At what point does the human diagnostician's eye no longer remain the "gold standard"?

Pierre Seroul
Newtone Technologies
pseroul@newtone.fr

The “gold standard test” can be defined as the best diagnostic test under reasonable conditions [1]. Starting from this definition, two aspects of a good diagnosis can be defined: its quality (Am I right?) and its consistency (Am I always taking the same decision?). Diagnosticians and computer vision algorithms have different approaches to establish assessments. Each of them has qualities to be the ideal gold standard or the worse.

No algorithms can yet tie the experts diagnosis quality

Any human being never stops learning along his life and from each case he is faced. He is able to adapt his point of view even if there are important variations in the conditions of observation. By integrating more environmental dimensions than any algorithm, human being can keep the same interpretation of a fact independently of the situation [2]. This can be easily demonstrated by the well known Adelson’s checkerboard (Figure 1). Any computer vision algorithm interprets square A and B to have the exact same color whereas a young child, as soon as he knows colors, can tell that A is black and B is white. The difference of light produced by the shadow of the cylinder is automatically integrated in the child brain with no prior information on the lighting conditions. This task would be very hard to achieve by algorithms without adding many inputs such as 3D description of scene and shapes, lights and viewing conditions and a full set of hypothesis about light scattering on each object [3].

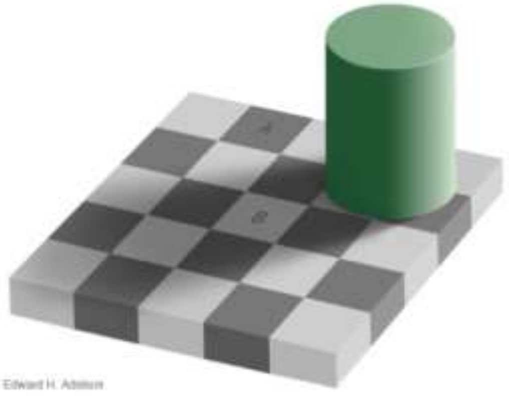

Figure 1 - Adelson's Checkerboard

In most clinical application, physicians remain the gold standard and the quality of their judgment is not yet equaled by algorithms. This is the case, for example, in differential diagnosis of pigmented skin lesions. Informal exchanges with dermatologists from three hospitals [4] lead us to understand that diagnosticians not only work on images but integrate many other parameters such as risk factors. Of course, these factors are well known and evaluated. Several dermoscopic algorithms exist to combine some of them like the ABCD rule [5] or the Menzies method [6] but none of these techniques can replace the dermatologist to establish if the lesion is a melanoma or a benign nevus [7]. Even full systems like the MelaFind®, combining acquisition hardware, image processing and training algorithms enhanced during 15 years, are not yet considered as diagnostic tool by the FDA: "MelaFind® should be used in conjunction with dermatological clinical expert assessment...[It] should not be used to confirm clinical diagnosis of melanoma."[8].

Algorithms will always perform the same diagnosis from the same conditions

But, because diagnosticians use no quantitative and automatic process to interpret the severity of a disease, they are prone to subjectivity. Their diagnosis depends on their experience, on their state of mind or on their fatigue level… The consistency of their interpretation can be affected. The main advantage of an algorithm to be use as a gold standard is that it always provides the same interpretation from the same image.

This is especially true if we are not dealing with boolean interpretation (malign or benign) but with measurement or quantification. Many studies had shown that if you ask several experts to drawn the borders of a lesion, results can be truly different [9]. Without having biological information extracted from the pathological section, none of the borders of the experts can be considered as the ground truth. None of the experts is more accurate than the other. In this case, it is not possible to insure that a diagnostician provides a result closer to the ground truth than an algorithm. It depends on the definition of the truth. In these cases, algorithms provide, at least, exactly the same border between two trials while clinicians cannot be so accurate.

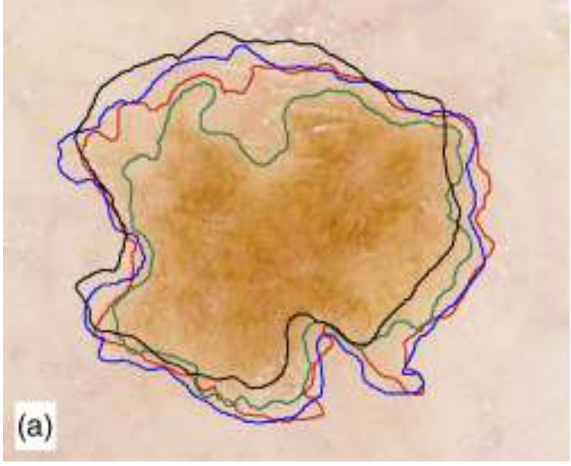

Figure 2 - Four differents dermatologists-drawn borders for a sample dermoscopy image [5]

Note that a diagnostician can be not only a physician but any kind of expert who performs quality control. In industry, assessment of compliance is more and more processed by image algorithms. It is usually possible to establish norms about color [10] and shape of a product under controlled conditions. Setting thresholds on an image can detect defects or deviation compared to a standard [11]. Image processing provides a quantitative assessment which is much more accurate than the qualitative estimation that an operator can make.

