From plastic to gold: A unified classification scheme for reference standards in medical image processing

Thomas M. Lehmann*

Department of Medical Image Processing
Institute of Medical Informatics, Medical Faculty
Aachen University of Technology (RWTH), Aachen, Germany

ABSTRACT

Reliable evaluation of medical image processing is of major importance for routine applications. Nonetheless, evaluation is often omitted or methodically defective when novel approaches or algorithms are introduced. Adopted from medical diagnosis, we define the following criteria to classify reference standards:

1. Reliance, if the generation or capturing of test images for evaluation follows an exactly determined and reproducible protocol.
2. Equivalence, if the image material or relationships considered within an algorithmic reference standard equal real-life data with respect to structure, noise, or other parameters of importance.
3. Independence, if any reference standard relies on a different procedure than that to be evaluated, or on other images or image modalities than that used routinely. This criterion bans the simultaneous use of one image for both, training and test phase.
4. Relevance, if the algorithm to be evaluated is self-reproducible. If random parameters or optimization strategies are applied, reliability of the algorithm must be shown before the reference standard is applied for evaluation.
5. Significance, if the number of reference standard images that are used for evaluation is sufficient large to enable statistically founded analysis.

We demand that a true gold standard must satisfy the Criteria 1 to 3. Any standard only satisfying two criteria, i.e., Criterion 1 and Criterion 2 or Criterion 1 and Criterion 3, is referred to as silver standard. Other standards are termed to be from plastic. Before exhaustive evaluation based on gold or silver standards is performed, its relevance must be shown (Criterion 4) and sufficient tests must be carried out to found statistical analysis (Criterion 5). In this paper, examples are given for each class of reference standards.

Keywords: Software Validation, Reference Standard, Gold Standard, Silver Standard, Evaluation, Quality Assessment, Image Processing, Research Design, Virtual Radiography

1. INTRODUCTION

In the past decades, modern imaging techniques have revolutionized the study of anatomy and physiology of man. Consequently, sophisticated computational methods for the extraction of salient information from medical image data are increasingly applied. Numerous algorithms have been developed especially for use in biomedicine and health care. However, such algorithms have mostly remained research tools and only a few of them have been transferred into clinical applications. It has been shown that a main reason for this poor track record is the lack of validation of this

* lehmann@computer.org; phone +49 241 80-88793; fax +49 241 80-82426; http://www.irma-project.org; Institute of Medical Informatics, Aachen University of Technology, Pauwelsstr. 30, D - 52057 Aachen, Germany.

Medical Imaging 2002: Image Processing, Milan Sonka, J. Michael Fitzpatrick,
Editors, Proceedings of SPIE Vol. 4684 (2002) © 2002 SPIE · 1605-7422/02/$15.00

methods.[3] In other words, reliable evaluation of medical image processing (such as registration, segmentation, classification, and quantitative measurements) is of major importance with respect to any routine application.

Most methods that have been proposed for evaluation of image processing can be classified into three groups, the analytical approach, the empirical goodness, and the empirical discrepancy.[17] The first class of methods addresses the algorithm only by theory and does not require its implementation. Hence, analytical analysis is unsuitable for reliable evaluations as required in medical applications. Approaches from the second class are based on some desirable properties ("goodness"), often established according to human intuition. For instance, the delineation of a complex-shaped structure is judged by medical personnel in two classes, true or false. However, a reference standard is not applied. The third class of evaluation methods relies on distance measures to a specific reference, which is usually called the gold standard.[17]

The importance of a ground truth (gold standard) for the evaluation of image processing procedures was also emphasized by NGUYEN & ZIOU.[15] Evaluation without a ground truth is termed parameter-free.[1] It is most questionable whether parameter-free evaluation sufficiently reflects the complex process of object representation in medical images. Another category of evaluation techniques distinguishes contextual and non-contextual approaches.[15] Non-contextual methods evaluate an algorithm by tests involving images with adjustable properties or systematic changes of parameters.[6] However, the evaluation is directly related to the algorithm, while contextual methods evaluate the suitability of an algorithm only for a certain application by means of the procedure-specific parameters subsequently determined by the application.

In summary, evaluation of medical image processing algorithms should be established by contextual methods of empirical discrepancy. In other words, it should rely on a reference standard. To reflect the high variability of medical images, a steady evaluation of image processing algorithms must rely on large sets of gold standard images. So far, a comprehensive collection of realistic image data together with confirmed segmentations and measurements is missing[2] or only available for specific tasks and modalities. 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 labeling of cryosections[4] 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*.

