ABSTRACT

With medical imaging technologies growth, the question of their assessment on the impact and benefit on patient care is rising. Development and design of those medical imaging technologies should take into account the concept of image quality as it might impact the ability of practitioners while they are using image information. Towards that goal, one should consider several human factors involved in image analysis and interpretation, e.g. image perception issues, decision process, image analysis pipeline (detection, localization, characterization...). While many efforts have been dedicated to objectively assess the value of imaging system in terms of ideal decision process, new trends have recently emerged to deal with human observer performances. This task effort is huge considering the variability of imaging acquisition methods and the possible pathologies. This paper proposes a survey of some key issues and results associated to this effort. We first outline the wide range of medical images with their own specific features. Next, we review the main methodologies to evaluate diagnostic quality of medical images from subjective assessment including ROC analysis, and diagnostic criteria quality analysis, to objective assessment including metrics based on the HVS, and model observers. At last, we present another evaluation method: eye-tracking studies to gain basic understanding of the visual search and decision-making process.

Index Terms – Diagnostic accuracy, medical images, ROC analysis, model observer, eye-tracking

1. INTRODUCTION

Over the past 30 years, information technology has facilitated the development of digital medical imaging. This development has mainly concerned Computed Tomography (CT), Magnetic Resonance Imaging (MRI), the different digital radiological processes for vascular, cardiovascular, and contrast imaging, mammography, diagnostic ultrasound imaging, nuclear medical imaging with the Single Photon Emission Computed Tomography (SPECT) and the Positron Emission Tomography (PET), optical imaging including video-endoscopies, microscopy, etc. [1][2].

The emergence of new technologies raises important questions concerning optimization of the acquisition, storage, transfer, and display of image, choice of appropriate display media and format, optimization of image compression, and optimization of image processing and computer-aided detection (CADe) and diagnosis (CADx) [3], etc. It is only through systemic and objective evaluation of the entire imaging system -from hardware to human interpretation of images- that these questions can be addressed [4][5].

But, how are we to measure diagnostic quality? Part of the answer stands in the fact that the pathological condition is what determines the information which must be retained in any given medical data. This information may be large in volume, but not sufficiently contrasted with the surrounding tissue, or as opposed to that it may be small, linear or punctiform details closed to the resolution gathered by the acquisition device. In fact, both of these categories of information may be required within one image, for diagnostic purposes. Diagnostic quality is therefore highly dependent on the protocol of both respective gathering technique and concerned pathological condition that makes the evaluation of medical images quality a complex task.

Numerous studies were conducted to develop methods that evaluate diagnostic accuracy in medical images in order to improve radiologists’ performance and reduce their interpretation variability. Evaluation methodologies cover a broad range, from subjective assessment (see section 3) including the widely used Receiver Operating Characteristic (ROC) techniques that measure diagnostic accuracy to objective assessment (see section 4) including metrics such as JND models that simulate human visual system perception and model observers that perform classification...
tasks (e.g. with or without disease). In section 5, we present recent studies using eye-tracking devices that analyze image-interpretation workflow in order to understand perceptual and cognitive processes that are the foundation of medical image interpretation. Such studies can be useful to optimize image quality, image presentation and radiologist training to improve diagnostic efficiency.

2. SPECIFICITIES OF MEDICAL IMAGES

Medical imaging was initiated and developed due to the diversity of physical phenomena being used (X-rays, γ-rays, ultrasound waves, magnetic nuclear resonance). Medical imaging was further developed with the increased use of computers in the acquisition process (real-time treatment of a large amount of information) as well as for image reconstruction (tomography).

Each medical imaging modality (digital radiology, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound imaging (US)) has its own specific features corresponding to the physical and physiological phenomena studied, as shown below in Figure 1. More information related to the guiding principles of the principal medical imaging devices and to the most relevant features of the medical images, may be found in [2].

