}_i - \vec{v}_{i+1}|}(\vec{C} - \vec{D})\right] + \frac{h_{\text{in}} + h_{\text{out}} - y}{h_{\text{in}} + h_{\text{out}}}\left[\vec{A} + \frac{x}{|\vec{v}_i - \vec{v}_{i+1}|}(\vec{B} - \vec{A})\right]\right) \tag{3}$$

with $\Delta x = \sum_{j=0}^{i} |\vec{v}_{j-1} - \vec{v}_j|$, $x \in [0, |\vec{v}_i - \vec{v}_{i+1}|]$, and $y \in [0, h_{\text{in}} + h_{\text{out}}]$. This transform reads image values from non-integer coordinates using linear interpolation.[10] Resulting in $z(x,y)$, the stripe z' is subsequently mirrored and duplicated along it's x-axis, which corresponds to the direction along the contour. With respect to (1), $z(x,y)$ is normalized to mean zero and Fourier-transformed. Again, the magnitude, phase, and mean, $r_z(u,v)$, $\varphi_z(u,v)$, and μ_z, respectively, are stored in the database.

2.2. Two-Dimensional Silver Standards

To ensure highest flexibility, the creation of silver standard images $s(x,y)$ is done by combining a map of textures $m_t(x,y)$ and a map of gray levels $m_g(x,y)$. In particular, $s(x,y)$ is defined as the sum of texture and gray value maps

$$s(x,y) = \lceil m_t(x,y) + m_g(x,y) \rceil \tag{4}$$

where $\lceil \cdot \rceil$ denotes the clipping to the value range of the output image. Additionally, a casting to the desired data type is often required.

2.2.1. Segmentation Map

Both maps are based on a binary segmentation map $m_s(x,y)$, which is derived from the contour of the desired segment. In general, any closed contour can be used to create a silver standard image. Again, such a contour can be drawn from a user sketch or from automatic segmentation of real data. Based on the polygonal representation of the contour, m_s labels interior and exterior by one and zero, respectively.

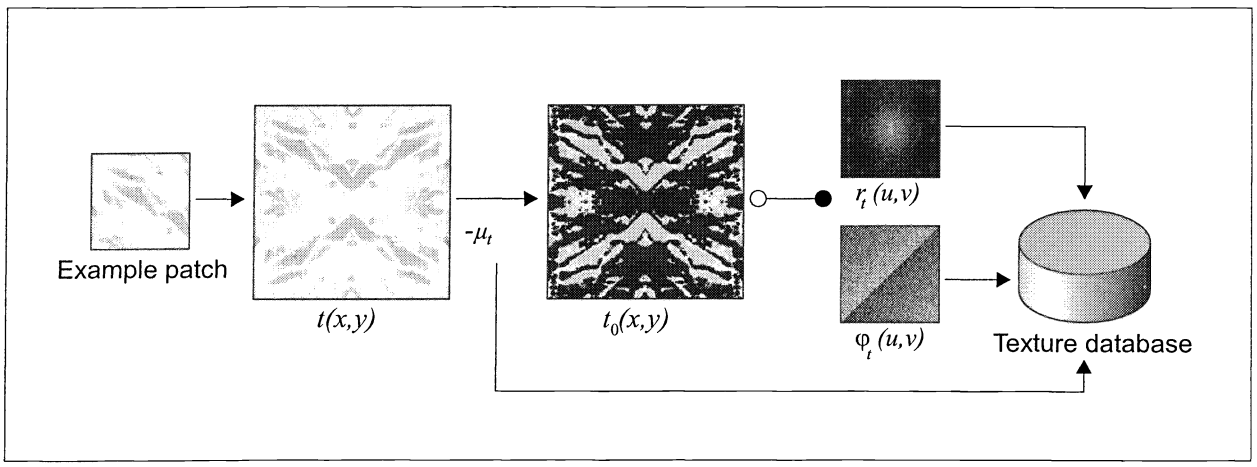

Figure 1. Fourier-based texture decomposition.

2.2.2. Texture Map

For each of inside and outside, a synthetic texture patch $\tilde{t}(x,y)$ is created according to the selected type of tissue. A randomly chosen amplitude $\hat{r}_t(u,v)$ is combined with a randomly chosen phase $\hat{\varphi}_t(u,v)$ that is additionaly perturbed by additive white noise $n_{0,\sigma}(u,v)$ with $\sigma = \frac{1}{10}\pi$. The inverse Fourier transform yields

$$\tilde{t}(x,y) \;\bullet\!\!-\!\!\circ\; \hat{r}_t(u,v) \cdot \exp\Big(- j\big(\hat{\varphi}_t(u,v) + n_{0,\sigma}(u,v)\big)\Big) \tag{5}$$

that is asymmetric but can be tiled $C^{(0)}$-continuously. According to the labels from $m_s(x,y)$, a texture map $m_t(x,y)$ is filled with the appropriate textures $\tilde{t}$ for inside and outside, $\tilde{t}_{\text{in}}(x,y)$ and $\tilde{t}_{\text{out}}(x,y)$, respectively. Tiling is used for regions are larger than the texture patches (Fig. 3a).