In this paper, we provide reliable criterions which, if acknowledged, result in true gold standards for the evaluation of image processing algorithms. Using the catchy criterions, any method of evaluation can be classified to be from plastic, silver, or gold. The consistency of this classification scheme is exemplified for each category.

2. A UNIFIED CLASSIFICATION SCHEME FOR REFERENCE STANDARDS

The term "gold standard" (ground truth) originates from medicine. Usually, a novel diagnostic procedure is validated against the true diagnosis, which is called the gold standard, to be scientifically acceptable. Based on a sufficient large set of data, this evaluation is performed by means of sensitivity and specificity. For example, histological findings are a reliable gold standard to evaluate radiological caries diagnosis. More general, a robust gold standard in medical diagnosis should be established by a method that is itself precise, i.e. reproducible, it should reflect the patho-anatomical appearance of the disease, and it should be established independently of the diagnostic method under evaluation.[16]

This definition of a diagnostic gold standard can be transferred to medical image processing, when the points raised above are regarded on a more abstracted level. Hence, we demand that a true gold standard in image processing satisfies the following criteria:

1. Reliance: The generation or capturing of test images must follow an exactly determined and reproducible protocol. In particular, it must not rely on interactive components such as manual delineation.
2. Equivalence: With respect to structure, noise, or other parameters of importance, the images or relations considered within an algorithmic gold standard must equal real-life data. For example, an x-ray phantom must not be made from a homogeneous material.

3. Independence: Any gold standard must rely on a different procedure or another image than that to be evaluated. For instance in image registration, artificial misalignment of the original image must not be taken to obtain the counterpart. Note that this criterion also bans the simultaneous use of one image for both, training and test phase.

This properties can be used to classify any reference standard, which is applied to validate image processing algorithms. If a reference standard fulfills all three criterions, it is named a gold standard. Any standard only satisfying two criteria is referred to as silver standard, if reliance is one of both. Other standards should termed to be from plastic, and such methods should not be used for validation. Note that this nomenclature is conform with the meager references in the field of validation of medical image processing[5,11]. In addition to the proper choice of the reference standards, two other properties must be satisfied to result in a meaningful evaluation of an image processing method:

4. Relevance: The algorithm to be evaluated must be self-reproducible. A quantitative measure that is extracted from an image (e.g., size of an object) must not change according to modified parameters in image acquisition that do not affect this measure (e.g., position of an object) and when the biomedical situation remains unchanged (i.e., imaging the same object).

5. Significance: The number of gold standard images that are used for evaluation must be sufficient large to enable statistically founded analysis and reasonable generalization. For example, the variance of leaving-one-out cross-validation must be small to ensure the generality of the classification result.

3. APPLICATION

The unambiguous criterions defined above can be applied to evaluate reference standards, which are in use in medical image processing. To avoid any defamation of other scientists, the author is critically reflecting only his own work. In particular, three examples are analyzed: image registration prior to subtraction of dental radiographs, segmentation of axo-somatic boutons at the cell membrane in microscopy, and collimation field detection in plain radiography.

3.1. Image registration prior to subtraction of dental radiographs

The detection of small changes in serial radiographs has been achieved using subtraction methods for more than two decades.[10] However, subtraction of radiographs requires perfect match of imaged structures. Using direct digital radiography based on either CCD-sensors or storage phosphor plates, a-posteriori registration is required after acquisition of radiographs, at least to compensate for planar rigid transforms (i.e. translation and rotation). Novel approaches are based on higher sophisticated models, such as affine or projective geometry, in order to compensate less standardized imaging geometry by more computational complexity. So far, results of computerized a-posteriori registration are demonstrated on numerous situations using in-vivo, in-vitro, or phantom images in combination with miscellaneous artificial manipulations. However, reliable validation is lacking. Hence, systems for computer-based subtraction are not comparable and therefore, about 20 years after the first development of digital subtraction imaging in dental radiology, subtraction is performed rather seldom in today's clinical routine.[10]

Regarding the evaluation of a planar registration technique, a straight forward reference standard is generated by transforming an image according to a known geometry, applying the algorithm under evaluation, and comparing the presetted geometry with the determined parameters (Fig. 1). Although the reference standards used are reproducible, they are neither equivalent nor independent: If serial radiographs are acquired without individual adjustment aids, structured noise is obtained from altered x-ray beam pathways but it is not incorporated into the reference standard. In addition, the reference is build from two images, which are obtained from the same exposure. Therefore, any evaluation that is based on such a kind of plastic standard in fact is meaningless for diagnostic routine.