![Figure 1. Sagittal slices of the brain by different imaging modalities: a) magnetic Resonance Imaging (MRI), b) computed Tomography (CT), c) positron Emission Tomography (PET), d) ultrasound (US)](image)

Then, the pixel or voxel values depend on the chemical and physical characteristics of the tissues studied. These characteristics often correspond to a physiological phenomenon. Imaging mechanisms based on spontaneous contrasts often provide anatomic information, while imaging mechanisms known as “functional”, often use markers reflecting fluid motion or metabolic exchanges. These mechanisms efficiently depict important details on how well the body functions.

As the data gathering techniques vary, the images produced by each are different: in spatial resolution, contrast, and type of noise as well. For this reason, we often refer to studies evaluating the legibility of the diagnosis by a radiologist. This is the subject of the next session.

3. SUBJECTIVE QUALITY ASSESSMENT

Subjective quality assessment methods are run by a group of experts who rate the quality of the diagnosis as well as the visual quality of the image. They work in the adapted conditions, similar to those applied during clinical routine, allowing them to observe all the necessary details.

In fact, the choice of assessment method depends on the task that is to be evaluated. In other words, it depends on the information that can be extracted from the images themselves. Usually, there are two different types of task:
- detection tasks that call upon a binary answer (presence or absence of pathology);
- estimation tasks that lead to an estimation of a quality grade according to specific diagnosis criteria.

Two important subjective assessment approaches are then suggested to judge on the quality of compressed medical images:
- the assessment of diagnosis reliability, classified as an assessment of detection task: the most common method used is one based on the ROC (Receiver Operating Characteristics) analysis;
- the assessment of the diagnosis criteria quality, classified as an assessment of estimation task.

3.1. Experiment protocol

The most popular experiment is the so-called multiple-reader multiple-case (MRMC) paradigm. This experiment involves multiple cases of various difficulties with known disease truth status and multiple readers of various skill levels. There are potential biases in the design of the MRMC experiment that one should minimize to make the experiment more effective and powerful [6]. For example, there are suspicious advantages when a physician reads images of a patient for the second time. Therefore there is a potential bias in favor of the modality that is read second compared with the modality that is read first. For imaging modalities that are not used together clinically, physicians have to read images of each modality independently to minimize this bias. For computer-aided diagnosis assessment, a sequential design is more appropriate (physicians read each case first without the computer aid and then, immediately after, read the case again with the computer aid), because this study design mimics the clinical use of computer aid [4][7].
3.2. Analyzing the diagnosis reliability – ROC

Diagnosis can be seen as a binary decision (patients being either normal or pathological) where four different situations may occur, depending on whether the observer takes one or the other decision according to the established reality of the gold standard. These four situations are summarized below in Table 1.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician’s Answer</td>
<td>Positive</td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

Table 1. Diagnostic test with TP the True Positive fraction (disease correctly classified as positive), FP the False Positive fraction, FN the False Negative fraction and TN the True Negative fraction

Two diagnostic indices have been defined that characterize how correct the answers on classification are: sensitivity $Se$, specificity $Sp$.

$$Se = \frac{TP}{TP + FN} \quad Sp = \frac{TN}{TN + FP}$$

The $Se$ and $Sp$ values are directly related to the prevalence of the disease, according to the Bayes theory. They therefore remain coherent even when the disease is only slightly present. Nevertheless, every assessment method using these diagnostic values requires a gold standard that provides the reference diagnosis. This is not always an easy task, especially when the images being studied normally form the gold standard (such as X-ray images of the coronary arteries and numerous MRI, etc.). To solve this problem, a commonly used approach is to call upon a group of physicians who are then asked to establish, according to consensus, a diagnosis on the images being studied. It is also possible to ask these physicians to give their individual opinions on the image, and then keep only the images on which all physicians raise a similar diagnosis.

In the medical field, there is no set way to distinguish a normal subject from a pathological one. In practice, the physician works spontaneously using density guide outlining the probabilities of having normal or pathological cases according to a law looked upon as a Gaussian function, as indicated in Figure 2.