By option, a synthetic contour zone is similarly textured combining randomly chosen magnitude and phase, $\hat{r}_z(u,v)$ and $\hat{\varphi}_z(u,v)$, respectively. However, pertubation of the phase is avoided in order to preserve distinct structures of the contour zone. The resulting synthetic texture

$$\tilde{z}(x,y) \;\bullet\!\!-\!\!\circ\; \hat{r}_z(u,v) \cdot \exp\big(- j\hat{\varphi}_z(u,v)\big) \tag{6}$$

is transformed into the texture map by inversion of (3). The lenght of the contour might require its $C^{(0)}$-continouous tiling along the x-axis (Fig. 3c).

2.2.3. Gray Value Map

For the interior and exterior of the contour, mean gray values μ_t are randomly read from the database and denoted $\hat{\mu}_{\text{in}}$ and $\hat{\mu}_{\text{out}}$, respectively. They are used to fill the gray value map $m_g(x,y)$ according to the labels provided by $m_s(x,y)$. Along the contour, this results in artificially high gradients. Therefore, a blurring partial volume effect, which is inherent to most imaging devices, is simulated by weakening of this gradient. A linear transition zone of width $2w$ is created in the vicinity of the contour. For each pixel in $m_g(x,y)$, its distance to the contour is determined. First, the nearest vertex v_i to (x,y) is determined in the cartesian space. Second, the two distances to the adjacent edges e_i and e_{i-1} are computed by decomposing the (x,y) into edge-base coordinates a_j and b_j, $j \in \{i, i-1\}$, which are given by

$$(x,y) = \vec{v}_j + a_j \cdot (\vec{v}_{j+1} - \vec{v}_j) + b_j \cdot \vec{n}_j \tag{7}$$

The distances $|b_j|$ are only valid if $a_j \in [0,1]$. The resulting distance is chosen as the minimum of the cartesian distance $|\vec{v}_i - (x,y)^T|$ and the valid distances $|b_j|$. It is used to calculate the mean gray value in the vicinity of the contour (Fig. 3b) resulting in a linear transition from $\hat{\mu}_{\text{in}}$ to $\hat{\mu}_{\text{out}}$.

If the contour zone is assessed by its own texture referring to (6), a mean value is randomly read from the database and accordingly denoted $\hat{\mu}_{\text{zone}}$. Hence, a linear transition of width $2w$ is created twice in the gray value map $m_g(x,y)$. One transition is defined from the area of $\hat{\mu}_{\text{in}}$ to that of $\hat{\mu}_{\text{zone}}$ and the other from $\hat{\mu}_{\text{zone}}$ to $\hat{\mu}_{\text{out}}$. However, the width of the textured contour zone should be larger that the width of the linear transition: $h_{\text{in}} + h_{\text{out}} > 2w$. This is exemplified in Figure 3d.

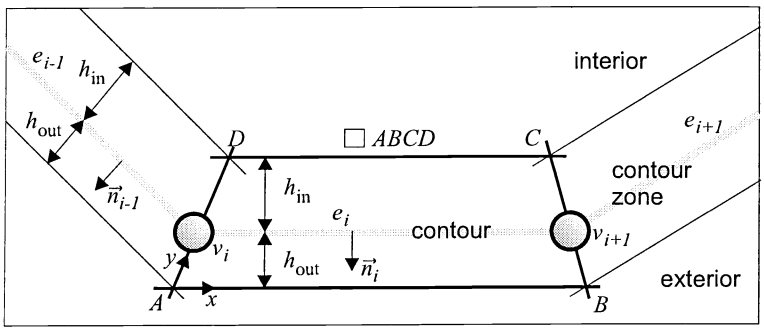

Figure 2. Construction of the contour zone.

2.2.4. Synthesis of Silver Standards

Disregarding the pertubation of the phase patches, a collection of N_{in}, N_{out}, and N_{zone} sample textures from interior, exterior, and the contour zone, respectively, can be used to create

$$N = N_{\text{in}}^3 \cdot N_{\text{out}}^3 \cdot N_{\text{zone}}^3 \tag{8}$$

silver standard images for each exemplary contour. Therefore, as much as 512 combinations result from only two samples of each texture.

2.3. Three- and Four-Dimensional Silver Standards

Modalities providing image material of higher dimension usually provide the data in either two-dimensional stacks, e.g. computed tomography (CT), or sequences of still images, e.g. video frames. Therefore, the generation of silver standard images is straightforwardly extended to image stacks, image sequences, or four-dimensional sequences of image stacks.

2.3.1. Collection of Texture Samples

Texture samples are collected in two-dimensional sub-images as described in Section 2.1.1 and accordingly reconstructed afterwards. For three- and four-dimensional images, a certain contour zone is not provided because a reliable solution for linear parametrizations of surfaces of arbitrary topology does not exist in those dimensions. For color images, the texture extraction is performed indepentently for all three different color channels. Here, the color model in use is expected to allow the computation of a mean value, which does not hold for cyclic channels such as hue in the hue, saturation, and value (HSV) color space. However, the red, green, and blue (RGB) color space agrees with this assumption.