For multivariate quality control, based on color perception, gloss analysis or shape characterization, training algorithms (Neural Network, Support Vector Machine [12]…) or multivariate analysis (Principal Component Analysis, Exploratory Factor Analysis [13]…) can model and simulate the diagnosis of a bench of experts. In this case, algorithms provide results that appear to be a consensus between all the experts [14]. The quality of the diagnosis is thus enhanced and perfectly reproducible under controlled conditions.

Combining image processing and human perception enhances diagnosis quality and consistency

Maybe the question is not about finding a gold standard between diagnostician and image processing but how to combine them in order to go further in term of numbers of cures and, thus, to define a new gold standard as combination of both.

In more and more applications, image processing does not replace diagnosticians but helps them to take a decision. It is a new way to guide the physician or the quality control operator by providing quantitative information or by adding augmented reality.

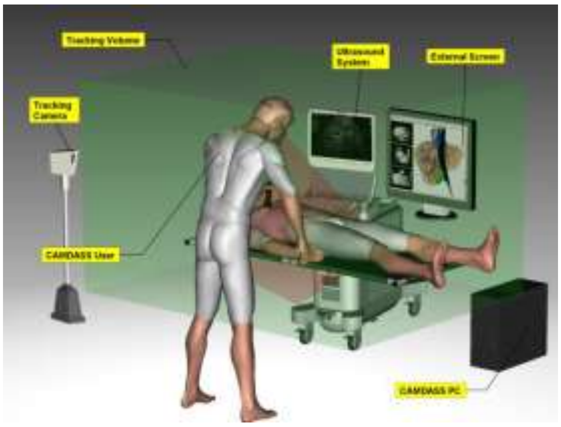

Figure 3 - Artistic representation of CAMDASS [8]

Algorithms are not used randomly. They are built from the observation and the comprehension of the human reasoning [15]. It can be seen as the automation of the experts thought. Providing selected information to the diagnostician is a way to confront his first feeling with an objective measurement. It is a good way for him to have a second guidance on his work and to choose if he has to take it into account.

For non-experts, augmented reality is a beautiful way to give key information at the good moment in order to decide what to do. The project CAMDASS (Computer Assisted Medical Diagnostic and Surgery System) provides 3D guidance in diagnosing medical problems for astronauts [16]. Their lack of experience and knowledge cannot lead to a good appreciation and image processing is not developed enough to adapt to each situation and patient. But together, they are able to achieve correct diagnosis.

The number of projects in which the aim of image processing is not to replace diagnosticians but to assist them is growing. We start to realize that combining image processing and human being can achieve the task of improving what a gold standard is.

References:

1.http://en.wikipedia.org/wiki/Gold_standard_(test)

2. Arnheim, Rudolf, Visual Thinking, *University of California Press, Berkeley, 1969.*

3. Elias, M. and Elias, G. Radiative transfer in inhomogeneous stratified scattering media with use of the auxiliary function method. *J. Opt. Soc. Am. A, 2004, 21(4):580–589.*

4. ANR TecSan 2010 0: Melascan Project.

5. Nachbar F and al., The ABCD rule of dermatoscopy. High prospective value I n the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol 1994 Apr;30 (4):551-9.*

6. Scott W. Menzies and al., Frequency and Morphologic Characteristics of Invasive Melanomas Lacking Specific Surface Microscopic Features. *Arch Dermatol. 1996;132(10):1178-1182*

7. Dolianitis C, Kelly J, Wolfe R and Simpson P. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. *Arch Dermatol. 2005 Aug;141(8):1008-14.*

8. PMA P090012 FDA Summary of Safety and Effectiveness Data

9. Garnavi R, Aldeen M and Celebi M. E. Weighted performance index for objective evaluation of border detection methods in dermoscopy images. *Skin Research and Technology 2011; 17: 35–44*

10. CIE Draft Standard DS 014-6/E:2012

11. Thomas A.D.H, Rodd M.G, Holt J.D, Neill C.J. Real-time Industrial Visual Inspection: A Review. *Real Time Imaging 1995 June; Volume 1 Issue 2:139-158*

12. Kecman V., Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models, *The MIT Press, Cambridge, MA, 2001*

13. Fabrigar and al., Evaluating the Use of Exploratory Factor Analysis in Psychological Research, *Psychological Methods 1999, Vol. 4. No. 3.272-299*

14. Kano M, Nakagawa Y. Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers & Chemical Engineering 2008 Jan, Volume 32, Issue 1-2:12-24*

15. Duch W, Oentaryo R.J. and Pasquier M., Cognitive Architectures: Where do we go from here?, *Proceedings of the conference on Artificial General Intelligence 2008, p 122-136*

16. Nevatia Y, Chintamani K, Meyer T, Blum T, Runge A, Fritz N. Computer Aided Medical Diagnosis and surgery system: Toward Automated Medical Diagnosis For Long Term Space Missions.