![Figure 2. Distribution of medical diagnosis](image)

The sensitivity and specificity values vary according to two different phenomena: where the decision threshold settled by the physician lies and the detailed precision of the pathology (the more subtle it is, the more overlapping there is between the Gaussian probability density curves). The ROC approaches solve the problem concerning the decision threshold for the simple reason that they specifically allow the evaluation of a system without considering the chosen decision threshold. In other words, the ROC approach consists of representing the sensitivity value as a function of the specificity for all threshold values possible, and then joining these two points on the curve, as shown in Figure 3. Each point on the curve therefore represents a compromise between sensitivity/specificity corresponding to a particular decision threshold. The ROC curve summarizes the entire range of sensitivity/specificity compromises for all the different threshold values.

![Figure 3. ROC curve](image)

In order to artificially vary the decision threshold, the most common technique in medical imaging is to ask the physician to report his diagnostic confidence in terms of 4-, 5-, or 6-point ordinal scale or in terms of a quasi-continuous ordinal scale (1-100) [8]. It has been shown that if readers are able to use quasi-continuous scales, then the results can benefit ROC-curve fitting [9].

For a 5-point ordinal scale (the disease is most certainly present (1), probably present (2), the case is being disputed (3), the pathology is probably absent (4), certainly absent (5)), there are 5 possible levels of answer. Amongst these responses, 4 couples ($Se$, 1-$Sp$) are created by simulating different decision thresholds that a physician could have had if the answer was strictly binary. A variety of methods for fitting ROC curves to observer data and testing the statistical significance of apparent differences are reported in [10]. It is possible to interpret ROC curve in a qualitative manner. The shape of the curve only characterizes the detection performances. A curve that has merged within the diagonal of the two axes (Figure 4.a) corresponds to a non-discriminating system. In other words, the answer given by the observer is in no way related to the presence or absence of an anomaly. A curve shape like the one represented in Figure 4.b corresponds to a perfectly discriminating system, for which there exists a specific decision threshold that
separates both groups distinctly. In practice, effective ROC curves have a shape halfway between these limits. For example in Figure 4.c, the curve corresponds to system A that reveals better detection performance than those corresponding to system B.

![ROC Curves](image)

**Figure 4. Interpretation of ROC curves:**
(a) non-discriminating system, (b) perfectly discriminating system, (c) usual system (system A > system B)

The AUC (Area Under Curve) represents the probability that we can correctly identify the image containing an anomaly when both an image with and another without an anomaly are simultaneously presented to the observer. It is the usual figure of merit to characterize the value of a system.

ROC analysis is appropriate mainly for binary diagnostic assignments (normal/pathological) without requirement about location specification. To consider location precision, the LROC (Location ROC) analysis (adapted to images that have only zero or one lesion), requires the physician to both detect and locate the lesion if it is present. A LROC curve plots the proportion of true-positive images with correctly located lesions at each rating criterion, as a function of its corresponding false-positive rate [11]. Moreover, the traditional ROC analysis cannot be applied in the case of a diagnosis involving multiple lesions on a same image. With the FROC (Free Response Operating Characteristic) and AFROC (Alternative FROC) approaches it is possible to analyze which, and how many, locations reporting as sites of possible lesions as well as the likelihood rating assigned to each reported target possibility. The ordinate of FROC and AFROC curves plots the fraction of anomalies correctly detected and located at each rating criteria. The abscissa plots the average number of false positives in each image for the FROC curves or the fraction of images that contained at least one false positive (abscissa of ROC) for the AFROC curves [11][12].