2.3.2. Segmentation Map

Again, the binary segmentation map $m_s(\vec{x})$ contains the value zero and one for all image elements $\vec{x}$ (pixels, voxels, or stixels) outside and inside the desired object, respectively. Although in general, user sketches might be used as source of $m_s(\vec{x})$, this procedure requires intensive user interaction in three and four dimensions. Therefore, visually inspected automatic segmentations are applied instead of. Note that all kind of contour representations can be easily converted into $m_s(\vec{x})$.

2.3.3. Contour Map

As mentioned above, a certain contour zone is not available for multi-dimensional silver standards. However, a binary contour image $m_c(\vec{x})$ is created that contains the surface of the object. The proper concept of neighborhood depends on the dimensionality of data and the properties of the modality.

- The 8-neighborhood is applied if the distance between two silces in a volumetric data set is so large that the contour of tangentially cut objects is not visible, or if the data contains an image sequence without motion artefacts.

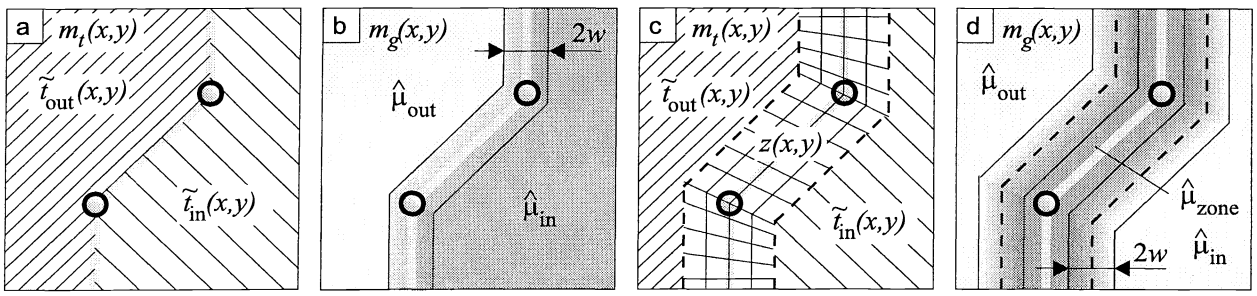

Figure 3. Texture- and gray value maps with or without a certain contour zone are shown in (c), (d) or (a), (b), respectively.

- The 26-neighborhood is applied to volume data with approximate cartesian image space, i.e. high inter-slice-resolution, as well as to four-dimensional sequences of image stacks without motion artefacts, where each stack of images is processed separately.
- The 80-neighborhood is applied to sequences of nearly-cartesian volume data containing motion artefacts, because this requires to process coherently in all of the four dimensions.

An image element in $m_c(\vec{x})$ is set to one if the local vicinity of $\vec{x}$ in $m_s(\vec{x})$ concurrently contains zeros and ones. According to the chosen neighborhood, this vicinity is of dimension 3×3 in one slice, $3 \times 3 \times 3$ in volumetric image data, and $3 \times 3 \times 3 \times 3$ in a coherent four-dimensional image space.

2.3.4. Distance Map

An incomplete distance map $m_d(\vec{x})$ marks all voxels or stixels that lay within a Manhattan distance of w to the contour in $m_c(\vec{x})$. In case of the 8-neigborhood, m_d is separately created for each slice. For 26-neighborhood-coded volumetric surfaces, m_d is created in the three-dimensional image space. A four-dimensional space is used concerning contours in the 80-neighborhood.

2.3.5. Gray Value Map

For the creation of the gray value map $m_g(\vec{x})$, mean gray values $\hat{\mu}_{\text{in}}$ and $\hat{\mu}_{\text{out}}$ are randomly determined as described in Section 2.2.3. Whether each image element $\vec{x}$ belongs to the inside or outside of the desired object is marked in $m_s(\vec{x})$. In addition, the distance to the contour is read from $m_d(\vec{x})$. If the image element is not in the vicinity of the contour, $m_g(\vec{x})$ is set to either $\hat{\mu}_{\text{in}}$ or $\hat{\mu}_{\text{out}}$. In the vicinity of the contour, a linear transition from $\hat{\mu}_{\text{in}}$ to $\hat{\mu}_{\text{out}}$ is created in $m_g(\vec{x})$.

2.3.6. Texture Map

The texture map $m_t(\vec{x})$, that is related to data of three and four dimensions, is created with respect to the desired modality. For modalities resulting in a reproducable appearance of tissue for each slice, only one synthetic texture is created for each of interior and exterior. This particularly holds for CT were fixed units of measurements are set according to standardized phantoms. Here, texture samples only represent the variability of tissue. For modalities where the above standardization is not possible, e.g. in case of magnetic resonance imaging (MRI), a synthetic texture is created for each slice. Furthermore, a random selection of means is performed for each slice. Now, texture samples represent both, the variability of tissue and the variability of appearance induced by the modality. Hence, a larger number of samples is needed. The final silver standard is defined by (4), again including gray value clipping and datatype casting.

