A Multiple-Instance Learning Framework for Diabetic Retinopathy Screening

Gwénolé Quellec, Mathieu Lamard, Michael D Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, Guy Cazuguel

A multiple-instance learning framework for diabetic retinopathy screening

Gwénolé Quellec [a,*], Mathieu Lamard [b,a], Michael D. Abràmoff [c], Etienne Decencière [d], Bruno Lay [e], Ali Erginay [f], Béatrice Cochener [b,a,g], Guy Cazuguel [h,a]

[a] *Inserm, UMR 1101, SFR ScInBioS, Brest F-29200, France*
[b] *Univ Bretagne Occidentale, Brest F-29200, France*
[c] *Departments of Ophthalmology and Visual Sciences, Electrical and Computer Engineering, and Biomedical Engineering, University of Iowa, Iowa City, IA 52242, USA*
[d] *Centre for Mathematical Morphology, MINES ParisTech, ARMINES, Fontainebleau F-77300, France*
[e] *ADCIS, Saint-Contest F-14280, France*
[f] *Service d'Ophtalmologie, Hôpital Lariboisière, APHP, Paris F-75475, France*
[g] *Service d'Ophtalmologie, CHU Brest, Brest F-29200, France*
[h] *INSTITUT TELECOM, TELECOM Bretagne, UEB, Dpt ITI, Brest F-29200, France*

A novel multiple-instance learning framework, for automated image classification, is presented in this paper. Given reference images marked by clinicians as relevant or irrelevant, the image classifier is trained to detect patterns, of arbitrary size, that only appear in relevant images. After training, similar patterns are sought in new images in order to classify them as either relevant or irrelevant images. There-fore, no manual segmentations are required. As a consequence, large image datasets are available for training. The proposed framework was applied to diabetic retinopathy screening in 2-D retinal image datasets: Messidor (1200 images) and e-ophtha, a dataset of 25,702 examination records from the Oph-diat screening network (107,799 images). In this application, an image (or an examination record) is rel-evant if the patient should be referred to an ophthalmologist. Trained on one half of Messidor, the classifier achieved high performance on the other half of Messidor ($A_z = 0.881$) and on e-ophtha ($A_z = 0.761$). We observed, in a subset of 273 manually segmented images from e-ophtha, that all eight types of diabetic retinopathy lesions are detected.

1. Introduction

The continually increasing amount of images stored in medical databases makes data-driven knowledge acquisition (i.e. machine learning) more and more attractive. Unfortunately, many image-based machine learning techniques require intensive interactions with clinicians for supervision and evaluation purposes (typically, through manual segmentation of regions of interest): this is a serious bottleneck. Therefore, extraction of a limited subset of the available data, for clinician interpretation and algorithm supervision, is common practice (this subset is referred to as the reference dataset). A different approach is explored in this paper: we think that similar or better detection performance can be achieved through limited clinician interpretation, so long as a larger reference dataset is available. In this paper, an extreme scenario is considered: clinician interpretation is limited to a Boolean label per image. Given a target concept (e.g. presence of a pathology), clinicians are simply asked to indicate whether or not each image in the reference dataset contains at least one image pattern indicating that the pathology is present. Obviously, if the burden of clinician interpretation is significantly reduced, clinicians would be willing to interpret more images. Moreover, assigning Boolean interpretations to medical images is often part of the clinical protocol, so after deidentification, images may be used directly for research without additional work for clinicians.

One medical application where many images are available is Diabetic Retinopathy (DR) screening. DR is the leading cause of blindness in the working population of the European Union and the United States (Klonoff and Schwartz, 2000). 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 (Abràmoff and Suttorp-Schulten, 2005; Philip et al., 2007; Massin et al., 2008; Li et al., 2011). As a consequence, large datasets of digitized eye fundus photographs have been collected. One of these datasets, Messidor[1] (1200 images), is freely available. However, Messidor images have not been manually segmented, so they cannot be used to supervise most state-of-the-art DR detection methods (Walter et al., 2007; Philip et al., 2007; Chaum

* Corresponding author. Address: LaTIM, Bâtiment 2bis (I3S), CHU Morvan, 5, Av. Foch, 29609 Brest Cedex, France. Tel.: +33 2 98 01 81 29; fax: +33 2 98 01 81 24.
E-mail address: gwenole.quellec@inserm.fr (G. Quellec).

[1] http://messidor.crihan.fr/index-en.php.

Table 1
Glossary.

Reference image	Image that has been categorized by clinicians and that is now used to categorize new images by analogy reasoning
Irrelevant image	Reference image categorized as irrelevant
Relevant image	Reference image categorized as relevant
Reference dataset	Dataset of reference images
Image patch/patch	Rectangular subset of pixels in an image
Image feature	Individual measurable property of a phenomenon being observed in an image patch
Signature	D-dimensional vector of image features
Reference signature	Signature of a patch in a reference image
Irrelevant signature	Signature of a patch in an irrelevant image
Possibly-relevant signature	Signature of a patch in a relevant image
Local relevance score	Real number indicating the risk that a patch is relevant
Global relevance score	Real number indicating the risk that an image is relevant

et al., 2008; Niemeijer et al., 2009; Sánchez et al., 2010; Giancardo et al., 2011; Antal et al., 2011; Quellec et al., 2011b; Oliveira et al., 2011); the readers are referred to Niemeijer et al. (2010) for an overview of the DR detection methods involved in the ROC international challenge.[2] Only smaller image datasets come with manual lesion segmentations: Kauppi et al.'s (2007) DIARETDB1 (89 images), Niemeijer et al.'s (2010) ROC (100 images), Giancardo et al.'s (2012) HEI-MED (169 images). Manual-segmentation-free training algorithms are therefore needed to make the most of larger datasets such as Messidor. The proposed framework is also applied to e-ophtha, an even larger dataset (25,702 examination records, 107,799 images) collected in the Ophdiat screening network (Massin et al., 2008).

In this paper, a general framework for automatically finding relevant patterns in images is presented. Its purpose is to automatically define (1) a *local relevance score* for patches of varying sizes in reference images (see Table 1) and, at the end of the training phase, (2) a local relevance score for patches in new images not included in the training set. Once a local relevance score is defined for each patch, a *global relevance score* is defined for the entire image, using a novel multi-resolution fusion strategy.

The proposed framework fits into the Multiple-Instance Learning (MIL) paradigm. Multiple-instance learning is a type of supervised learning which deals with uncertainty of instance labels (Maron and Lozano-Pérez, 1998). Standard supervised learners receive a set of instances which are labeled relevant or irrelevant. Multiple-instance learners, on the other hand, receive a set of *bags of instances* that are labeled relevant or irrelevant. Each bag may contain many instances. A bag is labeled irrelevant if all the instances in it are irrelevant. On the other hand, a bag is labeled relevant if there is at least one instance in it which is relevant. Uncertainty lies in the fact that the learner does not know which instances in the relevant bags are relevant. From a collection of labeled bags, the multiple-instance learner tries to induce a concept that will label individual instances correctly. If each image is regarded as a bag and each patch in an image is regarded as an instance, then our problem can be formulated as an MIL problem. An overview of Multiple-Instance Learning, in the context of image analysis, is given in Section 3. In this paper, we propose to solve this MIL problem using content-based image retrieval, a paradigm presented in the following section.

2. An introduction to content-based image retrieval

Content-Based Image Retrieval (CBIR) is the application of computer vision techniques to the problem of searching for digital images in large databases (Smeulders et al., 2000; Datta et al., 2008). In CBIR, each image is represented by a *signature*, i.e. a vector of visual features (e.g. color, texture, shape features). Comparing two images amounts to comparing their signatures using a *distance metric*, or conversely a *similarity metric*, between signatures. Based on these metrics, the nearest neighbors of an input image are retrieved using efficient *search algorithms*. CBIR is popular in medical applications (Müller et al., 2004). In particular, it has been applied to the automatic diagnosis of retinal pathologies in eye fundus images (Chaum et al., 2008; Quellec et al., 2011a).