Figure 2 shows a mechanical device, which has been used to produce reference standards for image registration.[9] A-priori known displacements are obtained by means of micrometer screws. A great variety of phantoms and specimens can be mounted to the device. Rotation of the specimen result in structured noise. Hence, this kind of reference standard is equivalent. Both, baseline and follow-up radiograph are captured independently. However, the precision of micrometer screws (finest scale > 100 μm) is low with respect to the high resolution of the digital sensor used for intraoral imaging (pixel size < 50 μm). Therefore, reference standards obtained using this device are not reproducible and hence, they are referred to being from plastic.

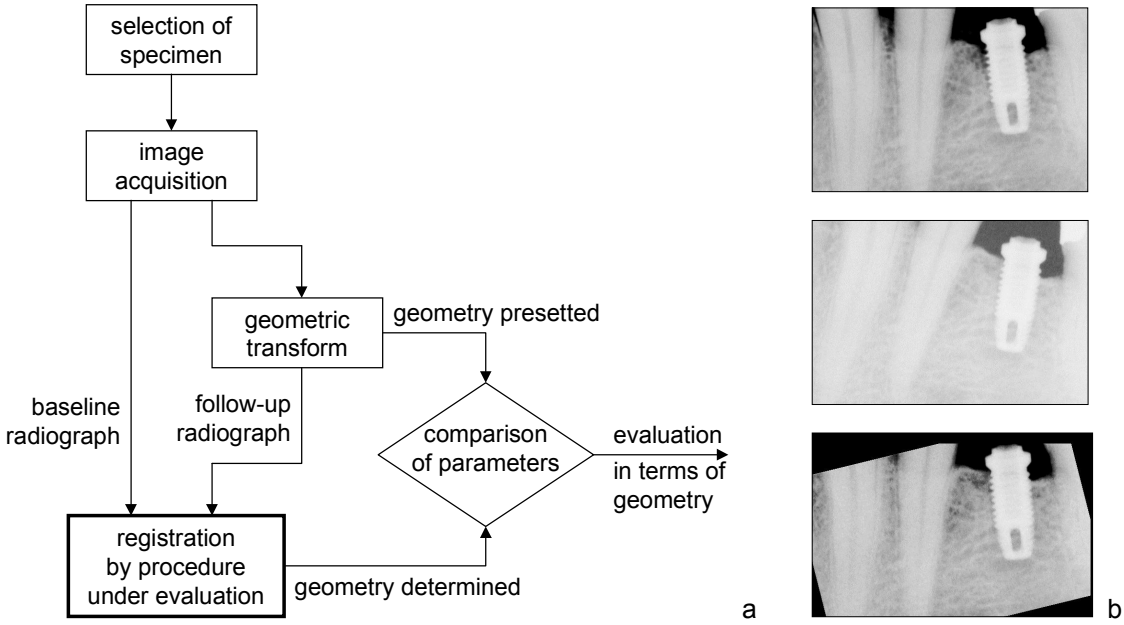

Fig. 1: Reference standard based on simulation. (a) flow chart, (b) baseline radiograph, computed follow-up with altered geometry and contrast, and result of registration.

In a recent investigation, a high-resolution volume representation of a formalin-preserved segment of a human maxilla was synthesized from a set of 51 digital radiographs equidistantly covering the entire sampling aperture by means of Tuned-Aperture Computed Tomography (TACT©).[13] Two-dimensional (2D) projection renderings of this three-dimensional model were generated yielding arbitrary but well known 2D projections with, and without structured noise. Because of the similarity of their appearance to actual 2D radiographs, such synthesized images are termed virtual radiographs. By means of virtual radiographs, 2D registration techniques can be evaluated either in terms of geometry or in terms of image similarity (Fig. 3). Since virtual projections are computed, such reference standard is reproducible. Any desired geometry can easily be realized including various focal spot dimensions. Also, Poisson-distributed photon

