34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Studying Disagreements among Retinal Experts through Image Analysis

Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener, Zakarya Droueche, Bruno Lay, Agnès Chabouis, Christian Roux, Guy Cazuguel

***Abstract*— In recent years, many image analysis algorithms have been presented to assist Diabetic Retinopathy (DR) screening. The goal was usually to detect healthy examination records automatically, in order to reduce the number of records that should be analyzed by retinal experts. In this paper, a novel application is presented: these algorithms are used to 1) discover image characteristics that sometimes cause an expert to disagree with his/her peers and 2) warn the expert whenever these characteristics are detected in an examination record. In a DR screening program, each examination record is only analyzed by one expert, therefore analyzing disagreements among experts is challenging. A statistical framework, based on Parzen-windowing and the Patrick-Fischer distance, is presented to solve this problem. Disagreements among eleven experts from the Ophdiat screening program were analyzed, using an archive of 25,702 examination records.**

I. INTRODUCTION

Diabetic Retinopathy (DR) is the leading cause of blindness in the working population of the European Union and the United States [1]. Because early detection and timely treatment of DR can prevent visual loss and blindness in patients with diabetes, several DR screening programs have been initiated in recent years [2], [3], [4]. As a consequence, large archives of DR examination records, each containing several eye fundus photographs, are available. In this paper, these image archives are used to study disagreements among retinal experts. If we can detect image characteristics that cause an expert to disagree with his/her peers, then a warning can be raised whenever these characteristics are found in an examination record. This personalized decision support tool does not imply an additional workload for retinal experts.

In order to discover these characteristics, we propose to project all stored examination records into a common image analysis space. Many image analysis algorithms have been presented in recent years to detect the early signs of DR [5], [6] of to assess image quality [7], [8]. We propose to associate each dimension in image analysis space with one of these algorithms. Then, using all examination records analyzed by one expert, as well as the decisions he/she made for each record, the decisions of this expert can be modeled at each location in image analysis space. Note that, in a DR screening program, the main decision experts have to make is whether or not a patient should be referred to an ophthalmologist for further examinations, treatment, etc. So we focused on these decisions. Finally, once a decision model is available for each expert, disagreements among experts can be studied through a comparative analysis of their decision models.

G. Quellec, M. Lamard, B. Cochener, Z. Droueche, C. Roux and G. Cazuguel with Inserm, UMR 1101, Brest, F-29200 France gwenole.quellec@inserm.fr

M. Lamard and B. Cochener are with Univ Bretagne Occidentale, Brest, F-29200 France

B. Cochener is with CHRU Brest, Service d'Ophtalmologie, Brest, F-29200 France

Z. Droueche, G. Cazuguel and C. Roux are with Institut Mines-Telecom; Telecom Bretagne; UEB; Dpt ITI, Brest, F-29200 France

B. Lay is with ADCIS, Saint-Contest, F-14280 France

A. Chabouis is with Hôpital Lariboisière - APHP, Service d'Ophtalmologie, Paris, F-75475 France

II. IMAGE ANALYSIS SPACE

In this section, we present how a d-dimensional image analysis space $\mathbb{A}$ was designed ($d = 6$). The first two dimensions were associated with standard lesion detection algorithms (§II-B), the third dimension was associated with a weakly-supervised anomaly detector (§II-C) and the last three dimensions were associated with image quality metrics (§II-D).

A. Training set

A training set of examination records was used to design the image analysis space: $\mathcal{T} = \{R_i, i = 1..N\}$. Each examination record $R_i \in \mathcal{T}$ was associated with a binary decision $\delta_i \in$ {'non-referable patient', 'referable patient'} assigned by one retinal expert.

B. Lesion Detection Dimensions

The first two dimensions of $\mathbb{A}$, namely a_1 and a_2, were associated with lesion detection algorithms: a_1 was associated with a microaneurysm detector D_1 [9], a_2 was associated with an exudate detector D_2 [6].