The central challenge CBIR systems have to face is bridging the semantic gap between low-level visual features and the high-level concept of semantic similarity (Smeulders et al., 2000; Datta et al., 2008). Several solutions have been proposed to bridge this gap. One solution is relevance feedback (Ramaswamy et al., 2009; Azimi-Sadjadi et al., 2009): (1) users are asked to assess the relevance of the retrieved images, (2) the similarity metric is modified accordingly, (3) new images are retrieved and (4) this process is iterated until convergence. A second solution is multiple-example query (Donald and Smeaton, 2005; Zhang and Le, 2009): users are asked to provide several images representing a target concept and the system tries to generalize. A third solution is off-line training: the similarity metric is trained in a dataset of images previously annotated by experts (Quellec et al., 2010a,).

In some CBIR systems, a *bag* of signatures is extracted from each image: each signature in this bag is associated with a patch in the input image. Similarity metrics and search algorithms are modified accordingly (Rubner et al., 2000; Ko and Byun, 2002; Sivic and Zisserman, 2003). In such systems, MIL is one solution to bridge the semantic gap (Rahmani et al., 2008). Note that the opposite is done in this paper: CBIR is used to solve an MIL problem. The final goal is not to retrieve similar images but to categorize each input image as relevant or irrelevant.

3. Medical image analysis based on Multiple-Instance Learning (MIL)

3.1. A state of the art of medical image analysis based on MIL

To solve general MIL problems, Maron and Lozano-Pérez (1998) proposed *Diverse Density*. *Diverse Density* measures the intersection of the relevant bags minus the union of the irrelevant bags. By maximizing *Diverse Density*, the optimal set of image feature weights, for a given concept, is found. This approach was applied to image retrieval (Maron and Ratan, 1998): image features are extracted from fixed-size patches (square-shaped patches, square-shaped patches with their 4-connexity neighbors, etc.). An alternative MIL formulation was proposed by Andrews et al. (2003): a Support-Vector Machine (SVM) processes the signature labels as unobserved integer variables, subjected to constraints defined by the bag labels. The goal is to maximize the soft-margin over hidden label variables and a discriminant function. DD-SVM, an image categorization system proposed by Chen et al., extends the *Diverse Density* framework and Andrew's SVM framework in the case of segmented regions of arbitrary shape instead of square-shaped patches (Chen et al., 2004). ACCIO!, another MIL-based image retrieval system, was recently proposed by Rahmani et al. (2008). In ACCIO!, signatures are also extracted within segmented regions or in the neighborhood of salient points.

3.2. Limitations of existing MIL methods for medical applications

One major limitation of ACCIO! and DD-SVM is the assumption that relevant image patches either contain salient points or can be automatically segmented. This is not necessarily true in medical

[2] http://roc.healthcare.uiowa.edu/.

applications, where relevant patterns (e.g. lesions) sometimes have fuzzy edges and their contrast with the background is sometimes low. In the proposed framework, unlike ACCIO! and DD-SVM, relevant patterns can be characterized even if they cannot be segmented. Robustness is increased by removing the segmentation step. Also, because this framework does not rely on any segmentation algorithm, increasing image dimension or changing image modality is trivial. This in an important feature in medical images where acquisition devices are regularly updated. *Diverse Density* and Andrew's SVM framework also have one major limitation: fixed size patches are described, whereas images may contain relevant patterns of varying size. To overcome these limitations, a multi-resolution and segmentation-free approach is adopted in this paper.

3.3. Proposed improvements

In the *Diverse Density* and ACCIO! formulations, learning the target concept(s) involves image feature weighting (Maron and Ratan, 1998; Rahmani et al., 2008). Let a *reference signature* denote a signature extracted from a reference image (see Table 1). In Andrews' DSM formulation, learning the target concept(s) involves training an SVM (Andrews et al., 2003); said differently, it involves assigning a weight to each reference signature, namely the optimal Lagrange multiplier (Schölkopf and Smola, 2002). A unifying approach is presented in this paper: both image features and reference signatures are weighted. We will show that combining image feature weighting and reference signature weighting significantly improves performance. Moreover, signatures are extracted in patches of varying size. Classifiers can be trained independently for each patch size, which defines *single-resolution* relevance scores (*resolution* refers to the distance between two neighboring patches, and is therefore related to patch size). A classifier can also be trained across several resolutions simultaneously: a *multi-resolution* extension is proposed for both image feature weighting and reference signature weighting.

4. Outline of the proposed framework

The proposed framework is summarized in Fig. 1 and the main terms are defined in Table 1.

The first step is to build up a reference dataset. Each image in this dataset must be categorized, as relevant or irrelevant, by clinicians.

The second step is to train the automatic categorization system. Each reference image is divided into possibly-overlapping rectangular image patches (see Section 5). A signature, i.e. a vector of image features, is extracted from each patch. Then, a local relevance score is computed for each patch (see Section 6). The local relevance score of a patch is based on comparisons between the signature of this patch and the signature of all other patches in the reference dataset (the *reference signatures* for short); this is where CBIR comes into play. The local relevance score is expected to be high if the patch includes relevant patterns and low otherwise. Local relevance scores are defined by two sets of weights: (1) weights assigned to image features and (2) weights assigned to reference signatures. These weights are tuned in order to ensure that patches extracted from irrelevant images are assigned a low relevance score (see Section 7). A 2-step process is used to optimize the weights:

1. the image feature weights are optimized while keeping the reference signature weights constant,
2. the reference signature weights are optimized while keeping the image feature weights constant.

The third step is to categorize new images. Each new image is divided into possibly-overlapping rectangular image patches. A signature is extracted from each patch. Then, a local relevance score is computed for each patch using the optimal weights. Finally, a global relevance score is computed for the new image: this global relevance score combines all local relevance scores (see Section 8).

In the following sections, the framework is presented in the case of fixed-size image patches and two-dimensional images. A multiscale extension is presented in Appendix A and a generalization to higher-dimensional images is presented in Appendix B.

5. Image patches

Let I be an image of size $M \times N$. Let $j \in \mathbb{N}_1$ be a scale factor, where $\mathbb{N}_1$ denotes the set of non-zero natural numbers.

The size of a patch in I is given by $\frac{M}{j} \times \frac{N}{j}$. Its top-left corner is at location (x, y):

$$\begin{cases} x = \frac{M}{j}\left(u + \frac{\delta_u}{K}\right), & u = 0 \ldots j-1, \quad \delta_u = 0 \ldots K-1 \\ y = \frac{N}{j}\left(v + \frac{\delta_v}{K}\right), & v = 0 \ldots j-1, \quad \delta_v = 0 \ldots K-1 \end{cases} \tag{1}$$

where $K \in \mathbb{N}_1$ controls the number of patches (see Fig. 2). If $K = 1$, patches at a given scale form a partition of I. If $K > 1$, these patches overlap. Let $I_{j,x,y}$ denote the patch of I at scale j and location (x, y).

The scale factor has to be adapted to the typical size of the patterns indicating the presence of the target pathology(-ies). If the patches are too large, then the relevant patterns they contain will have little impact on the extracted image features. On the contrary, if the patches only contain small portions of the relevant patterns, there may be enough data to reliably compute some features (texture features or color features) but not all features (global shape parameters, for instance). Therefore, the scale factor has to be trained and the optimal scale factor probably depends on the image features that are used. If the presence of the pathology is indicated by several types of patterns, with different typical sizes, then several scale factors should be used simultaneously (see appendix A).

6. A definition for the local relevance score

Let I be an image and let $I_{j,x,y}$ be a patch of I. Let $\boldsymbol{s}(I_{j,x,y}) = \{\boldsymbol{s}_1(I_{j,x,y}), \boldsymbol{s}_2(I_{j,x,y}), \ldots, \boldsymbol{s}_D(I_{j,x,y})\}$, (or $\boldsymbol{s} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_D\}$ for short) be the signature of $I_{j,x,y}$.

In order to define a local relevance score for $I_{j,x,y}$, we search for a neighborhood of signature $\boldsymbol{s}$, noted $\mathcal{N}(\boldsymbol{s})$, among the reference signatures. Precisely, $\mathcal{N}(\boldsymbol{s})$ contains the $k \in \mathbb{N}_1$ nearest neighbors of $\boldsymbol{s}$, with respect to a Minkowski metric $\|.\|_m$ ($m \in \mathbb{N}_1$) and with the restriction that no more than $k_{max} \in \mathbb{N}_1$ neighbors are selected per reference image. Minkowski metrics include the Manhattan distance ($m = 1$), the Euclidean distance ($m = 2$) and the Max distance ($m = \infty$). The three neighborhood parameters (m, k and k_{max}) need to be trained (see Section 9.5).