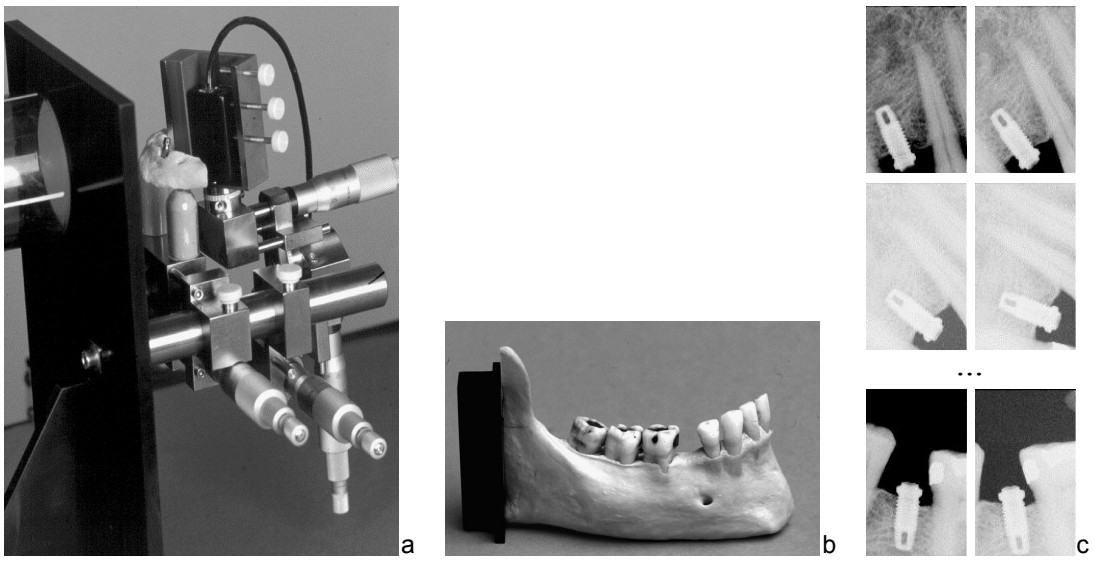

Fig. 2: Reference standard based on controlled geometry. (a) mechanical device, (b) mounted specimen, (c) example images.