2.4. Quantitative Assessment of Segmentation Quality

Silver standards $s(\vec{x})$ are tantamount to synthetic images with a-priori known ground truth of segmentation. They are applied for the evaluation of segmentation algorithms. Hence, quantitative similarity measures are required, which reflect the distance or agreement of a current result $r(\vec{x})$ of segmenting the silver standard image $s(\vec{x})$. In contrast to other applications,[11] suitable similarity measures are simultaniously defined in two, three, and four dimensions. Since the a-priori known binary segmentation map $m_s(\vec{x})$ was used in (4) to create the silver standard image s, advantageous measures are based on binary images.

2.4.1. Overlap Measure

Let $m_r(\vec{x})$ denote the binarization of $r(\vec{x})$. The overlap measure $O \in [0,1]$ is defined as the number of pixels in the intersection of all set pixels in $m_r(\vec{x})$ and $m_s(\vec{x})$ normalized by the number of pixels in their union. The overlap O cumulatively measures the agreement of the segmentation with the silver standard

$$O = \frac{\left| m_r(\vec{x}) \cap m_s(\vec{x}) \right|}{\left| m_r(\vec{x}) \cup m_s(\vec{x}) \right|} = \frac{\sum_{\vec{x}} \begin{cases} 1 & \forall \quad \vec{x} \mid m_r(\vec{x}) = 1 \wedge m_s(\vec{x}) = 1 \\ 0 & \text{else} \end{cases}}{\sum_{\vec{x}} \begin{cases} 1 & \forall \quad \vec{x} \mid m_r(\vec{x}) = 1 \vee m_s(\vec{x}) = 1 \\ 0 & \text{else} \end{cases}} \tag{9}$$

2.4.2. Asymmetric Distance Measures

For the computation of local geometric distance measures, a geometric representation of the segmentation is required. We assume a simplex-mesh that is a set V of I vertices v_i at position $\vec{v}_i$. Affine simplices connect these vertices according to the dimension of the image space. Such a representation is obtained if active contour models are used for segmentation.[12] Of course, polygonal approximation[13] or triangulation[14] are also applicable to region- or volume-orientated segmentation and the silver standard contour. However, these approximative methods yield artifacts that are not separable from errors induced by the segmentation method. Therefore, secondary simplex-mesh representations should not be used for quantitative distance measures.

Based on the complete Manhattan distance transform of $m_s(\vec{x})$, which is denoted $m_d^c(\vec{x})$ and obtained similar to $m_d(\vec{x})$ as described in Section 2.3, the Manhattan distance to the silver standard contour is read for each $\vec{v}_i$, which is supposed to lay on this contour if the segmentation method is accurate. Hence, asymmetric mean Manhattan and Hausdorff Manhatten distances are given by

$$\bar{d}_M^a = \frac{1}{I}\sum_{i=1}^{I} m_d^c(\vec{v}_i) \tag{10}$$

and

$$H_M^a = \max_{i=1}^{I}\left(m_d^c(\vec{v}_i)\right), \tag{11}$$

respectively.

2.4.3. Symmetric Distance Measures

If both, $m_r(\vec{x})$ and $m_s(\vec{x})$ are originally given as simplex-meshes, symmetric cartesian distances can be quantified. Symmetry is obtained because all vertice positions, either determined by the segmentation or the silver standard, are compared to each other. Let W denote the contour of the silver standard with J vertices w_j at positions $\vec{w}_j$. The computation of the distance $d(\vec{x}, C)$ from an image position $\vec{x}$ and a contour C consists of the following steps:

- The vertex v_i in C with smallest cartesian distance to $\vec{x}$ is determined and added to a list of distances.
- For all simplices e_j that contain v_i, the projection of $\vec{x}$ along the simplex' normal vector $\vec{n}_j$ onto e_j is computed. If the base of this projection lays within e_j, the length of the projection vector is added to the above list of distances.
- From the list of distances, the smallest value is returned as $d(\vec{x}, C)$.

Using this computation of $d(\vec{x}, C)$, the symmetric mean cartesian distance $\bar{d}_c^s$ and the symmetric Hausdorff cartesian distance H_c^s is defined by

$$\bar{d}_c^s = \frac{1}{I+J}\left(\sum_{i=1}^{I} d(\vec{v}_i, W) + \sum_{j=1}^{J} d(\vec{w}_j, V)\right) \tag{12}$$

and

$$H_c^s = \max\left(\max_{i=1}^{I} d(\vec{v}_i, W), \max_{j=1}^{J} d(\vec{w}_j, V)\right), \tag{13}$$

respectively. While (10) and (11) are only effected by deviations from the detected to the original contour, (12) and (13) also assess whether the detected contour follows all details of the original.

3. APPLICATIONS

Silver standards can be applied to directly evaluate the non-contextual ability of a certain segmentation procedure according to a specific task or to contextually evaluate a system for quantification of medical images that is based on a certain segmentation procedure. Concerning contextual evaluation, the ground truth results from quantitative measurements of the silver standard contour. We have applied silver standard images to evaluate a finite element balloon model for segmentations in two, three, and four dimensions.[12] Non-contextual as well as contextual segmentation tasks were chosen for this evaluation.