$\mathcal{N}(\boldsymbol{s})$ is divided into $\mathcal{N}^-(\boldsymbol{s})$ and $\mathcal{N}^+(\boldsymbol{s})$, respectively the set of irrelevant and possibly-relevant signatures in the neighborhood of $\boldsymbol{s}$ (see Table 1).

Let ϕ_m be a similarity metric derived from $\|.\|_m^m$, the Minkowski metric of order m, raised to the power of m:

$$\phi_m(\boldsymbol{s}, \boldsymbol{t}) \triangleq \|\boldsymbol{s} - \boldsymbol{t}^{max}(\boldsymbol{s})\|_m^m - \|\boldsymbol{s} - \boldsymbol{t}\|_m^m \tag{2}$$

where $\boldsymbol{t}$ is a signature and $\boldsymbol{t}^{max}(\boldsymbol{s})$ is the most distant signature in the neighborhood of $\boldsymbol{s}$:

$$\boldsymbol{t}_{max}(\boldsymbol{s}) \triangleq \underset{\boldsymbol{t} \in \mathcal{N}(\boldsymbol{s})}{\operatorname{argmax}} \|\boldsymbol{s} - \boldsymbol{t}\|_m \tag{3}$$

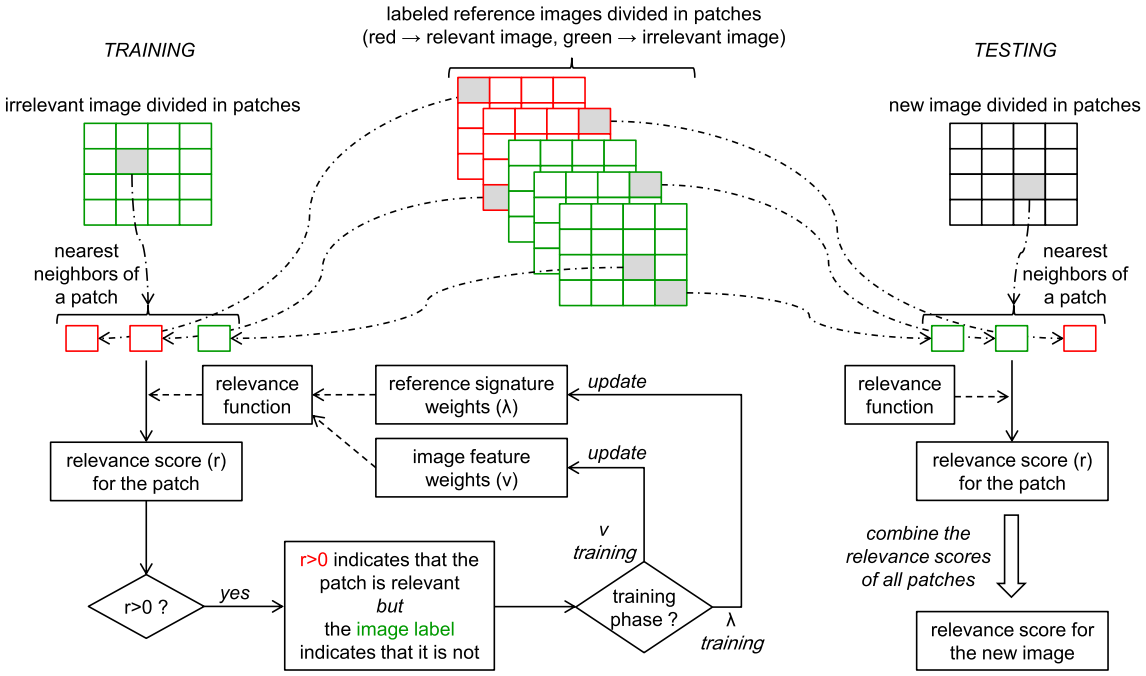


Fig. 1. Outline of the proposed framework.

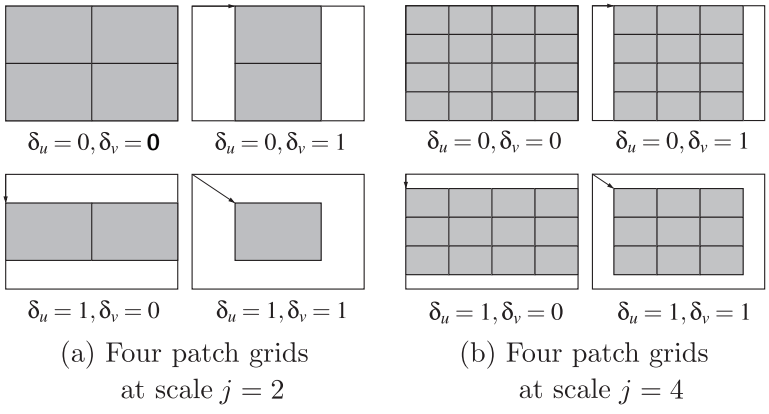


Fig. 2. Image patch geometry. Let I be an image of size $M \times N$. Fig. (a) and (b) show the location of the patches of I at scale $j = 2$ and $j = 4$, respectively, for $K = 2$.

The relevance probability of $\boldsymbol{s}$, noted $p(\boldsymbol{s})$, derives from ϕ_m as follows:

$$\begin{cases} \phi_m^+(\boldsymbol{s}) \triangleq \sum_{\boldsymbol{t} \in \mathcal{N}^+(\boldsymbol{s})} \phi_m(\boldsymbol{s}, \boldsymbol{t}) \\ \phi_m^-(\boldsymbol{s}) \triangleq \sum_{\boldsymbol{t} \in \mathcal{N}^-(\boldsymbol{s})} \phi_m(\boldsymbol{s}, \boldsymbol{t}) \\ \phi_m(\boldsymbol{s}) \triangleq \phi_m^+(\boldsymbol{s}) + \phi_m^-(\boldsymbol{s}) \\ p(\boldsymbol{s}) \triangleq \frac{\phi_m^+(\boldsymbol{s})}{\phi_m(\boldsymbol{s})} \end{cases} \tag{4}$$

By analogy reasoning, $p(\boldsymbol{s})$ increases with the fraction of possibly-relevant signatures in the neighborhood of $\boldsymbol{s}$, and also with the similarity of these signatures to $\boldsymbol{s}$. Finally, the relevance score for $I_{j,x,y}$ is defined as follows:

$$r(I_{j,x,y}) \triangleq p(\boldsymbol{s}) - p_0 \tag{5}$$

where p_0 is the chance relevance probability: p_0 is the threshold on $p(.)$ between relevant and irrelevant signatures. Initially, p_0 is the fraction of relevant images in the reference dataset. An adaptive variation on $r(I_{j,x,y})$ is presented in Section 7 for training purposes. The reasons why we chose this definition are enumerated below. Firstly, we focused on a neighborhood of $\boldsymbol{s}$ (rather than the entire set of reference signatures) for complexity reasons. In particular, fast algorithms for finding the k nearest neighbors of a signature, according to Minkowski metrics, are available, which allows large reference datasets to be used. The Approximate Nearest Neighbors (ANN) library[3] was used in this study. Secondly, the number of neighbors per reference image was limited to $k_{max} \in \mathbb{N}_1$ in order to avoid a selection bias. Suppose some signature $\boldsymbol{t}$ belongs to the neighborhood of $\boldsymbol{s}$. Then, without the above-mentioned limitation, the neighborhood of $\boldsymbol{s}$ would likely contain other signatures from the same image as $\boldsymbol{t}$: signatures extracted from patches in the same region of the eye, depicting the same tissue and photographed with similar lighting conditions. Thirdly, ϕ_m similarity metrics were chosen because they can be decomposed into a sum of feature-specific similarities $\phi_{m,d}$, where d denotes the feature index:

$$\phi_{m,d}(\boldsymbol{s}, \boldsymbol{t}) = (\boldsymbol{s}_d - \boldsymbol{t}_d^{max}(\boldsymbol{s}))^m - (\boldsymbol{s}_d - \boldsymbol{t}_d)^m \tag{6}$$

This property will prove convenient for training purposes (see Section 7.2).