For each lesion detector D_k, $k \in \{1, 2\}$, a set of $l_{k,i}$ lesion candidates was detected in each examination record $R_i \in \mathcal{T}$: $\mathcal{L}_{k,i} = \{(I_j, x_j, y_j, s_j, p_j), j = 1..l_{k,i}\}$. Lesion candidate $(I_j, x_j, y_j, s_j, p_j)$ is a connected component of s_j pixels detected in image $I_j \in R_i$, at location (x_j, y_j), with probability p_j.

To define the k^{th} dimension of $\mathbb{A}$, each set $\mathcal{L}_{k,i}$ was converted into a single number, $a_{k,i} \in \mathbb{R}$, as explained below:

1) F_k, the joint Cumulated Distribution Function (CDF) of all (s_j, p_j) tuples in the training set ($j = 1..l_{k,i}, R_i \in \mathcal{T}$) was estimated,
2) for each examination record $R_i \in \mathcal{T}$, $H_{k,i}$, a b-bin histogram of $\{F_k(s_j, p_j), j = 1..l_{k,i}\}$ was built ($b \in \mathbb{N}^*$),
3) for each examination record $R_i \in \mathcal{T}$, each bin in $H_{k,i}$ was divided by $|R_i|$, the number of images in R_i,
4) a linear discriminant analysis [10] was performed in $\mathbb{R}^b$ to best separate referable patients from non-referable

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

patients (according to the δ_i decisions) ; let w_k denote the normal to the discriminant hyperplane,

5) for each examination record $R_i \in \mathcal{T}$, $a_{k,i}$ was obtained by projecting $H_{k,i}$ onto w_k:

$$a_{k,i} = w_k \cdot H_{k,i}, \quad k \in \{1, 2\} \tag{1}$$

C. Anomaly Detection Dimension

The following dimension of $\mathbb{A}$, namely a_3, was associated with a weakly-supervised anomaly detector D_3 [11]. Unlike the above microaneurysm and exudate detectors, D_3 computes $a_{3,i}$ directly from an examination record $R_i \in \mathcal{T}$, without segmentation steps.

Given annotated examination records $(R_i, \delta_i) \in \mathcal{T} \times \{$'non-referable patient', 'referable patient'$\}$, D_3 was trained to detect patterns of arbitrary size that only appear in images $I \in R_i$ such that $\delta_i =$'referable patient'. D_3 was trained as follows:

1) each image I in the training set ($I \in R_i, R_i \in \mathcal{T}$) was divided into sub-images $J \subset I$ of various sizes,
2) each sub-image was characterized with texture, color and shape features [12], [13],
3) the n nearest neighbors of each sub-image $J \subset I$ were searched in $\mathcal{T} \setminus I$ ($n \in \mathbb{N}^*$),
4) using the percentage of nearest neighbors coming from patient records R_j such that $\delta_j =$'referable patient', an local anomaly index $\alpha(J)$ was computed for J,
5) the distance measure between sub-image characterizations was updated in order to reduce false alarms [11],
6) steps 3) to 5) were repeated until convergence,
7) for each examination record $R_i \in \mathcal{T}$, $a_{3,i}$ was computed using all local anomaly indices computed for sub-images of R_i (see step 4)):

$$a_{3,i} = \sqrt{\frac{1}{|R_i|} \sum_{I \in R_i, J \subset I} |\alpha(J)|^2} \tag{2}$$

Note that anomalies detected by D_3 are typically larger than isolated microaneurysms or exudates. They include exudate clusters, large hemorrhages, intraretinal microvascular abnormalities, nevi, etc.

D. Image Quality Dimensions

The last dimensions of $\mathbb{A}$, namely a_4, a_5 and a_6, were associated with three image quality metrics based on mathematical morphology [14]:

- $Q_4(I)$ = the average intensity in the morphological gradient of I (note that gradient images are often used by autofocus optical systems),
- $Q_5(I)$ = the average intensity in the residual image obtained after alternate sequential filtering of I (this image contains all dark patterns of I, including blood vessels),
- $Q_6(I)$ = the average intensity in the same residual image, after removing blood vessels through mathematical morphology.