Table 1. Application to immunohistochemically processed micrographs.

silver	#	$\bar{d}_c^s$ in pixel		H_c^s in pixel		O in percent	
contour		μ	σ	μ	σ	μ	σ
A	48	1.44	0.64	4.99	1.88	97.55	1.07
B	50	1.80	0.83	14.69	9.50	97.59	1.17
C	49	1.71	0.75	6.34	2.35	97.66	1.03
D	50	1.71	0.78	6.04	3.76	97.39	1.21
$\sum$	197	1.67	0.77	8.06	6.63	97.55	1.13

3.1. Two-Dimensional Silver Standards

The quantification of axo-somatic boutons at motoneuron cell-surface membranes is based on immunohistochemical staining of thin slice specimens of the spinal cord of adult rats (Fig. 4a).[15] Therefore, this task was selected to exemplify the application of two-dimensional silver standard images. A reliable segmentation of the cell membrane is required to extract precisely and analyze exactly the synaptic profiles (Fig. 4b). The micrographs are 512×480 pixels in size. From 22 micrographs, 50 exemplary texture patches of size 128×128 for each of the intracellular and extracellular space were collected. Automated segmentation of 19 cells was used to extract 50 texture patches from the cell membrane region to describe the contour zone. The linearized patches were of height $h_{\text{in}} + h_{\text{out}} = 8$ and width 256 pixels. For each of four exemplary contours A–D, 50 silver standard images were created and segmented.

Based on real data, the failure rate of the balloon-based segmentation method is about 7%.[15] In this application, the considered algorithm completely failed on three silver standards (2%). This validates a sufficient variability of the silver standards. The distance and agreement measures $\bar{d}_c^s$, H_c^s, and O were computed for the remaining 197 silver standards in which the segmentation did not stop prematurely (Tab. 1). The large overlap $O = 97.55\%$ shows that the localization of the cell membrane is easily resolved by the balloon-based segmentation method. However, the delineation usually is not sufficiently accurate in all parts of the detected contour. This is indicated by the average Hausdorff distance of $H_c^s = 8.06$ pixels, whereas the mean distance of $\bar{d}_c^s = 1.67$ pixels shows that most parts of the contour are localized correctly. Note that the applied segmentation procedure covers local detection errors by automatically assigning local confidence measures, which are used as weights in the quantification of staining along the cell membrane.[16] Such confidences are also demanded by HAYNOR.[7]

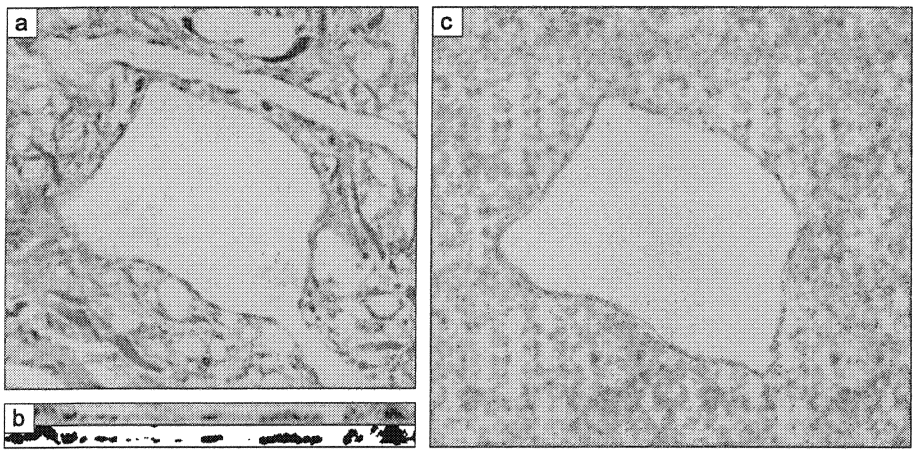

Figure 4. Immunohistochemistally stained motoneuron (a), linearized and binarized region of interest (b), and a silver standard for this image class (c).

Table 2. Application to CTs of vertebrae and intervertebral discs.

silver contour	#	$\bar{d}_c^s$ in pixel		H_c^s in pixel		O in percent	
		μ	σ	μ	σ	μ	σ
A	10	2.64	0.09	18.51	7.62	87.23	0.31
B	10	3.23	0.08	16.38	1.31	84.16	0.31
C	10	2.83	0.09	29.89	11.34	86.59	0.38
$\sum$	30	2.90	0.26	21.59	9.90	85.99	1.36

3.2. Three-Dimensional Silver Standards

CT slices of the spinal cord are routinely acquired for surgery of a herniated disk. Since treatment planning requires precise segmentation of bony structures (Fig. 5a), this task was selected to exemplify the application of silver standard images in three dimensions. Texture patches were extracted from slices of three-dimensional volume data sets A–C showing the vertebrae and the intervertebral discs (Fig. 5b). The slices are 512×512 pixels in size, and 29, 36, or 38 slices were available. For the outside region, patches of 128×128 pixels show a combination of fatty and watery tissue. Taking into account that especially for elderly patients, the intervertebral disc has a similar appearence to bony structures, interior example patches of 32×32 pixel are extracted either from compact bone or the intervertebral disc. Therefore, the silver standard slices show neither spongiouos structures nor bone marrow (Fig. 5c).