[3] http://www.cs.umd.edu/~mount/ANN/.

7. Localized weight updating

Let $\mathcal{I}^-$ denote the set of all irrelevant images in the reference dataset $\mathcal{I}$. Let $I \in \mathcal{I}^-$ be an irrelevant image. Let $I_{j,x,y}$ be a patch of I and let $\boldsymbol{s}$ be its signature.

Because I is irrelevant, we know that all its patches are irrelevant. In particular, $I_{j,x,y}$ is irrelevant. As a consequence, the relevance score for $I_{j,x,y}$ should be non-positive: $r(I_{j,x,y}) \leqslant 0$ (i.e. $p(\boldsymbol{s}) \leqslant p_0$). Two sets of weights are introduced in the relevance score definition: (1) weights assigned to image features and (2) weights assigned to reference signatures (see Section 4). These weights are tuned to make sure $r(I_{j,x,y}) \leqslant 0$ whenever $I_{j,x,y}$ is known to be irrelevant. Note that relevant images contain a mixture of irrelevant and relevant patches, and we do not know which ones are relevant, so these images cannot be used to supervise weight adaptation.

A weight $\boldsymbol{v}_d \in \mathbb{R}$ is associated with each feature d and a weight $\lambda(\boldsymbol{t}) \in \mathbb{R}$ is associated with each reference signature $\boldsymbol{t}$. Weight initialization is presented in Section 7.1. A weight updating procedure is presented both for image feature weights (see Section 7.2) and and for reference signature weights (see Section 7.3).

To allow weight updating, Eq. (4) becomes:

$$\begin{cases} \phi^+_m(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t}\in\mathcal{N}^+(\boldsymbol{s})} \lambda(\boldsymbol{t}) \sum\limits_{d=1}^{D} \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s},\boldsymbol{t}) \\ \phi^-_m(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t}\in\mathcal{N}^-(\boldsymbol{s})} \lambda(\boldsymbol{t}) \sum\limits_{d=1}^{D} \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s},\boldsymbol{t}) \\ \phi_m(\boldsymbol{s}) \triangleq \phi^+_m(\boldsymbol{s}) + \phi^-_m(\boldsymbol{s}) \\ p(\boldsymbol{s}) \triangleq \frac{\phi^+_m(\boldsymbol{s})}{\phi_m(\boldsymbol{s})} \end{cases} \tag{7}$$

Note that Minkowski metrics can still be used to find the most similar reference signatures, but they must be applied to signatures $\left\{ {}^{\boldsymbol{s}_d}/_{\boldsymbol{v}_d} | d = 1 \dots D \right\}$ and $\left\{ {}^{\boldsymbol{t}_d}/_{\boldsymbol{v}_d} | d = 1 \dots D \right\}$, instead of $\boldsymbol{s}$ and $\boldsymbol{t}$.

7.1. Weight initialization

Let $\boldsymbol{\sigma}_d$ denote the standard deviation of feature d, at scale j, computed over all the patches in all the reference images. Image feature weight $\boldsymbol{v}_d$ is initialized to $1/\boldsymbol{\sigma}_d$. Reference signature weight $\lambda(\boldsymbol{t})$ is initialized to 1.

7.2. Image feature weight updating

In a first scenario, image feature weights $\boldsymbol{v}_d, d = 1 \dots D$, are updated while reference signature weights $\lambda(\boldsymbol{t})$ remain constant. Given that reference signature weights are constant, and therefore do not depend on d, Eq. (7) above can be simplified as follows:

$$\begin{cases} \phi^+_{m,d}(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t}\in\mathcal{N}^+(\boldsymbol{s})} \lambda(\boldsymbol{t}) \phi_{m,d}(\boldsymbol{s},\boldsymbol{t}) \\ \phi_{m,d}(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t}\in\mathcal{N}^+(\boldsymbol{s})\cup\mathcal{N}^-(\boldsymbol{s})} \lambda(\boldsymbol{t}) \phi_{m,d}(\boldsymbol{s},\boldsymbol{t}) \\ p(\boldsymbol{s}) \triangleq \frac{\sum_{d=1}^{D} \boldsymbol{v}_d \phi^+_{m,d}(\boldsymbol{s})}{\sum_{d=1}^{D} \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s})} \end{cases} \tag{8}$$

Let us remind the reader that weight updating is performed on irrelevant images only: $\boldsymbol{s}$ is a signature describing a patch of $I \in \mathcal{I}^-$. If $p(\boldsymbol{s}) > p_0$, then $\boldsymbol{v}_d$ is updated. Note that $\boldsymbol{v}_d$ is global: it is not tuned to the neighborhood of $\boldsymbol{s}$ specifically, otherwise the system would not be able to generalize this update to new images. As a consequence, local optimal solutions (obtained as described in Section 7.2.1) are combined into a global solution as described in Section 7.2.2.

7.2.1. Locally optimal image feature weight updates

If $p(\boldsymbol{s}) > p_0$, we multiply $\boldsymbol{v}_d$ by a real coefficient $\boldsymbol{x}_d \in \mathbb{R}$ in such a way that $p(\boldsymbol{s}) = p_0$ after multiplication. The locally optimal set of $\boldsymbol{x}_d$ coefficients is solution to the following equation (see Eq. (8)):

$$p(\boldsymbol{s}) - p_0 = \frac{\sum_{d=1}^{D} \boldsymbol{v}_d \phi^+_{m,d}(\boldsymbol{s})}{\sum_{d=1}^{D} \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s})} - \frac{\sum_{d=1}^{D} \boldsymbol{x}_d \boldsymbol{v}_d \phi^+_{m,d}(\boldsymbol{s})}{\sum_{d=1}^{D} \boldsymbol{x}_d \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s})} \tag{9}$$

If we define $S(\boldsymbol{s}) \triangleq \sum_{d=1}^{D} \boldsymbol{v}_d \phi_{m,d}(\boldsymbol{s})$ and $S^+(\boldsymbol{s}) \triangleq \sum_{d=1}^{D} \boldsymbol{v}_d \phi^+_{m,d}(\boldsymbol{s})$, then Eq. (9) simplifies to:

$$\sum_{d=1}^{D} \boldsymbol{x}_d \boldsymbol{v}_d \Big[\Big((p(\boldsymbol{s}) - p_0) \phi_{m,d}(\boldsymbol{s}) + \phi^+_{m,d}(\boldsymbol{s}) \Big) S(\boldsymbol{s}) - \phi_{m,d}(\boldsymbol{s}) S^+(\boldsymbol{s}) \Big] = 0 \tag{10}$$

7.2.2. Global updating of image feature weights

This process is replicated for each irrelevant signature $\boldsymbol{s}^{(e)}$ such that $p(\boldsymbol{s}^{(e)}) > p_0, e = 1 \dots E$. As a consequence, a system of equations has to be solved. This system has D unknowns, $\boldsymbol{x}_1 \dots \boldsymbol{x}_D$, and E equations. It can be written as follows:

$$\begin{cases} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ \boldsymbol{A}_{e,d} \triangleq \boldsymbol{v}_d \Big[\Big((p(\boldsymbol{s}^{(e)}) - p_0) \phi_{m,d}(\boldsymbol{s}^{(e)}) + \phi^+_{m,d}(\boldsymbol{s}^{(e)}) \Big) S(\boldsymbol{s}^{(e)}) \\ \qquad - \phi_{m,d}(\boldsymbol{s}^{(e)}) S^+(\boldsymbol{s}^{(e)}) \Big] \\ \boldsymbol{b}_e \triangleq 0 \end{cases} \tag{11}$$

with $e = 1 \dots E$ and $d = 1 \dots D$. Since E is generally larger than D, we propose to solve this system in the least mean squared sense, after weighting each equation e by $(p(\boldsymbol{s}^{(e)}) - p_0)$. For equation weighting, $\boldsymbol{A}$ is replaced by $\tilde{\boldsymbol{A}}$ in system (11):

$$\begin{aligned} \tilde{\boldsymbol{A}}_{e,d} = & (p(\boldsymbol{s}^{(e)}) - p_0) \boldsymbol{v}_d \Big[\Big((p(\boldsymbol{s}^{(e)}) - p_0) \phi_{m,d}(\boldsymbol{s}^{(e)}) + \phi^+_{m,d}(\boldsymbol{s}^{(e)}) \Big) S(\boldsymbol{s}^{(e)}) \\ & - \phi_{m,d}(\boldsymbol{s}^{(e)}) S^+(\boldsymbol{s}^{(e)}) \Big] \end{aligned} \tag{12}$$