Low $Q_k(I)$ values, $k \in \{4, 5, 6\}$, are supposedly associated with low quality images. Therefore, $a_{k,i}$, the projection of an examination record R_i onto a_k, was defined as the minimal $Q_k(I)$ value in R_i:

$$a_{k,i} = \min_{I \in R_i} Q_k(I), \quad k \in \{4, 5, 6\} \tag{3}$$

E. Test set

$\mathbb{A}$, the d-dimensional image analysis space, was designed using a training set $\mathcal{T}$ of examination records (§II-B, II-C). Before proceeding to the next step, each examination record in a test set $\overline{\mathcal{T}}$ was also projected into $\mathbb{A}$.

III. DISAGREEMENT MAPS IN IMAGE ANALYSIS SPACE

In this section, we describe how a *disagreement map* was built for each retinal expert in image analysis space $\mathbb{A} = \mathbb{R}^d$ (§III-C). In that purpose, we first modeled the decisions of each expert individually, as well as the decisions of the group as a whole, through the design of *decision maps* (§III-B).

Let M denote the number of retinal experts. Let X_e denote one expert, $e = 1..M$. To model the decisions of each expert individually, the test set $\overline{\mathcal{T}}$ was partitioned into $2M$ sets of examination records: $\mathcal{X}_e^-$ and $\mathcal{X}_e^+$, $e = 1..M$. Subset $\mathcal{X}_e^-$ (respectively $\mathcal{X}_e^+$) contains all examination records $R_i \in \overline{\mathcal{T}}$ analyzed by expert X_e such that $\delta_i =$'non-referable patient' (respectively $\delta_i =$'referable patient'). To model the decisions of the group as a whole, the test set $\overline{\mathcal{T}}$ was also partitioned into 2 sets of examination records: $\mathcal{X}^- = \{R_i \in \overline{\mathcal{T}} : \delta_i =$'non-referable patient'$\}$ and $\mathcal{X}^+ = \{R_i \in \overline{\mathcal{T}} : \delta_i =$'referable patient'$\}$.

To build a decision map for expert X_e, we estimated f_e^- and f_e^+, the Probability Density Function (PDF) of the d-dimensional distributions from which $\mathcal{X}_e^-$ and $\mathcal{X}_e^+$ were drawn, respectively. Assuming each subset is an independent and identically distributed sample, the associated PDF was estimated using the Parzen-window method [15]. Similarly, PDFs f^- and f^+ were estimated from subsets $\mathcal{X}^-$ and $\mathcal{X}^+$, respectively.

A. Parzen-Window Density Estimation

Parzen-windowing is a non-parametric PDF estimation method. Originally defined for one-dimensional data, it was extended to any d-dimensional space by Murthy [16]. In the Parzen-window method, the estimate PDF $\hat{f}$ of a subset $\mathcal{X}$ is given by the following formula:

$$\begin{aligned} \hat{f}: \quad \mathbb{A} &\to \mathbb{R} \\ x &\mapsto \hat{f}(x) = \frac{1}{|\mathcal{X}|} \sum_{x' \in \mathcal{X}} \frac{1}{h(|\mathcal{X}|)^d} K\left(\frac{x - x'}{h(|\mathcal{X}|)}\right) \end{aligned} \tag{4}$$

where K is a weighting function (or kernel function) and h is a positive valued function of $|\mathcal{X}|$ that tends to 0 as $|\mathcal{X}|$ increases. The following functions were used for K and h: the Gaussian kernel and the Koontz's function [17], respectively; these choices are standard. A one-dimensional illustration is given in figure 1.