According to the reproducability of the CT imaging technique, three-dimensional silver standards are created using one choice of $\hat{\mu}_{\text{in}}$, $\hat{\mu}_{\text{out}}$, and synthetic textures for each silver standard image. Since the slice distance is four to five times higher than the intra-slice pixel spacing, $m_c(\vec{x})$ and $m_d(\vec{x})$ are created in layers using the 8-neighborhood. For each of the three data sets, ten silver standards were created and segmented with the balloon-based method. The contour is initialized at the border of the image space and shrinks towards the structures of interest. Topological changes are automatically performed to represent holes induced by spinal aperture, joint openings, and the root canal. In order to simulate a partial volume effect, $w = 2$ was selected.

The obtained distance and agreement measures are summerized in Table 2. A large mean Hausdorff distance of $H_c^s = 21.59$ is shown. This indicates that in almost all data sets some major segmentation errors occur locally. Consequently, routine application of this method requires visual inspection by clinicians. The mean distance $\bar{d}_c^s = 2.90$ reflects the problem of automatic delineation of CTs in pixel accuracy.

3.3. Four-Dimensional Silver Standards

A temporal sequence of volumetric MRI scans of the beating heart was used as an exemplary four-dimensional data set. The segmentation of the left ventricular endocard is used to quantify the overall blood flow as well as local wall movements. The balloon-based segmentation procedure allows coherent four-dimensional segmentation, where the

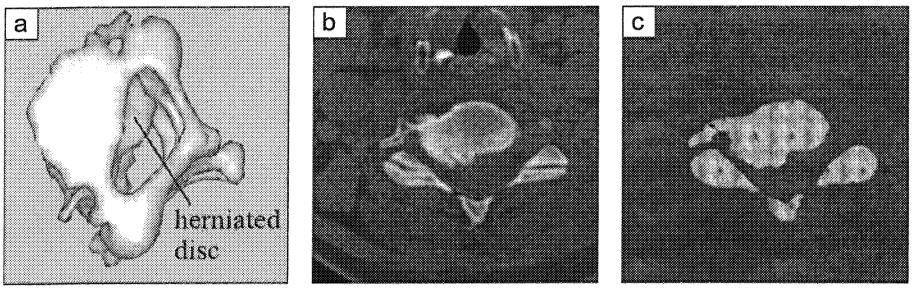

Figure 5. Segmentation of the spinal chord (a), slice of the CT data set (b), and a silver standard for this image class (c).

Table 3. Application to temporal MRIs of the left ventricle.

silver contour	#	$\bar{d}_c^s$ in pixel		H_c^s in pixel		O in percent	
		μ	σ	μ	σ	μ	σ
A	10	1.16	0.03	6.29	0.29	79.98	0.33

correspondence of vertices in different volume data sets is known over the entire cardiac cycle (Fig. 6a, midsystolic view). A data set of 256×256 pixels in 20 slices and 11 points of time over one cardiac cycle was manually clipped (Fig. 6b). Based on automatic segmentation of the left ventricle, ten silver standards of myocard and blood cavity tissue were created (Fig. 6c). For the myocard, ten texture samples of 8×8 pixels in size were extracted from different slices (ventricular level down to the apex) and points of time (enddiastolic, midsystolic, endsystolic, middiastolic). Another ten texture samples show the cavity in patches of 16×16 pixels. Since the baseline of gray values and hence, the appearance of tissue is difficult to reproduce in MRIs, $\hat{\mu}_{\text{in}}$, $\hat{\mu}_{\text{out}}$, and different texture patches were used for each slice of the silver standard.

The segmentation of ten silver standards shows reproducible results (Tab. 3). The Hausdorff distance $H_c^s = 6.29$ is mainly caused by papillary muscles near to the myocard that may or may not be segmented as part of the cavity. A Hausdorff distance $H_c^s = 6$ is already measured if at least one vertex out of more than 1200 vertices lays three pixels apart from the silver standard contour in each dimension with the distance of two slices already estimated to four pixels. Note that the overlap measure cannot be compared between image material of different dimensions. With a mean distance similar to that of the micrographs (Sec. 3.1), the overlap measure is significantly reduced. This is caused by the varying number of pixels that are excluded from or included into the segmentation when parts of the contour are shifted for a fixed distance in image material of different dimensions.