Let F be the function to minimize:

$$F(\boldsymbol{x}) \triangleq \Big(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b} \Big)^T \Big(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b} \Big) = \boldsymbol{x}^T \tilde{\boldsymbol{A}}^T \tilde{\boldsymbol{A}} \boldsymbol{x} \tag{13}$$

To ensure that image feature weights are non-negative, $\boldsymbol{x}_d$ must be non-negative, $d = 1 \dots D$. Moreover, for Eq. (9) to be valid, solution $\boldsymbol{x} = \boldsymbol{0}$ must be avoided. So, we enforce the sum of all image feature weights to be positive: without limitations, the sum is set to 1. System (11), together with the above-mentioned constraints, can be rewritten as a quadratic program:

$$\begin{cases} min & \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{l}^T \boldsymbol{x} \\ \boldsymbol{Q} & = 2 \tilde{\boldsymbol{A}}^T \tilde{\boldsymbol{A}} \\ \boldsymbol{l} & = \boldsymbol{0} \\ s.t. & \boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{x}^T \boldsymbol{v} = 1 \end{cases} \tag{14}$$

$\boldsymbol{Q}$ is written as a $\boldsymbol{M}^T\boldsymbol{M}$ product, $\boldsymbol{M} \in \mathcal{M}_{E,D}(\mathbb{R})$. As a consequence, $\boldsymbol{Q}$ is a positive-semidefinite matrix. If $\boldsymbol{Q}$ is also invertible, then system (14) has a unique solution, which can be found in polynomial time with the ellipsoid method.[4]

7.3. Reference signature weight updating

In the previous section, we have seen how to tune image feature weights $\boldsymbol{v}_d$ while keeping reference signature weights $\lambda(\boldsymbol{t})$ constant. The complementary scenario is presented in this second scenario. Let $I \in \mathcal{I}^-$ be an irrelevant image. Let $I_{j,x,y}$ be a patch of I and let $\boldsymbol{s}$ be its signature.

If $p(\boldsymbol{s}) > p_0$, then for each possibly-relevant signature $\boldsymbol{t}$ in the neighborhood of $\boldsymbol{s}$, $\lambda(\boldsymbol{t})$ is updated. Like $\boldsymbol{v}_d, \lambda(\boldsymbol{t})$ is global: local opti-

[4] CGAL/QP_solver—http://www.cgal.org.

mal solutions (obtained as described in Section 7.3.1) are combined into a global solution as described in Section 7.3.2.

7.3.1. Locally optimal reference signature weight updates

If $p(\boldsymbol{s}) > p_0$, then we multiply the weight of each possibly-relevant signature $\boldsymbol{t}$ in the neighborhood of $\boldsymbol{s}$ by a coefficient $x(\boldsymbol{s}) \in \mathbb{R}$ in such a way that $p(\boldsymbol{s}) = p_0$ after multiplication. Because image feature weights are constant, this is equivalent to multiplying $\phi_m^+(\boldsymbol{s})$ by $x(\boldsymbol{s})$, where $\phi_m^+(\boldsymbol{s})$ is defined in Eq. (7). The locally optimal $x(\boldsymbol{s})$ coefficient is solution to the following equation (see Eq. (7)):

$$p(\boldsymbol{s}) - p_0 = \frac{\phi_m^+(\boldsymbol{s})}{\phi_m(\boldsymbol{s})} - \frac{\phi_m^+(\boldsymbol{s})x(\boldsymbol{s})}{\phi_m^+(\boldsymbol{s})x(\boldsymbol{s}) + \phi_m^-(\boldsymbol{s})} \tag{15}$$

or, equivalently, to the following equation:

$$x(\boldsymbol{s}) = \frac{\phi_m^+(\boldsymbol{s})\phi_m^-(\boldsymbol{s}) - (p(\boldsymbol{s}) - p_0)\phi_m^-(\boldsymbol{s})\phi_m(\boldsymbol{s})}{\phi_m^+(\boldsymbol{s})\phi_m^-(\boldsymbol{s}) + (p(\boldsymbol{s}) - p_0)\phi_m^+(\boldsymbol{s})\phi_m(\boldsymbol{s})} \tag{16}$$

7.3.2. Global updating of reference signature weights

A possibly-relevant signature $\boldsymbol{t}$ likely appears in the neighborhood of several reference signatures (see Section 6). Let $\boldsymbol{s}^{(e)}$ denote the irrelevant signatures such that $p(\boldsymbol{s}^{(e)}) > p_0$ and such that $\boldsymbol{t}$ appears in the neighborhood of $\boldsymbol{s}^{(e)}, e = 1 \dots E$. $\lambda(\boldsymbol{t})$, the weight of signature $\boldsymbol{t}$, is multiplied by a coefficient $y(\boldsymbol{t}) \in \mathbb{R}$ chosen in order to try and satisfy Eq. (16) for each signature $\boldsymbol{s}^{(e)}$. The following system should therefore be solved:

$$\begin{cases} \boldsymbol{a}y(\boldsymbol{t}) = \boldsymbol{b} \\ \boldsymbol{a}_e \triangleq 1 \\ \boldsymbol{b}_e \triangleq x(\boldsymbol{s}^{(e)}) \end{cases} \tag{17}$$

where $e = 1 \dots E$. This system has one unknown, $y(\boldsymbol{t})$, and E equations. Like in image feature weighting, system (17) is solved in the least mean squared sense, subject to $y(\boldsymbol{t}) \geqslant 0$, after weighting each equation e by $(p(\boldsymbol{s}^{(e)}) - p_0)$. For equation weighting, $\boldsymbol{a}$ and $\boldsymbol{b}$ are replaced by $\tilde{\boldsymbol{a}}$ and $\tilde{\boldsymbol{b}}$, respectively:

$$\begin{cases} \tilde{\boldsymbol{a}}_e \triangleq (p(\boldsymbol{s}^{(e)}) - p_0) \\ \tilde{\boldsymbol{b}}_e \triangleq (p(\boldsymbol{s}^{(e)}) - p_0)x(\boldsymbol{s}^{(e)}) \end{cases} \tag{18}$$

Let G be the function to minimize:

$$\begin{aligned} G(y(\boldsymbol{t})) &\triangleq \left(\tilde{\boldsymbol{a}}y(\boldsymbol{t}) - \tilde{\boldsymbol{b}}\right)^T \left(\tilde{\boldsymbol{a}}y(\boldsymbol{t}) - \tilde{\boldsymbol{b}}\right) \\ &= \tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{a}}y(\boldsymbol{t})^2 - 2\tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{b}}y(\boldsymbol{t}) + \tilde{\boldsymbol{b}}^T\tilde{\boldsymbol{b}} \end{aligned} \tag{19}$$

If constraint $y(\boldsymbol{t}) \geqslant 0$ is relaxed, then $G(y(\boldsymbol{t}))$ is minimized when:

$$G'(y(\boldsymbol{t})) = 2\tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{a}}y(\boldsymbol{t}) - 2\tilde{\boldsymbol{a}}^T\boldsymbol{b} = 0 \tag{20}$$

or, equivalently, when:

$$y(\boldsymbol{t}) = \frac{\tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{b}}}{\tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{a}}} = \frac{\sum_{e=1}^{E}\left(p(\boldsymbol{s}^{(e)}) - p_0\right)^2 x(\boldsymbol{s}^{(e)})}{\sum_{e=1}^{E}\left(p(\boldsymbol{s}^{(e)}) - p_0\right)^2} \triangleq \widehat{y(\boldsymbol{t})} \tag{21}$$

If $\widehat{y(\boldsymbol{t})} < 0$, then constraint $y(\boldsymbol{t}) \geqslant 0$ is infringed when $y(\boldsymbol{t}) = \widehat{y(\boldsymbol{t})}$. In that case, it can be easily checked that $G_{|y(\boldsymbol{t}) \geqslant 0}$ is strictly increasing ($G'(y(\boldsymbol{t})) > 0, \forall y(\boldsymbol{t}) \geqslant 0$). Therefore, if $\widehat{y(\boldsymbol{t})} < 0$, the optimal solution is $y(\boldsymbol{t}) = 0$. If $\widehat{y(\boldsymbol{t})} \geqslant 0$, then the optimal solution is $y(\boldsymbol{t}) = \widehat{y(\boldsymbol{t})}$. Unlike image feature weighting, no quadratic programming solver is required for reference signature weighting.