The significant amount of theoretical development and practical application of ROC techniques, especially in the area of radiology, were facilitated to a large extent by the distribution of freely available ROC-analysis computer software [13][14][15]. ROC analysis is not the only method that can be used to evaluate diagnostic quality, it is the most rigorous and the most widely accepted. Other approaches also have been suggested and used to varying degrees. One such approach is an experiment in which pairs of images are presented side-by-side and the observer is asked to distinguish or rank-order each image [16]. In recent year, this approach has been used successfully to evaluate image compression techniques, in which observers’ ability to distinguish images compressed to varying degrees from the original (not-compressed) image is assessed. The rationale, based on the concept of just noticeable differences (JNDs), is that if the observer is not able to distinguish an compressed image reliably from its no-compressed original, then the image compression causes only “visually lossless” changes to the image and, therefore, the changes should not affect diagnostic performance [17][18][19]. This approach has been proposed as a way to plan for a large-scale ROC study, to decide whether a ROC study is justified.

3.3. Analyzing the quality of diagnostic criteria

The analysis of the diagnosis reliability is often rather “general”. This is because it refers to a set of test images. This analysis is also “unique” because it focuses on analyzing one specific pathology. In clinical practice, each diagnosis is very specific and may deal with more than one pathology. Thus, all local information (often anatomic information) in the image is analyzed according to specific criteria in order to best integrate the diagnostic process. Evaluating local attributes of an image is therefore an important part of the diagnosis that reflects the clinical practice.

For this type of analysis it is crucial to define both the grading scale and the diagnostic criteria that are to be evaluated. Moreover, they must be defined very precisely (as indicated in Figure 5) in order to avoid inaccuracies. The criteria selected are specific to the organ being studied. The notation scale is also adapted to the characteristics of the acquisition system.

### Details: Pulmonary vascular tree thinness

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Upper pulmonary vasculature</th>
<th>Lower pulmonary vasculature</th>
<th>Retrocardiac lung</th>
<th>Retrosternal lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scale</td>
<td>5 Very sharp and thin with a normal aspect</td>
<td>4 Thin</td>
<td>3 Moderate or badly zoned</td>
<td>2 Occluded or blunted</td>
</tr>
<tr>
<td></td>
<td>1 Indistinguishable</td>
<td>1 Indistinguishable</td>
<td>1 Indistinguishable</td>
<td>1 Indistinguishable</td>
</tr>
</tbody>
</table>

### Contours: Anatomic contours sharpness

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Edge of descending thoracic aorta</th>
<th>Pulmonary hilum</th>
<th>Carina, right and left main bronchi</th>
<th>Intercostal spurs</th>
<th>Vertical border</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scale</td>
<td>5 Very sharp and thin contour, taut muscle drawing</td>
<td>4 Taut contour, possible muscle drawing</td>
<td>3 Not complete muscle drawing</td>
<td>2 Hard muscle drawing</td>
<td>1 Impossible muscle drawing</td>
</tr>
</tbody>
</table>

### Face-shaped artifacts

<table>
<thead>
<tr>
<th>Level scale</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Absent</td>
<td></td>
</tr>
<tr>
<td>4 Hardly visible</td>
<td></td>
</tr>
<tr>
<td>3 Visible but does not disturb visibility of other criteria</td>
<td></td>
</tr>
<tr>
<td>2 Visible and disturb visibility of one or more criteria</td>
<td></td>
</tr>
<tr>
<td>1 Significant with one or more criteria unavailable</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5. Anatomic criteria defined in [20] for quality assessment of normal PA chest radiographs**
This notation of anatomic criteria useful to any diagnosis (defined as Diagnostic Quality Pattern or DQP in [21]) allows us to determine a local score (and thus a quality map) or a global score (an average), for every compression rate.

In order to analyze the diagnosis reliability and the quality of specific diagnostic criteria, we need to ask physicians to assess the diagnosis legibility on a particular set of images. In practice, such studies are laborious and difficult to conduct [1]. It is indeed very complicated to gather different images making up a representative sample of the studied pathology, and/or a sample enabling us to judge the result on each type of criteria given for the evaluated system, in order to make significant comparisons.