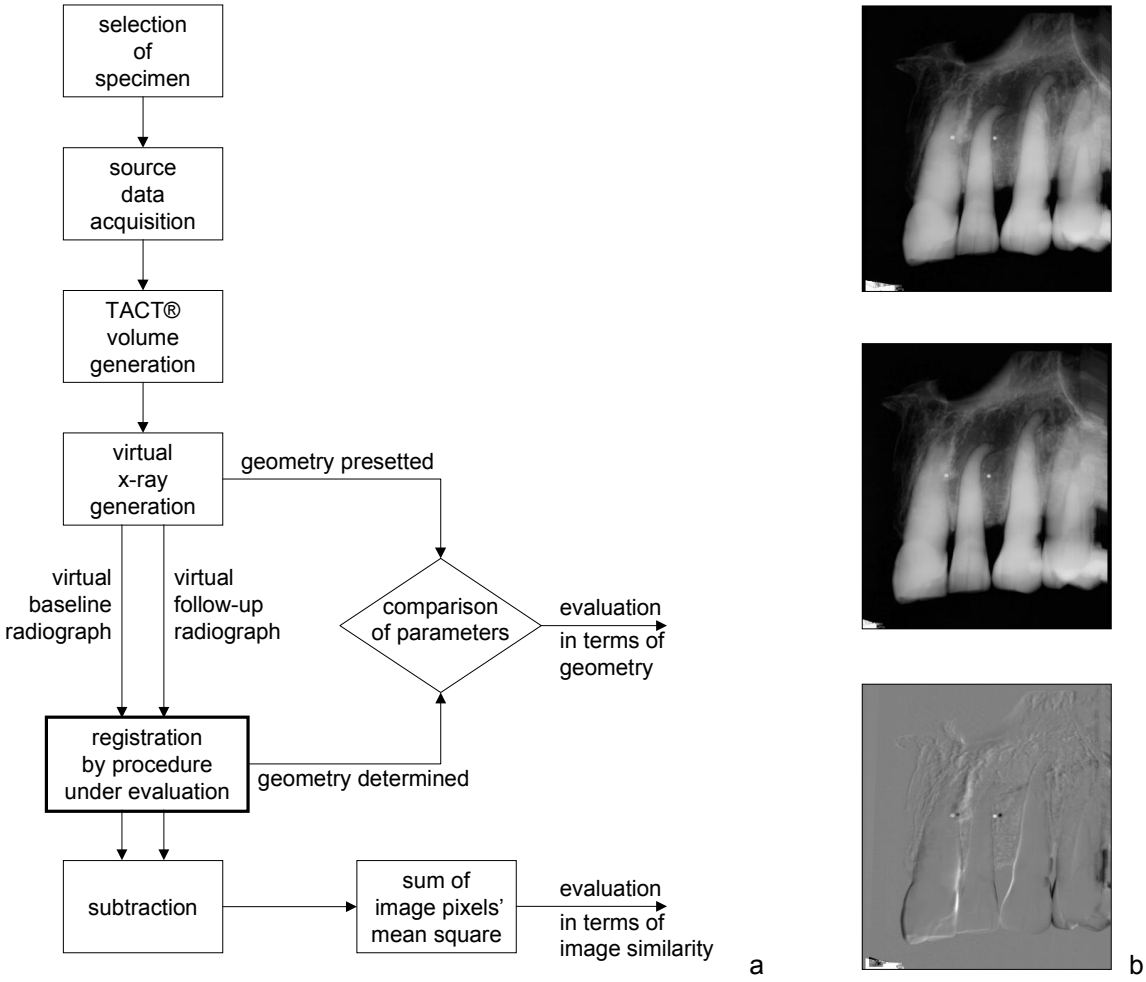

Fig. 3: Reference standard based on virtual radiographs. (a) flow chart, (b) virtual baseline and virtual follow-up radiograph with 5° angle disperancy and subtraction image with structured noise.

noise is easily modeled during rendering. Since virtual radiographs are based on computed tomography but applied as reference standard for plain radiography, independence is given. Therefore, virtual radiographs provide a suitable gold standard as reference for 2D registration techniques prior to subtraction of intraoral radiographs. However, the virtual constellations of source, patient, and sensor must cover clinical situations and a sufficient number of TACT volumes should reflect inter-patient variability.

3.2. Segmentation of axo-Somatic boutons at the cell membrane in microscopy

Traumatic injury to human spinal cord induces a complex pattern of functional alterations. So far, the pathophysiological mechanisms underlying these acute and chronic alterations are only partially understood. By now they are examined on a light-microscopic level, where an objective and unbiased approach to the quantification of synaptic input to the somal surface of motoneurons has been established.[12] Based on micrographs displaying single neurons, a finite element balloon model is applied to determine the exact location of the cell membrane. A synaptic profile is extracted next to the cell membrane and normalized with respect to the intracellular brightness. Thereafter, measures of absolute staining S_{abs}, absolute homogeneity H_{abs}, two-dimensional allocations of boutons $A_{2D,}$ as well as their one-dimensional projection A_{1D} are determined automatically from the synaptic profiles.

Several random-effected parameters may influence this complex system of image processing. For example, the balloon-based segmentation method for delineation of the cell membrane is followed by a simulated annealing for local optimization. More general,

- interactive components (e.g., placement and recording of the object, image orientation, manual setting of start points),
- algorithmic components (e.g., order of processing such as row by column, sorted lists, fixed start points), or
- random components (e.g., stochastic optimization, simulated annealing, evolution strategies, random start points)

may strongly effect the quantitative measurements. Whenever such components are included into an image analysis system, the relevance of the system must be shown before evaluation against any reference standard is meaningful. However, this is often disregarded when novel algorithms are published. The coefficient of variation

$$C_v = \frac{\sigma}{\mu} \tag{1}$$

is a suitable measure to assess the relevance of an algorithm. For a sufficient number of repetitions, C_v normalizes the standard deviation of measures σ to their mean μ. In general, the coefficient of variation for a reliable measure must be smaller than 5%, which means that 95% of the measures differ less than 1/20 from their mean.[8]

With respect to the quantification of synaptic boutons at the cell membrane, stochastic optimization, manual placement of the cell for image recording as well as microscope settings may have a non-negligible effect on the quantitative measurements. To determine the systems variation caused by stochastic optimization, an arbitrarily chosen image was analyzed 20 times based on different random seeds. For all measures, the coefficient of variation was smaller 3%(Fig. 4). Therefore, stochastic optimization of the final balloon position is sufficiently reproducible. To determine the variation induced by manual placement of the cell, an arbitrarily chosen cell was captured at 20 positions shifting the specimen from the center in all eight directions until the membrane nearly reached the edge of the image frame, and clockwise rotating the cell in steps of approximately 36°. Note that the measured 2D allocation varies almost 5% (Fig. 4). The microscope allows manual settings of focus and illumination without a reliable scale. Since the algorithm was designed to normalize different illuminations, the quantitative measures should stay constant. To determine the variation induced by microscope settings, an arbitrarily chosen cell was captured 20 times with the illumination covering the entire scale from dark to light. Since the coefficient of variation of both allocation measures is larger than 5% (Fig. 4), the system is only capable to measure relative allocations within a series of cells that are subsequently micrographed but, despite the build-in normalization, the system is weak in the determination of absolute measures.