At the end of this procedure, the weight of several possibly-relevant signatures has changed. As a consequence, p_0, the chance relevance probability, should be updated:

$$p_0 = \frac{\sum_{I \in \mathcal{I}^+, x, y} \lambda(\boldsymbol{s}(I_{j,x,y}))}{\sum_{I \in \mathcal{I}, x, y} \lambda(\boldsymbol{s}(I_{j,x,y}))} \tag{22}$$

where $\mathcal{I}^+$ denotes the set of all relevant images in $\mathcal{I}$.

8. A definition for the global relevance score

Let I be an image. The global relevance score for I, noted $r(I)$, is defined as follows ($m \in \mathbb{N}$):

$$r(I) = \sqrt[m]{\sum_{x,y} \left|r(I_{j,x,y})\right|^m} \tag{23}$$

$r(I)$ is the m-norm of the vector consisting of the local relevance scores computed for patches in I. $r(I)$ is expected to be large if I is relevant and small otherwise.

9. Application to diabetic retinopathy screening

The proposed framework was applied to Diabetic Retinopathy (DR) screening. 25,702 DR screening records were used to test the performance of *global* relevance assessment (see Section 9.1). 273 manually segmented images (see Section 9.2) were used to test the performance of *local* relevance assessment. The system was trained in an independent dataset of 1200 DR screening images (Messidor—see Section 9.3).

9.1. e-ophtha—examination record dataset

e-ophtha contains all examination records collected in the Ophdiat screening network[5] during two consecutive years (2008 and 2009). 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 ophthalmologist, out of 11 participating ophthalmologists, in Lariboisière Hospital (Paris, France). Each record contains four eye fundus photographs on average (two per eye) as well as demographic and biological data. 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. For the purpose of this study, images were automatically resized and cropped, by bilinear interpolation, to a definition of 780×780 pixels (see Fig. 3a). Resizing images may affect classification performance. However, it has one major advantage: a single classifier needs to be trained, as opposed to one classifier per image size, which means that a larger training set can be used and that the same classifier can be used in several datasets. Overall, 107,799 images were collected.

Each examination record was marked (by one ophthalmologist) as relevant or irrelevant. Examination records were regarded as relevant if the patient should be referred to an ophthalmologist for further examinations, treatment, etc. In Ophdiat, patients are referred to an ophthalmologist because signs of DR (or other eye pathologies) have been detected or because images are ungradable. 6,391 records were marked as relevant and 19,311 were marked as irrelevant (prevalence: 25%). Over the 2005–2006 period, agreement between graders ranged from 92% to 99% in the Ophdiat network (Erginay et al., 2008).

9.2. e-ophtha—subset of manually segmented images

273 images were randomly selected from e-ophtha. One participating ophthalmologist annotated eight types of lesions in these

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

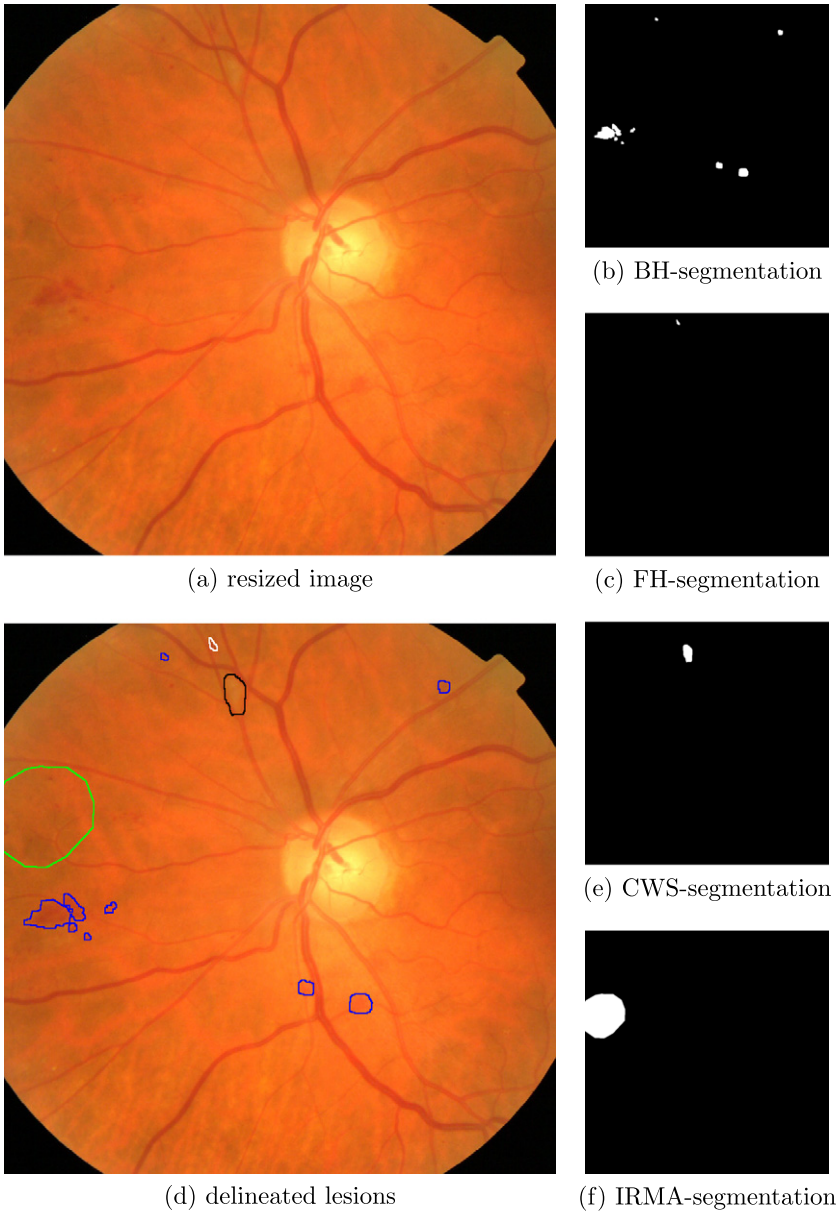


Fig. 3. Manually segmented image. (a) Represents the resized image. (d) Displays the manually segmented lesions: blot hemorrhages (in blue), flame hemorrhages (in white), cotton wool spots (in black) and intra-retinal microvascular abnormalities (in green). (b), (c) and (e) Represent the segmentation maps obtained for blot hemorrhages (BH), flame hemorrhages (FH), cotton wool spots (CWS) and intra-retinal microvascular abnormalities (IRMA). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

images (see Table 2). Note that she did not systematically segment all types of lesions present in each image. However, when she started segmenting one type of lesions in an image, she segmented all lesions of that type in this image. An example of manually segmented image is given in Fig. 3.

9.3. Messidor dataset

The Messidor dataset[6] contains 1200 eye fundus photographs collected in three ophthalmology departments, in France, between 2005 and 2006. Images were obtained with TRC-NW6 non-mydriatic retinographs (Topcon, Tokyo). 800 images were acquired with pupil dilation (one drop of 10% Tropicamide); 400 were acquired without dilation. Image sizes ranged from 1440 × 960 to 2304 × 1536 pixels. For the purpose of this study, images were automatically resized and cropped to a definition of 780 × 780 pixels. 654 images were marked as relevant and 546 were marked as irrelevant.

[6] http://messidor.crihan.fr/index-en.php.