4. OBJECTIVE QUALITY ASSESSMENT

To overcome the above problems of subjective studies, objective quality assessment methods are carried out by calculating a value representing the local or global quality of an image using mathematical algorithms known as “objective quality criteria” and without any human intervention. This value must best reveal the visual appreciation of a human observer.

4.1. Simple signal-based metrics

Although objective quality parameters based on contrast such as PSNR / MSE do not correlate well with perceived quality measurement and have little relationship with diagnostic utility [22], they are still widely used to evaluate medical image quality [23].

4.2. Metrics based on a model of the HVS

Numerous studies have been carried out over the last three decades to develop quality assessment methods exploiting known characteristics of the human visual system (HVS) [24]. Because detection task is essential in the diagnostic process, models of human vision that predict human detection performance and those based on the concept of just noticeable differences (JNDs) appear to simulate physician’s performance closely [25]. The idea is to input two images (e.g., a gold standard reference image and an image in question) into the model which yields a JND map of the magnitude and spatial location of visible differences between the images. Various stages of the model simulate the main visual factors such as luminance adaptation, contrast sensitivity, spatio-frequency decomposition and masking effect [2]. These models often correlate well with human performance and have been used to evaluate imaging system components such as phosphor for CRT display [26], to analyze the possible impact of defective pixels on the perception of medical images [27], reconstruction algorithms for parallel MRI [25]… Nevertheless, those models, widely used for quality assessment of conventional images, are just useful to provide error visibility. This is a strong limitation in the context of medical imaging since it is just one step of the diagnosis pipeline leading to expert decision.

4.3. Model observers

To predict objectively, the value of a system in terms of classification task, several numerical observers several numerical observers have been proposed [24][28]. Model observers perform a task that takes a decision between two classes: one contains noise only (H1) and the other contains lesion plus noise (H2). The output of a model observer is a test statistic $\lambda = T(g)$, where $T(g)$ is the observer’s discriminant function of the image column vector $g$. Model observers differ by their discriminant functions.

4.3.1. Ideal Observer (IO)
The ideal observer has a likelihood-ratio discriminant function [24]: $T_{io}(g) = \Pr(g|H_2)/\Pr(g|H_1)$

Amongst all the model observers, the ideal observer achieves maximum AUC. For this reason it is optimal in the sense of AUC.

However, IO requires full knowledge of the probability density functions (PDFs) of the image for each hypothesis (which are mainly high-dimensional functions), including descriptions of the object and complete information regarding the imaging process and the noise statistics. And IO is usually a nonlinear function of the image data $g$. These make it difficult to be computed except for a few cases. In order to make sure that the IO performance can easily be computed, a lot of papers calculate it under some assumptions [29], e.g. background statistics are Gaussian while the signal properties (e.g. amplitude, orientation, shape, etc.) are known exactly.

4.3.2. Linear Observer (LO)

In view of the IO’s mathematical intractability, linear discriminant functions are considered. The general form of the corresponding model observer referred as Linear Observer (LO) is $T_{lo}(g) = w^T g$, where $w$ is an M×1 column vector usually called template and assumed to be real [289]. Different LOs are proposed by different templates. For example, the template of Non-PreWhitening Matched Filter (NPWMF) is the difference between the ensemble means of the lesion-absent and lesion-present image data.

4.3.3. Hotelling Observer (HO)

Hotelling Observer (HO) is one of the LOs, and its template is [24]:

$$w_{ho} = S^{-1} \gamma \overline{g}$$
Where \( \mathbf{g} = (g_1, g_2) \) and \( \mathbf{g}_i = (g_{2i-1}, g_{2i}) \).

\[
S_i = \frac{1}{2} [K_i + K_{i2}] \cdot \mathbf{K}_{i2} = \left( \mathbf{g}_i - \mathbf{\bar{g}}_i \right) \left( \mathbf{g}_i - \mathbf{\bar{g}}_i \right)^T H_i, i = 1, 2.
\]