3.4. Colored Silver Standards

The exemplary contextual evaluation of an automated measurement system was based on color photographs of skin melanomas. The photos were taken under standardized lightning conditions and subsequent histologic findings secured the diagnosis. For each of the three color channels R, G, and B, five texture patches of 128×128 pixels in size were collected from a data set of seven images. A polygonal approximation of one manual segmentation was used as input contour. Resulting in an original size of 48748 pixes, this approximation was required to perform transforms to the contour. Then, 100 silver standard images were created. For each image, the contour was randomly shifted, rotated, and scaled. The offsets in x- and y-direction, angles of rotations and scales were uniformly distributed in $[-10, 10]$ pixels, $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and $[0.7, 1.4]$, respectively. The size of each contour was compared to the size of the segmentation of the silver standard. Preliminary convergence of the balloon was obtained in only one case, which was excluded from further analysis. A correlation coefficient of $\rho = 0.998$ showed a strong linear dependance of the true and the measured values. The slope was calculated to $m = 0.986$ and the intercept to $b = 128.7$. Including the incorrect segmentation, which was easily detected by visual inspection, the correlation coefficient is decreased down to $\rho = 0.950$.

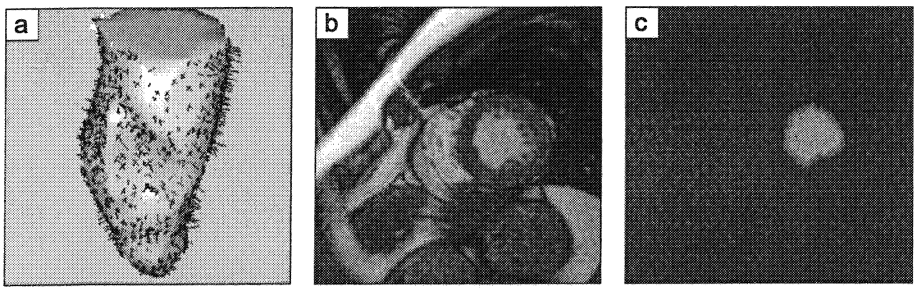

Figure 6. Midsystolic view of the left ventricular cavity with local wall movement (a), slice from the MR data set (b), and a silver standard for this image class showing only the cavity and the myocard (c).

Table 4. Systematic deviation induced by replacement of the ground truth with a segmented contour.

data	${}^1\bar{d}_c^s$	${}^2\bar{d}_c^s$	Δ_d	${}^1O$	${}^2O$	Δ_O
2-D	1.67	1.50	-0.17	97.55	97.80	0.27 %
3-D	2.90	2.17	-0.73	85.99	90.00	4.66 %
4-D	1.16	0.42	-0.74	79.98	96.09	20.12 %

4. RESULTS

The presented methodology for the creation of silver standards is suitable for contextual as well as non-contextual evaluation whenever automated segmentation procedures are included to medical image analysis systems. The modular generation of silver standards easily allows flexible adaptation. For example, silver standards can exclude or include a certain texture for the contour zone when used for evaluation of region-based or contour-based segmentation procedures, respectively. Anyway, only a small number of texture pathes is required for each type of tissue to create a plenty of silver standard images. In addition, our method is suitable for two, three, and four-dimensional data.

To test whether systematic deviations are induced in the creation of silver standards, the symmetric mean cartesian distance $\bar{h}_c^s$ as well as the overlap measure O were analysed more detailed. Note that caused by the maximum operator in (11) and (13), the Hausdorff distances are less suitable for this analysis. For each test, a segmentation obtained by an arbitrarily chosen silver standard image was considered as ground truth for all others. The superscripts "1" and "2" denote whether the measure is based on the original or the exchanged ground truth, respectively. Concerning the two-dimensional example, the distance was only marginally decreased and the overlap was raised by only 0.25 points (Tab. 4). In case of three-dimensional silver standards, an increased reduction was shown for $\bar{d}_c^s$ and O was enlarged by 4.0 points. Comparing nine arbitrarily chosen four-dimensional silver standard segmentations to the tenth, $\bar{d}_c^s$ was essentially decreased while O was substancially increased by 16.11 points. Therefore, it is concluded that two-dimensional silver standards don't show systematic deviations as compared to the original contours. The creation of three- and four-dimensional silver standards results in only minor systematic deviations, which have to be considered whenever silver standards are applied to evaluate segmentation techniques.

5. DISCUSSION

Gold standards are found seldom in medical image analysis as they often require invasive preparations or subsequent tissue extraction.[1,17] In this paper, a versatile but simple method to create silver standards for the evaluation of numerous segmentation procedures was introduced and its usefulness was successfully demonstrated. Regarding the manifold quantification tasks in medical imaging that require a precise localization and delineation of objects, our approach is expected to find a wide range of applications in medicine. In order to promote the use of automated segmentation in clinical routine, results of evaluation must be laid open to users[7] giving objective criteria for the optimization of the algorithm. The need for evaluation of well-known methods might interfere with the goal to develop novel methods. However, it is known from speech recognition that proliferous competitions require standardized evaluation.[2]

Our aim was to present a suitable method to create silver standard images that can be used to evaluate contextual as well as non-contextual segmentation tasks with realistic images of known ground truth. As shown by the non-contextual validation of an exemplary segmentation method, the silver standards reproduce original textures of medical images and reflect or even surpass the variability of image data acquired in routine applications. This is achieved by the combination of features from only a few example patches. The examplary contextual validation used only a simple feature, namely the size of a region. The ability to control the parameters of the reference contour here enables us to test not only the reproducibility of single values, but also that of distributions of values. Although the covariance of features[3] was not estimated in our particular example, the creation of silver standards does not limit the features that are changed for a contour in order to simulate multivariate populations.