9.4. Image features

Many image features have been proposed in the CBIR literature, the most common being color features, texture features, shape features and salient point descriptors (Datta et al., 2008). These features should be chosen carefully with respect to the problem under study. For diabetic retinopathy screening in eye fundus photographs, many patterns of interest have fuzzy edges, so salient point descriptors likely are not relevant. On the other hand, the relevance of wavelet-based texture features for diabetic retinopathy assessment has been shown in previous studies (Quellec et al., 2010a,). So wavelet-based features were used for texture characterization (Quellec et al., 2010a). For completeness, popular shape features, Zernike moments, were also used. Eighteen wavelet-

Table 2
Segmented lesions.

Lesion type	Number of annotated images	Number of annotated lesions
Microaneurisms (MA)	106	663
Dot hemorrhages (DH)	63	246
Blot hemorrhages (BH)	98	383
Flame hemorrhages (FH)	46	66
Exudates (EX)	47	2648
Cotton wool spots (CWS)	80	282
Intra-retinal microvascular abnormalities (IRMA)	25	57
Neovascularizations (NV)	25	50

based features and nine Zernike moments were extracted in each color channel, respectively. Each experiment was replicated in four different color spaces: RGB, Lab (Hunter and December, 1958), I1I2I3 (Ohta et al., 1980), HSV (Joblove and Greenberg, 1978). This selection of image features is of course arbitrary and a different selection would likely lead to very different results.

9.5. Training in one half of the Messidor dataset

The Messidor dataset was randomly divided into a training subset and a test subset; each subset contains one half of the Messidor dataset.

The following parameters were trained by 2-fold cross-validation on the training subset: the color space, the sequence of weight updating iterations, the scale factor j, the spatial-distribution parameter K, the neighborhood parameters $\{m, k, k_{max}\}$ (see Sections 4 and 6). In each fold, one half of the training subset was used as reference dataset ($\mathcal{I}$) and the classification performance was assessed, in terms of area under the ROC curve (A_z), on the other half.

A naive optimization process was used to train parameters $\{K, m, k, k_{max}\}$: each parameter was trained separately while keeping the other three constant (default constant values: $K = 1$, $m = 2, k = 5, k_{max} = 1$). For each $\{K, m, k, k_{max}\}$ combination, the system was trained using 32 different {color space, sequence of weight updating iterations} pairs (see Table 3): two A_z values were obtained for each pair (one per fold). The score of each $\{K, m, k, k_{max}\}$ combination was defined as the average of these 64 A_z values. The multiscale extension was used during the training process. Overall, the training process lasted approximately 15 h using an Intel Xeon E5520 processor running at 2.27 GHz.

The optimal value for $\{K, m, k, k_{max}\}$ was $\{2, 2, 10, 1\}$. These optimal parameters were used in the following experiments. Cross-validation results, obtained on the training subset with these parameters, are reported in Table 3, in terms of average A_z. In this table, each row is associated with one sequence of image feature (IF) weight updating and reference signature (RS) weight updating iterations.

Table 3
Cross-validation results. The bold value indicates the highest score.

Update sequence	RGB	Lab	I1I2I3	HSV
$\emptyset$	0.810	0.845	0.824	0.846
IF	0.867	0.896	0.865	0.875
RS	0.856	0.893	0.862	0.877
IF–RS	0.900	**0.902**	0.892	0.880
RS–RS	0.860	0.857	0.811	0.840
IF–IF	0.892	0.886	0.878	0.867
IF–IF–RS	0.885	0.896	0.892	0.879
IF–RS–RS	0.859	0.885	0.863	0.866
IF–RS–IF	0.900	0.900	0.894	0.881

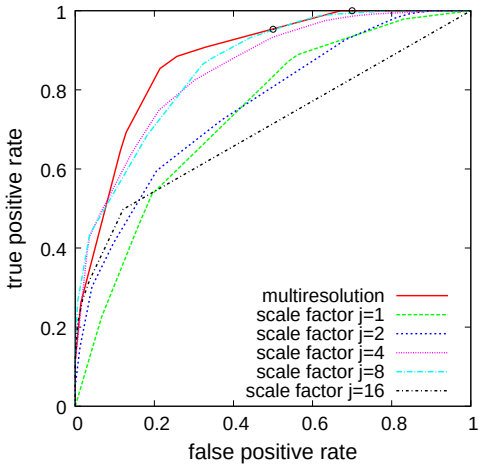


Fig. 4. Classification performance in the Messidor test subset. Black circles indicate interesting settings in a screening context.

The best performance was obtained using the Lab color space, with one iteration of image feature weight updating followed by one iteration of reference signature weight updating (IF–RS). At the end of the training process, the image feature weights and the reference signature weights were re-trained using the entire training subset as reference dataset ($\mathcal{I}$). Unless mentioned otherwise, the same reference dataset and the same weights were used in the following experiments.

9.6. Testing global relevance assessment in the other half of the Messidor dataset

The ROC curves obtained on the Messidor test subset are reported in Fig. 4. The corresponding areas under the ROC curve (A_z) are reported in Table 4 together with the total processing time per image (T), using one core of an Intel Xeon E5520 processor.

Up to scale factor $j = 8$, performance increases with the scale factor (see Fig. 4). However, performance drops at $j = 16$. This is likely due to the small number of pixels per patch (48×48), which may lead to unreliable texture feature values. When the false positive rate is set to 50% (respectively 70%), meaning that 50% (respectively 30%) of the healthy patient records do not need to be seen by an ophthalmologist, the false negative rate is 4.7% (respectively 0%).

A comparison with *Diverse Density* (Maron and Lozano-Pérez, 1998) and *Random Forests*[7] (Breiman, 2001), using the same image features and the same patch geometry, is also reported. We remind the reader that *Diverse Density* is a multiple-instance learning framework. *Random Forests* are not: each image must be characterized by a single feature vector. This feature vector was obtained by concatenating the signatures of all patches in alphabetical order of their location (x, y). The parameters of the *Random Forests* were trained by 2-fold cross-validation in the Messidor training subset: in particular, 50 trees were generated per forest. For scale factor $j = 1$, i.e. when there is only one patch per image, *Random Forests* outperform all other frameworks. However, the performance of *Random Forests* decreases as j increases. This is simply because lesions are not always at the same location in two semantically similar images.

In the Messidor dataset, the proposed framework outperformed a recently published DR detection algorithm by Agurto et al. (2010): an area under the ROC curve of $A_z = 0.84$ was achieved by Agurto's algorithm on a selection of 400 images from the Messidor dataset

[7] http://www.alglib.net/—dforest subpackage.

Table 4
Classification performance in the Messidor test subset. The bold value indicates the highest score for each method.

Method	Figure of merit	Scale factor (j)					
	0	1	2	4	8	16	Multiscale
Proposed method	A_z	0.738	0.760	0.852	**0.860**	0.702	*0.881*
	T (s)	0.551	0.652	1.04	3.35	12.8	13.2
Diverse Density	A_z	0.663	0.687	**0.708**	0.699	0.616	
	T (s)	0.550	0.648	0.992	2.84	10.4	
Random Forests	A_z	**0.757**	0.750	0.743	0.709	0.647	
	T (s)	0.550	0.646	0.982	2.77	9.91	

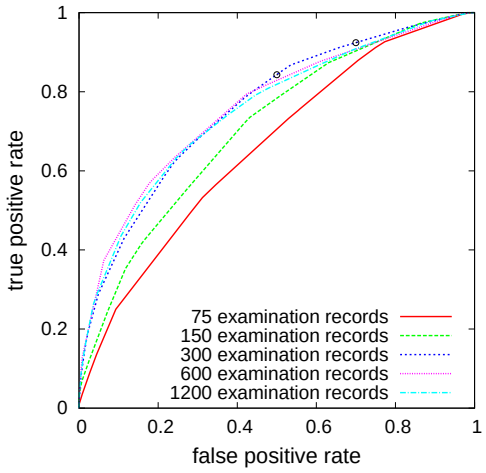


Fig. 5. Classification performance in e-ophtha with respect to the size of the reference dataset. Black circles indicate interesting settings in a screening context.

(computation times are not available). It also slightly outperformed Antal et al.'s (2011) framework, which relies on targeted lesion detectors ($A_z = 0.87$). Note that Giancardo et al. (2012) reported similarly high A_z scores in the Messidor dataset, but regarding the detection of a different pathology: diabetic macular edema.