So far, it has been shown that the image analysis system acts like a system with respect to linearity and shift invariance. However, it must also be shown that the quantitative measures in fact correlate with the true situation. In other words, the algorithm must be validated against a reference standard. In a first step, the balloon-based segmentation of the cell membrane was evaluated.[11] In medical image segmentation, the ground truth is often unknown. In other words, a gold standard does not exists. Hence, algorithms often are "evaluated" only by visual inspection (i.e., parameter-free) or a distance measure is computed with respect to the manual delineation in a low number of images. Since manual delineation is not reproducible, such kind of references is from plastic. A large number of reference images can be obtained from a small number of samples if realistic contours and textures, which represent the appearance of tissue imaged by any modality, are stochastically combined.[11] For that, texture samples are extracted manually from the inside, outside, and boundary zone of cell micrographs. Each sample is represented by its mean and the magnitude and phase of its Fourier-transform. For texture synthesis, mean, magnitude, and phase are arbitrarily combined and the

stochastic optimization				
	S_{abs}	H_{abs}	A_{2D}	A_{1D}
μ	20.23	5.75	41.74%	76.80%
σ	0.44	0.04	1.16%	1.57%
C_v	2.19%	0.61%	2.78%	2.04%

manual placement of cell				
	S_{abs}	H_{abs}	A_{2D}	A_{1D}
μ	20.64	5.57	39.49%	73.46%
σ	0.70	0.06	1.96%	2.11%
C_v	3.39%	0.97%	4.96%	2.87%

microscope settings				
	S'_{abs}	H'_{abs}	A_{2D}	A_{1D}
μ	54.49%	83.23%	15.18%	40.85%
σ	1.68%	1.64%	0.83%	2.41%
C_v	3.10%	2.00%	5.50%	5.89%

Fig. 4: Variation of quantitative measures induced by stochastic optimization, manual placement of cell, and microscope settings.

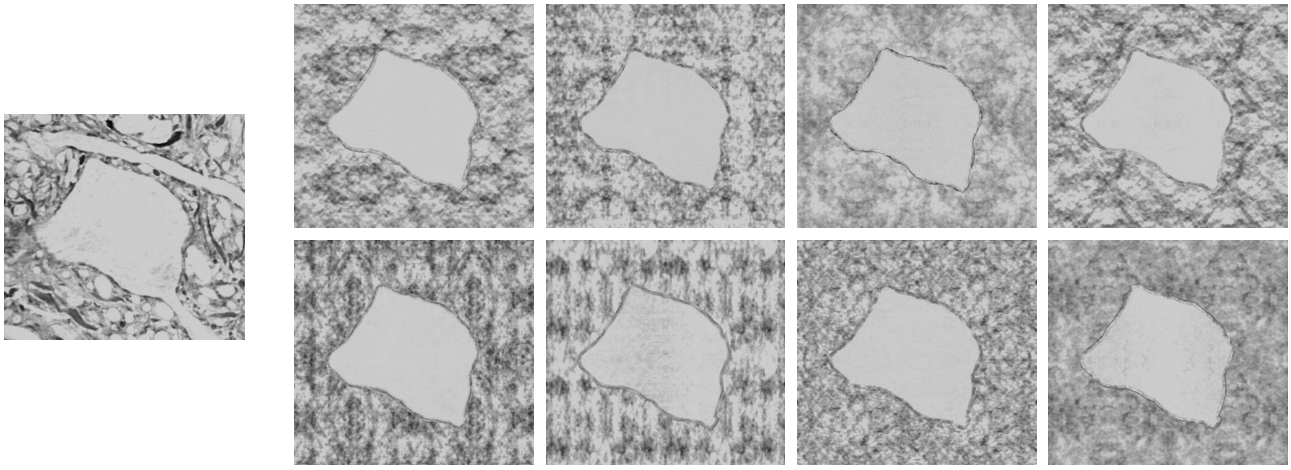

Fig. 5: Reference standards obtained from Fourier-based texture synthesis to evaluate segmentation procedures. Original micrograph (left) and synthetic images with exactly known ground truth (right).

phase is randomly modified before the texture is re-transformed into the spatial domain (Fig. 5). Since the texture synthesis follows an exactly determined and reproducible protocol, the reference standard is reliant. Using two samples for each of internal, external, and boundary texture, a total of 2^8=256 reference images is obtained disregarding the random alterations of phase. The textures can be mapped onto any given contour. Therefore, the standard is equivalent. However, since it is based on the same images and image modality, it is not independent. In summary, this standard can be termed to be from silver.