Fourier analysis has lots of applications for texture description as well as generation. Based on the detection of leukocytes in intravital microscopies, EGMONT-PETERSEN et al. have shown the superiority of artificial textures

generated by Fourier techiques over natural example patches when applied to the training of a neural network.[18] In our approach, the Fourier techique was improved and generalized to be suitable for any modality in any dimension. Consequently, our silver standards cannot be applied to evaluate segmentation techniques that are based on texture frequency analyses because their creation incorporates characteristic Fourier representations. A different characteristic description for texture patches is needed to evaluate Fourier-based segmentation. As the example of bony structures in CT volume data has shown, further extensions of the method should cover the creation of a contour zone for three-dimensional images even though this involves the most difficult linearization of non-trivial surface topologies.

In a common viewpoint, it is stated that rigorous evaluation of medical image processing algorithms ultimately requires the use of patient data.[2] Note that our silver standard images are confirm with this position. However, we disagree with the conclusion, which is often drawn from this standpoint, that the use of patient data entails substancial work by a number of experts. Although our silver standards are based on patient data, essential information is extracted and used to create realistic synthetic images with a-priori known ground truth. Hence, this approach might close the gap between reliable evaluation and acceptable user interaction.

REFERENCES

1. Y.J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition* **29**(8), pp. 1335–1346, 1996.
2. J.C. Gee, "Performance evaluation of medical image processing algorithms," *Procs. SPIE* **3979**, pp. 19–27, 2000.
3. R.M. Haralick, "Validating image processing algorithms," *Procs. SPIE* **3979**, pp. 2–16, 2000.
4. V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. on Medical Imaging* **16**(5), pp. 642–652, 1997.
5. V. Chalana, J.A. Hodgdion, and D.R. Haynor, "Unified data structures in a software environment for medical image segmentation," *Procs. SPIE* **3338**, pp. 947–958, 1998.
6. R.A. Robb, "Virtual endoscopy: Development and evaluation using the Visible Human data sets," *Computerized Medical Imaging and Graphics* **24**(3), pp. 133–151, 2000.
7. D.R. Haynor, "Performance evaluation of image processing algorithms in medicine: A clinical perspective," *Procs. SPIE* **3979**, p. 18, 2000.
8. T.B. Nguyen and D. Ziou, "Contextual and non-contextual performance evaluation of edge detectors," *Pattern Recognition Letters* **21**, pp. 805–816, 2000.
9. M. Borsotti, C. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognition Letters* **19**(8), pp. 741–747, 1998.
10. T.M. Lehmann, C. Gönner, and K. Spitzer, "Survey: Interpolation methods in medical image processing," *IEEE Trans. Medical Imaging* **18**(11), pp. 1049–1075, 1999.
11. T.M. Lehmann, A. Sovakar, W. Schmitt, and R. Repges, "A comparison of mathematical similarity measures for digital subtraction radiography," *Computers in Biology and Medicine* **27**(2), pp. 151–167, 1997.
12. J. Bredno, T.M. Lehmann, and K. Spitzer, "A general finite element model for segmentation in 2, 3, and 4 dimensions," *Procs. SPIE* **3979**, pp. 1174–1184, 2000.
13. S.C. Huang and Y.N. Sun, "Polygonal approximation using genetic algorithms," *Pattern Recognition* **32**(8), pp. 1409–1420, 1999.
14. G. Barequet, D. Shapiro, and A. Tal, "Multilevel sensitive reconstruction of polyhedral surfaces from parallel slices," *Visual Computer* **16**(2), pp. 116–133, 2000.
15. T.M. Lehmann, J. Bredno, V. Metzler, G. Brook, and W. Nacimiento, "Computer-assisted quantification of axo-somatic boutons at the cell membrane of motoneurons," *IEEE Trans. on Biomedical Engeneering*, accepted for publication.
16. V. Metzler, J. Bredno, T.M. Lehmann, and K. Spitzer, "A deformable membrane for the segmentation of cytological samples," *Procs. SPIE* **3338**, pp. 1246–1257, 1998.
17. R.T. Constable, K.M. Rath, A.J. Sinusas, and J.C. Gore, "Development and evaluation of tracking algorithms for cardiac wall motion analysis using phase velocity MR imaging," *Magnetic Resonance in Medicine* **32**(1), pp. 33-42, 1994.
18. M. Egmont-Petersen, U. Schreiner, S.C. Tromp, T.M. Lehmann, D.W. Slaaf, and T. Arts, "Detection of leukocytes in contact with the vessel wall from in-vivo microscope recordings usind a neural network," *IEEE Trans. Biomedical Engineering* **47**(7), pp. 941–951, 2000.