Amongst all the linear observers, HO maximizes the signal-to-noise ratio (SNR) associated with the test statistic often called detectability index and referred to as \( d^* \) or \( d' \):

\[
d' = \sqrt{\frac{\mathbb{E}[\mathbb{V}(g_i H_i)] - \mathbb{E}[\mathbb{V}(g H_i)]}{\frac{1}{2} \text{var}[\mathbb{V}(g_i H_i)] + \frac{1}{2} \text{var}[\mathbb{V}(g H_i)]}}.
\]

Hence HO is optimal in the sense of SNR. In addition, HO needs far less information over the image statistics; it requires only the first- and second-order statistics of the image data under each hypothesis. Nevertheless, HO template calculation is computationally burdensome.

### 4.3.4. Channelized Hotelling Observer (CHO)

To overcome the above problems of HO, Channelized Hotelling Observer (CHO) tends to be used extensively in practice. Actually, CHO works just like HO does. The only difference is that HO discriminant function is applied on the channelized image \( \mathbf{v} \) in place of the original image \( g \) in the CHO process: \( \mathbf{v} = \mathbf{U}^T \mathbf{g} \), where \( \mathbf{U} \) is an M×P matrix, each column vector \( (\mathbf{u}_p, p = 1, 2, ..., P) \) of which represents one channel. Thus the output channelized image \( \mathbf{v} \) will be a P×1 vector. Then its test statistic becomes:

\[
T_{\text{CHO}}(\mathbf{v}) = \mathbf{w}_{\text{HO},x}^T \mathbf{v}
\]

Where

\[
\mathbf{w}_{\text{HO},x} = \mathbf{S}_{2,x}^{-1} (\mathbf{v} - \mathbf{\bar{v}}), \quad \mathbf{S}_{2,x} = \frac{1}{2} (\mathbf{K}_{1,x} + \mathbf{K}_{2,x})
\]

\[
\mathbf{K}_{i,x} = \mathbb{E}[\mathbf{v} - \mathbf{\bar{v}}] (\mathbf{v} - \mathbf{\bar{v}})^T H_i, i = 1, 2.
\]

We could immediately see a dimensionality reduction which leads to less calculation amount, because the value of \( P \) is much smaller than the value of \( M \) [29][30].

Channel selection is an important step in the CHO, and it depends on the objective. For example, CHO based on the Laguerre-Gauss (LG) channels approximates the HO well in terms of SNR under the condition of lumpy backgrounds and circularly symmetric Gaussian signals [31].

Initially, the model observers are applied only on the 2D image. But in the clinic, we usually have access to an entire image volume presented as multiple slices in multiple views. Because single-slice format lacks sufficient anatomical context for an efficient diagnosis, model observers were developed to solve the multi-slice problem. One of the most recent works applies 2D CHO on each slice and then uses HO to combine the test statistic of each slice together to get a final test statistic [32].

While all the above observers have been designed keeping in mind optimal decision, it might be interesting to use such formalism in order to model human observer performance while performing decision tasks. Channelized model observer’s framework presents interesting features that could incorporate HVS properties. The first HVS property considered with the model observer is the frequency- and orientation- selective mechanism, which is commonly approximated by the assumption of discrete channels, like Gabor channels, Wavelet channels and Cortex channels [33]. For instance, [34] added a frequency-selective mechanism to the IO model; [29] integrated a simple HVS orientation-selectivity model (the steerable oriented channels) in the CHO. Their results show that this type of model observer does not overestimate the human performance as much as pure model observer.

In order to predict the human performance better and better, it’s natural to think about combining more and more HVS properties. A recent example is [35], which uses NPWMF with one of the most famous HVS models - Sarnoff Human Vision Discrimination Model (HVM) [36], which models not only frequency- and orientation- selective mechanism, but also eye optics, retinal sampling, luminance nonlinearity. More recently, [37] presents a CHO with a model of contrast gain control in human vision that incorporates contrast sensitivity function, multiple oriented bandpass channels, accelerating nonlinearities, and a divisive inhibitory gain control pool [38]. They showed that their HVS-based model observers do predict the human performances better.