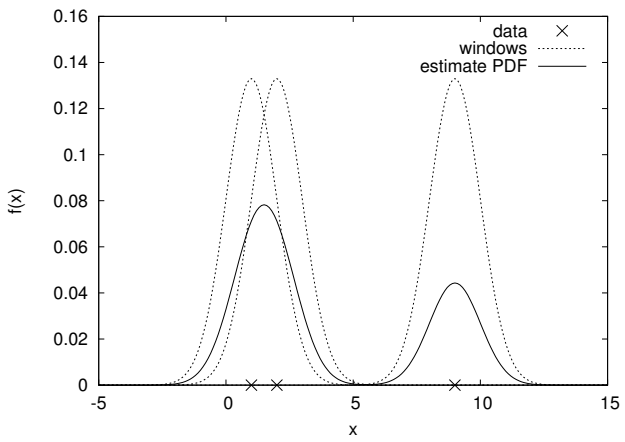


Fig. 1: 1-D Parzen-windowing: the estimate PDF is the average of the $|\mathcal{X}| = 3$ data-centered windows ($h(|\mathcal{X}|) = 1$).

B. Decision maps

Then, a decision map $\Phi_e \colon \mathbb{A} \to \mathbb{R}^2$ was defined for expert X_e as follows:

$$\begin{cases} \Phi_e(x) = (\Phi_e^-(x), \Phi_e^+(x)) \\ \Phi_e^-(x) = |\mathcal{X}_e^-| \hat{f}_e^-(x) \\ \Phi_e^+(x) = |\mathcal{X}_e^+| \hat{f}_e^+(x) \end{cases} \quad (5)$$

where $\hat{f}_e^-$ and $\hat{f}_e^+$ are the estimate PDFs obtained through Parzen-windowing (§III-A). Similarly, a collective decision map $\Phi \colon \mathbb{A} \to \mathbb{R}^2$ was defined for the group:

$$\begin{cases} \Phi(x) = (\Phi^-(x), \Phi^+(x)) \\ \Phi^-(x) = |\mathcal{X}^-| \hat{f}^-(x) \\ \Phi^+(x) = |\mathcal{X}^+| \hat{f}^+(x) \end{cases} \quad (6)$$

C. Disagreement Maps

Then, a disagreement map $\Delta_e \colon \mathbb{A} \to \mathbb{R}$ was defined for expert X_e as follows:

$$\Delta_e(x) = \frac{\Phi_e^+(x)}{\Phi_e^-(x) + \Phi_e^+(x)} - \frac{\Phi^+(x)}{\Phi^-(x) + \Phi^+(x)} \quad (7)$$

A positive (respectively negative) $\Delta_e(x)$ value indicates that expert X_e is overly sensitive (respectively specific) in the neighborhood of $x \in \mathbb{A}$.

D. Visualizing Decision and Disagreement Maps

Let us remind that $\mathbb{A}$ is a 6-dimensional space ($d = 6$ - §II). Therefore, dimension reduction, from $\mathbb{A} = \mathbb{R}^d$ to $\mathbb{R}^2$, is necessary for visualization purposes. We propose to find the 2-dimensional linear subspace of $\mathbb{A}$, defined by a linear transformation $L \in \mathbb{M}(6, 2)$, where the two binary decisions $\delta =$'non-referable patient' and $\delta =$'referable patient' are most separated.

Let $\mathcal{X}^- = \{R_i \in \mathcal{T} : \delta_i =$'non-referable patient'$\}$ and $\mathcal{X}^+ = \{R_i \in \mathcal{T} : \delta_i =$'referable patient'$\}$ denote a partition of the training set with respect to the binary decisions. The optimal linear transformation L is obtained by maximizing the Patrick-Fischer distance [18] between $\hat{f}^-$ and $\hat{f}^+$, the estimate PDF of the marginal distributions from which $L^T \cdot \mathcal{X}^-$ and $L^T \cdot \mathcal{X}^+$ were drawn, respectively. To facilitate the interpretation of L, each dimension of $\mathbb{A}$ was normalized (average=0, standard deviation=1) before maximizing the Patrick-Fischer distance.

IV. OPHDIAT ARCHIVE

A digital archive containing all examination records collected in the Ophdiat screening network[1] during two consecutive years (2008 and 2009) was used in this paper. Ophdiat consists of 29 DR screening centers in the Parisian area. 25,702 examination records were collected by trained technical staff and submitted to a remote server. Then, each examination record was analyzed by one retinal expert, out of 11 participating experts, in Lariboisière Hospital (Paris, France).