3.3. Automatic localization and delineation of collimation fields in digital and film-based radiographs

Collimation field detection is an important pre-processing step for automatic image analysis.[14] Especially for content-based image retrieval, the image area covered by shutters or collimation fields must not be considered for global color, texture, or structure analysis. Since shutter edges are easily visible and composed from linear or circular primitives, the ground truth can be determined manually for this particular case. Suppose the relevance for evaluation is given (i.e., Criterion 4 is satisfied), the evaluation must rely on a sufficient large number of images (Criterion 5). On the one hand, statistical methods can be applied to calculate the sample size. On the other hand, the number of reference images often is limited for some practical reasons. For the evaluation of shutter detection, a large number of 4,000 images have been arbitrarily chosen from clinical routine, 763 of them showing some shutter or collimation field edges. In order to satisfy Criterion 3, i.e., independence, eight subsets each of 500 images have been compiled. For each of the eight compilations, the parameters of the algorithm were optimized based on 500 images and the remaining 3,500 references were used for evaluation.[14] In terms of sensitivity, specificity, and precision, the results of optimization vs. evaluation for the same combination differed up to 13, 1, and 3 percentage points, respectively, while that of evaluations for different combinations altered up to 9, 2, and 1 percentage points, respectively (Fig. 6). This emphasizes a still insufficient number of reference images to allow absolute generalization of the results. Note, however, that most "evaluations" published in medical image processing are based on a significantly smaller number of images for algorithms with much more or more crucial parameters, often without the required separation of training and test data sets.

4. RESULTS

A great variety of reference standards have been used in the literature of medical image processing. With respect to the unambiguous criterions: reliance, equivalence, independence, relevance, and significance, most of them are from plastic.

image subset	training			evaluation		
	sensitivity	specificity	precision	sensitivity	specificity	precision
0		0.99	0.91	**0.52**	0.98	**0.89**
1	0.47	**0.98**	0.89	0.56	**0.99**	0.90
2	0.48	0.98	0.89	0.56	0.99	0.90
3	**0.74**	0.98	0.93	**0.61**	0.97	0.90
4	0.58	0.98	0.90	0.55	0.99	0.90
5	0.49	0.99	0.88	0.56	0.98	0.90
6	0.63	0.99	**0.93**	0.54	0.99	**0.90**
7	0.55	0.98	0.91	0.55	0.99	0.90

Fig. 6: Results of cross-validation. The marked values are referred to from the body text.

The main problem is the lack of a reliable ground truth, which arises from image acquisition, image contents, and task-specific medical problems. Thus, reference standards have to take these aspects into consideration. It became obvious throughout the examples, that a formal expression of the ground truth is inevitable. When regarding reliance, the test images have to be achieved in a formal and reproducible way, which has to be defined in cooperation of the medical expert and the technical developer to match the need of both medical and algorithmic relevance.

The reference standard must be designed task-specifically. Real-life image data from clinical routine is suitable for equivalence and therefore to guarantee the usability of a new procedure if and only if the ground truth is a-priori known. However, this is rather seldom in medical imaging. Hence, constructed, synthetic, or other artificial reference standards are frequently applied. However, a new algorithm only makes sense if it really works on the planned type of data. Hence, the standards must be equivalent to the routine situation, the algorithm is designed for.

When testing a new algorithm, it has been pointed out that results can be outstanding if the same image is used as reference and training data. But concerning real-life data with a high inter-pictorial variability, the new algorithm may be useless. Hence, independence is most important for any reference standards. Dependence also arises from algorithmic affinity, e.g., if Fourier-based techniques are used to design references for an Fourier-based image processing method.

These three criteria form a gold standard. Their fulfillment can easily be checked and thus, ensure a qualitative approach to validation of image processing algorithms. The validation requires a critical and distanced view on the new method by its developers.

Now considering the algorithm under quantitative evaluation, its reliability and robustness to parameter alterations has to be proven. This can be done analytical, as, for instance, described by HARALICK.[6] Also, it can be tested experimental by an appropriate selection of real-life images that contain the object of interest in different manners. The goal of this step is not to show, that the quantities determined by the algorithm correlate with the actual situation, but to show the quantities stay constant if the underlying biological situation remains unchanged.