However, the literature rarely explained and considered which HVS characteristics should be taken into account and which HVS model should be used. This type of model observers needs further studies and more efforts.

### 5. EYE-TRACKING STUDIES

Another experimental and fundamental tool to better understand and characterise the diagnosis pipeline is eye-position recording (see Figure 6). Eye-tracking studies have been used to gain basic understanding of the visual search and decision-making process [39][40] and also for system evaluation [41].

![Figure 6. Example of a typical eye-position pattern generated by an experienced radiologist. The circles represent gaze fixations. The lines represent saccades. The task of this study was to identify each elementary task implemented during diagnostic process of glioma detection and characterization on brain MRI] (39)
Eye-tracking studies are useful in general to understand how an imaging system affects interpretation efficiency and the decision-making process. Whereas ROC analysis assesses the final decision, eye-tracking studies provide information on how the observer reaches that decision. A major assumption behind eye-tracking studies is that the amount of time spent looking at features in the image reflects information processing, object encoding and recognition. By correlating eye-position parameters such as dwell time, number of returns to a location, and saccade length with image interpretation. Then we can hope to optimize image quality and image presentation to better match with the human eye-brain system. We can also hope to develop computer-aided diagnosis tools to assist the radiologist when their perceptual or cognitive abilities tend to fail (in detecting subtle or partially obscured lesions).

For example, in our last study [39], we investigated the use of an eye-tracking system to explore relationships between visual scanning patterns and the glioma diagnostic process during brain MRI analysis. The analysis of saccadic amplitudes reveals clear delineation of three sequential steps. During the first step (characterized by large saccades), a radiologist performs a short review on all sequences and on the patient report. In the second step (characterized by short saccades), a radiologist sequentially and systematically scans all the slices of each sequence. The fixation duration in one AOI depends on the number of slices, on the lesion subtlety and on the lesion contrast in the sequence to be analyzed. In order to improve the detection, localization and characterization of the glioma, the radiologist compares sequences during the third step (characterized by large saccades). This result enables one to understand the elementary tasks radiologists accomplish to analyze medical images. Moreover, fixation maps enable one to assign weights to lesion areas according to their subtlety and the sequence to be analyzed. Very little objects, such as fine linear details, or granulated textures are important facts to help differential diagnosis between high grade gliomas and other brain tumors, or to appreciate the shape of a suspected multiple sclerosis high signal area. Such results suggest criteria for optimization of image compression rates (depending on specific diagnostic features and lesion area). This knowledge is critical to the design of image viewing devices and picture archiving and communication system (PACS), especially for image compression. Moreover, perceptual feedback may be advantageous for radiologist training. The recorded scanpaths of experts can be used for teaching novice neuroradiologists the best exploration strategies in a very intuitive manner [40]. If we can understand visual search of medical images, we may be able to explain why errors of diagnosis occur and what can be done.

6. CONCLUSION

Diagnostic quality assessment of medical images is a multifaceted process that can be approached from a variety of perspectives. Nevertheless, the researchers have to keep in mind that in medical imaging, clinicians need not the most “beautiful” image, but rather the image that allows them to render the most accurate and timely interpretation.

Studies of medical image perception and the general interaction of clinicians with medical imaging examinations remains a critical element of improving health care. Continued investigation of this complex perceptual recognition and interpretation process will be needed to offer the most useful and effective presentation of imaging information to physicians and to improve their detection and classification of disease.

7. ACKNOWLEDGMENTS

Some results presented in this paper are supported by the “Région des Pays de La Loire” through the regional French project EQuIMO (Subjective and objective Evaluation of the Quality of Medical Images for an Optimal use of the display, archiving and transmission Systems).

8. REFERENCES