Besides demographic and biological data, each examination record contains two eye fundus photographs per eye on average: one centered on the fovea and one centered on the optic disk. Images were obtained with non-mydriatic retinographs: either CR-DGi (Canon, Tokyo) or TRC-NW6S (Topcon, Tokyo) retinographs. Depending on the settings of each retinograph, images with varying sizes were obtained: image sizes ranged from 1440×960 to 2544×1696 pixels. To ease the task of the image analysis algorithms, all images were automatically resized and cropped to a definition of 780×780 pixels (see Fig. 3). Overall, 107,799 images were collected.

In each record, one expert indicated whether or not the patient should be referred to an ophthalmologist for further examinations, treatment, etc. Normally, patients are referred to an ophthalmologist when they have DR, or another pathology, in at least one eye (one exception: mild nonproliferative DR does not trigger a referral). Referral was decided for 6,391 records (prevalence: 25%).

This archive was partitioned into a training subset $\mathcal{T}$ and a testing subset $\overline{\mathcal{T}}$ at random: $|\mathcal{T}| = |\overline{\mathcal{T}}| = 12,851$ records.

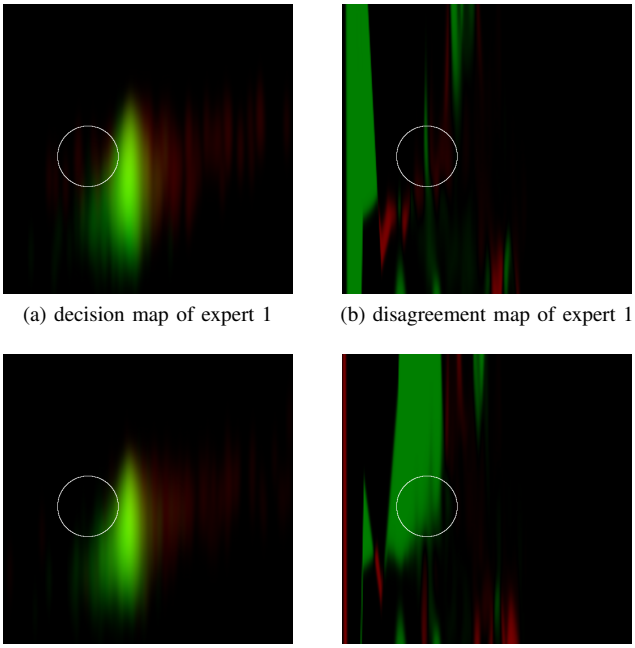
V. RESULTS

The following parameters were used to build the image analysis space: $b = 8$ bins (§II-B), $n = 5$ neighbors (§II-C). The following linear transformation L was found optimal:

$$L^T = 10^{-2} \begin{pmatrix} -4.79 & 1.18 & 1.01 & 11.4 & -74.3 & 65.7 \\ 96.6 & 1.51 & -25.0 & 1.86 & -6.10 & 0.173 \end{pmatrix} \quad (8)$$

The first dimension in the reduced image analysis space (first row in L^T) is mostly related to image quality: the most important input dimension is a_5 (alternate sequential filtering). The second dimension is mostly related to pathology detection: the most important input dimension is a_1 (microaneurism detection). Fig. 2 displays the decision and disagreement maps of two retinal experts (out of eleven) in this space. One can see, in Fig. (a) and (c), that the second dimension (in columns) clearly is more correlated with the decisions assigned by experts to each

[1] http://reseau-ophdiat.aphp.fr



(a) decision map of expert 1 (b) disagreement map of expert 1

(c) decision map of expert 2 (d) disagreement map of expert 2

Fig. 2: Decision and Disagreement maps. In decision maps, the first component (associated with 'non-referable patient' decisions) is in the green channel and the second component (associated with 'referable patient' decisions) is in the red channel. In disagreement maps, negative values (indicating that the expert is overly specific) are in green and positive values (indicating that he/she is overly sensitive) are in red.