This leads directly to the last aspect, the significance, which is only guaranteed by a sufficient high amount of test data. While most algorithms in medical image processing are usually evaluated based on a very small number of references, some databases with several thousands of references currently are developed. For instance, BROWN describes a medical image processing algorithm verification database containing 4,000 computed tomography data sets of the head.[2] However, we have shown that even 4,000 reference images might be insufficient to enable absolute generalization of results.

5. CONCLUSIONS

In medical image processing, novel methods, concepts, or algorithms are often presented without sufficient evaluation. In this paper, a reliable definition of a gold standard for evaluation of image analysis procedures is given using modular

criterions. It was shown that a gold standard can be formalized and brought to computational traceable unambiguous rules, which evolve from each other, and which may be taken into account even on the level of algorithmic design, to support the evaluation phase. We hope that these criteria will have essential impact to researchers in all fields of medical image processing, the design of novel algorithms and the development of routine applications.

Nonetheless, the proposed scheme is qualitative, i.e. the quality of a reference standard can not be calculated. This may be addressed by refining the criterions described above.

Once a reference standard is established, it supports the transfer from scientific work into medical routine because researchers, physicians and legal authority now posses means of reliably controlling and comparing new and advanced methods of medical image processing.

REFERENCES

1. M. Borsotti, C. Campadelli and R. Schettini, "Quantitative Evaluation of Color Image Segmentation Results," *Pattern Recognition Letters*, **19(8)**, pp. 741–747, 1998.
2. C. W. Brown, "Building a Medical Image Processing Algorithm Verification Database," *Proceedings SPIE Medical Imaging*, **3979**, pp. 772–780, 2000.
3. J. C. Gee, "Performance Evaluation of Medical Image Processing Algorithms," *Proceedings SPIE Medical Imaging*, **3979**, pp. 19–27, 2000.
4. 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.
5. R. M. Haralick, Oral Keynote Speech at SPIE's Medical Imaging Symposium, 2000.
6. R. M. Haralick, "Validating Image Processing Algorithms," *Proceedings SPIE Medical Imaging*; **3979**, pp. 2–27, 2000.
7. D. R. Haynor, "Performance Evaluation of Image Processing Algorithms in Medicine: A Clinical Perspective," *Proceedings SPIE Medical Imaging*, **3979:** p. 18, 2000.
8. D. Ingram and R. F. Bloch, *Mathematical Methods in Medicine. Part I: Statistical and Analytical Techniques*, Chinchester, U.K.: Wiley, 1984.
9. 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.
10. T. M. Lehmann, H.-G. Gröndahl and D. Benn, "Computer-based Registration for Digital Subtraction in Dental Radiology," *Dentomaxillofacial Radiology* **29(6)**, pp. 323–346, 2000.
11. T. M. Lehmann, J. Bredno and K. Spitzer, "Silver Standards Obtained from Fourier-Based Texture Synthesis to Evaluate Segmentation Procedures," *Proceedings SPIE Medical Imaging* **4322(1)**, pp. 214–225, 2001.
12. 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 Transactions on Biomedical Engineering*, **48(6)**: pp. 706–717, 2001.
13. T. M. Lehmann, P. F. Hemler and R. L. Webber, "Virtual Radiographs Computed from TACT® Volume Data as a Gold Standard for Image Registration Prior to Subtraction," *Dentomaxillofacial Radiology*, accepted for publication, 2002.
14. T. M. Lehmann, S. Goudarzi, N. I. Linnenbrügger, D. Keysers and B. Wein, "Automatic Localization and Delineation of Collimation Fields in Digital and Film-Based Radiographs," *Proceedings SPIE Medical Imaging,* contents of this volume, 2002.
15. T. B. Nguyen and D. Ziou, "Contextual and Non-Contextual Performance Evaluation of Edge Detectors," *Pattern Recognition Letters*, **21**, pp. 805–816, 2000.
16. A. Wenzel and H. Hintze, "The Choice of Gold Standard for Evaluation Tests for Caries Diagnosis," *Dentomaxillofacial Radiology* **28**, pp. 132–136, 1999.
17. Y. J. Zhang, "A Survey on Evaluation methods for Image Segmentation," *Pattern Recognition*, **29(8)**, pp. 1335–1346, 1996.