Fig. 3: Example of image where expert 2 is overly specific.

examination record. However, it can be seen, in Fig. (b) and (d), that the first dimension (in rows) has an influence on the disagreements among experts: image quality influences disagreements among experts. The white circle in Fig. 2 indicates a region in image analysis space where expert 2 is overly specific: it corresponds to low quality images where no abnormalities are visible (see Fig. 3).

VI. CONCLUSION

In this paper, some disagreements among retinal experts were discovered in a feature space generated by existing image analysis algorithms. Even though each examination record was only seen by one expert, the use of Parzen-windowing (see Fig. 1) in this space allowed us to compare the decisions of several experts. In order to study the disagreements among retinal experts further in future works, demographic information (age, duration of diabetes, etc.), as well as other retinal image analysis algorithms (vessel tortuosity assessment, cup-to-disk ratio assessment, etc.), will be included.

ACKNOWLEDGMENT

This work was supported in part by the French *Agence Nationale de la Recherche* (ANR), within the framework of the Teleophta project (see http://www.teleophta.fr).

REFERENCES

[1] D. Klonoff and D. Schwartz, "An economic analysis of interventions for diabetes," *Diabetes Care*, vol. 23, no. 3, pp. 390–404, 2000.

[2] M. D. Abràmoff and M. S. A. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project," *Telemed J E Health*, vol. 11, no. 6, pp. 668–674, 2005.

[3] S. Philip, A. D. Fleming, and K. A. Goatman et al., "The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme," *Br J Ophthalmol*, vol. 91, no. 11, pp. 1512–1517, 2007.

[4] P. Massin, A. Chabouis, A. Erginay, C. Viens-Bitker, A. Lecleire-Collet, and T. Meas et al., "OPHDIAT: a telemedical network screening system for diabetic retinopathy in the Ile-de-France," *Diabetes Metab*, vol. 34, no. 3, pp. 227–234, 2008.

[5] M. Niemeijer, B. van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, and C. I. Sanchéz et al., "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans Med Imaging*, vol. 29, no. 1, pp. 185–195, 2010.

[6] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Med Image Anal*, vol. 16, no. 1, pp. 216–226, 2012.

[7] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated assessment of diabetic retinal image quality based on clarity and field definition," *Invest Ophthalmol Vis Sci*, vol. 47, no. 3, pp. 1120–1125, 2006.

[8] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening," *Med Image Anal*, vol. 10, no. 6, pp. 888–898, 2006.

[9] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Trans Med Imaging*, vol. 27, no. 9, pp. 1230–1241, 2008.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[11] G. Quellec, M. Lamard, G. Cazuguel, M. D. Abràmoff, B. Cochener, and C. Roux, "Weakly supervised classification of medical images," in *Proc IEEE Int Symp Biomed Imaging*, 2012, in press.

[12] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Med Image Anal*, vol. 14, no. 2, pp. 227–241, 2010.

[13] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Trans Pattern Anal Mach Intell*, vol. 12, no. 5, pp. 489–497, 1990.

[14] J. Serra, Ed., *Image Analysis and Mathematical Morphology - Vol. II : Theoretical Advances*. Academic Press, 1988.

[15] E. Parzen, "On estimation of a probability density function and mode," *Ann Math Stat*, vol. 33, no. 3, pp. 1065–1076, 1962.

[16] V. K. Murthy, "Nonparametric estimation of multivariate densities with applications," in *Multivariate analysis*, P. R. Krishnaiah, Ed. Academic Press, 1966, pp. 43–56.

[17] W. L. G. Koontz and K. Fukunaga, "Asymptotic analysis of a nonparametric clustering technique," *IEEE Trans Comput*, vol. C-21, no. 9, pp. 967–974, 1972.

[18] E. A. Patrick and F. P. Fischer II, "Nonparametric feature selection," *IEEE Trans Inform Theory*, vol. IT-15, no. 5, pp. 577–584, 1969.