9.7. Testing global relevance assessment in e-ophtha

In this experiment, the system was trained using images from Messidor and tested in e-ophtha, an independent dataset.

Let us remind the reader that, in e-ophtha (see Section 9.1), relevance labels are not assigned to images alone, but to examination records as a whole: each examination $\mathcal{E}$ record usually contains four images (see Section 9.1), but this number can vary: it ranges from 1 to 19. To measure the performance of global relevance assessment in this dataset, the global relevance score for $\mathcal{E}$ was defined as follows ($m \in \mathbb{N}_1$):

$$r(\mathcal{E}) = \sqrt[m]{\frac{1}{\text{number of images in}\mathcal{E}} \sum_{I \in \mathcal{E}, x, y} \left| r(I_{j,x,y}) \right|^m} \tag{24}$$

We evaluated how the size $|\mathcal{I}|$ of the reference dataset impacts the performance of global relevance assessment. Five values for $|\mathcal{I}|$ were tested: 75, 150, 300, 600 and 1200. For $|\mathcal{I}| = 1200, \mathcal{I}$ was the entire Messidor dataset. For $|\mathcal{I}| = 600, \mathcal{I}$ was the Messidor training subset (see Section 9.5). For the remaining values of $|\mathcal{I}|, \mathcal{I}$ was selected at random in the Messidor dataset. The weights were retrained for each value of $|\mathcal{I}|$.

The ROC curves obtained on e-ophtha, at scale $j = 8$ (see Table 4), are reported in Fig. 5. The corresponding areas under the ROC curve (A_z) are reported in Table 5 together with the total processing time per image (T). It can be seen that, for $|\mathcal{I}| \geqslant 300$, the classification performance is little dependent on the size of the reference dataset.

Table 5
Classification performance in e-ophtha. The bold value indicates the highest score.

Figure of merit	Size of the reference dataset $\lvert\mathcal{I}\rvert$				
	75	150	300	600	1200
A_z	0.658	0.708	0.759	**0.761**	0.754
T (s)	3.07	3.16	3.25	3.34	3.44

Then, we evaluated how the proportion of irrelevant and relevant images in the reference dataset, noted p_0, impacts the performance of global relevance assessment. Three reference datasets of 300 images were selected at random in the Messidor dataset: one with $p_0 = 0.25$, one with $p_0 = 0.55$ (the prevalence in Messidor) and one with $p_0 = 0.75$. The performance on e-ophtha was respectively $A_z = 0.719, A_z = 0.759$ and $A_z = 0.702$.

When the false positive rate is set to 50% (respectively 70%), the false negative rate is 15.7% (respectively 7.6%) in the best scenario ($|\mathcal{I}| = 300, p_0 = 0.55$, see Fig. 5). Ideally, the false negative rate should be zero. Among these false negatives, 6 patients (respectively 1 patient) have proliferative DR; in e-ophtha, 123 patients have proliferative DR. Furthermore, among these false negatives, images were considered ungradable by ophthalmologists in 172 records (respectively 69 records); in e-ophtha, images were considered ungradable in 2,167 records.

9.8. Testing local relevance assessment in manually segmented images

In this experiment, a local relevance score was computed for each patch of the 273 manually segmented images (see Section 9.2) and and each patch of 200 randomly selected irrelevant images from e-ophtha. The Messidor training subset was used as reference dataset. To evaluate these local relevance scores, lesion-specific *local relevance labels* were assigned to each patch, as explained hereafter. Suppose the ophthalmologist segmented all lesions of type l in image $I, l \in \{MA, DH, BH, FH, EX, CWS, IRMA, NV\}$ (see Table 2). Then, an l-label map was built for I. Each patch of I having a non-zero intersection with the l-segmentation map (see Fig. 3) was labeled as l-relevant; the others were labeled as l-irrelevant. This process is illustrated in Fig. 6.

Now that local relevance labels are assigned to each patch, local relevance scores can be evaluated through a ROC analysis. A ROC curve was built for each lesion type l. To build this ROC curve, we used (1) the relevance score of each (l-irrelevant) patch in the set of 200 randomly selected irrelevant images and (2) the relevance score of each l-relevant patch in manually segmented images. For instance, 106 manually segmented images were used to evaluate *MA* detection. These ROC curves are reported in Fig. 7.

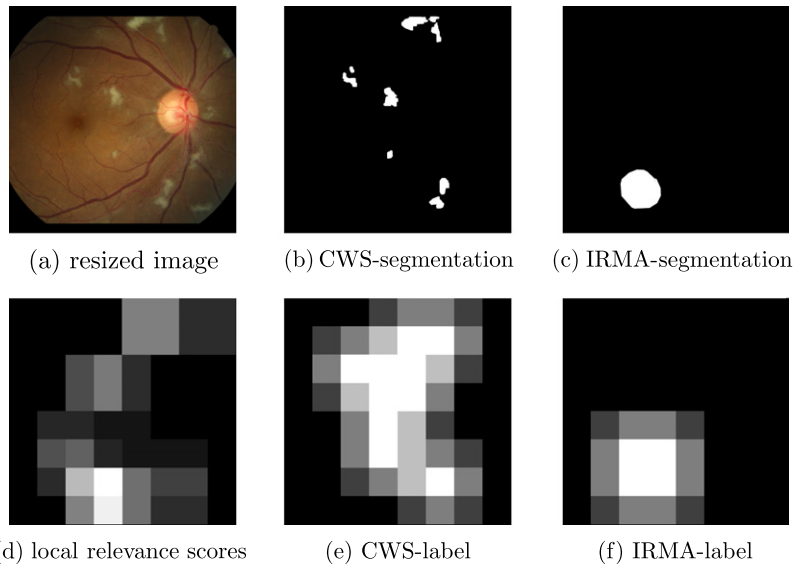


Fig. 6. Evaluating local relevance scores. The CWS- and IRMA-segmentation maps in images (b) and (c) are converted into CWS- and IRMA-label maps in images (e) and (f). Local relevance scores, computed by the proposed framework, are displayed in image (d). To obtain the CWS-label map, a binary label was assigned to each image patch. A patch was assigned a 'CWS-relevant' label if and only if it had a non-zero intersection with the CWS-segmentation map (image (b)). Note that, although labels are binary, the visual representation of the CWS-label map (image (e)) is not binary. The reason is that patches overlap (see Fig. 2b). In image (e), the intensity of a pixel is proportional to the number of patches that were assigned a 'CWS-relevant' label among all patches intersecting the pixel. As a result, the CWS-label map is a dilated version of the CWS-segmentation map. A similar process was used to build image (d) and (f).

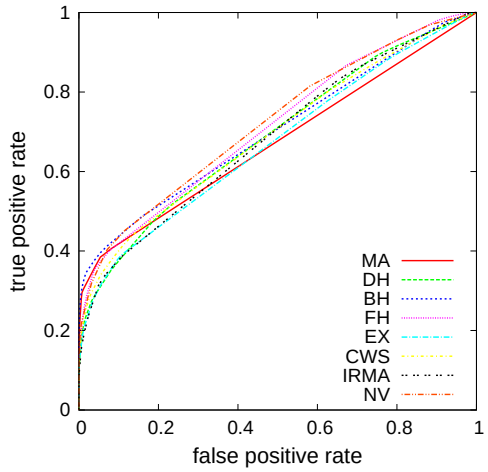


Fig. 7. Local relevance assessment. The area under the ROC curves range from $A_z = 0.671$ (EX – exudates) to $A_z = 0.721$ (NV – neovascularizations).

10. Discussion and conclusions

A Multiple-instance learning framework for finding relevant patterns in images was presented in this paper. It only needs a reference image dataset, in which images have been annotated by clinicians as *relevant* or *irrelevant*, for supervision. Given a set of image features, the importance of each image feature, and of each reference signature itself, is weighted by a novel weight updating procedure. Both a single- and a multi-resolution implementation were presented.

The proposed framework was successfully applied to diabetic retinopathy screening in two eye fundus image datasets: Messidor (1200 images) and e-ophtha (107,799 images). In the Messidor dataset, the proposed framework achieved an area under the ROC curve of $A_z = 0.881$. It compared favorably with *Random Forests*, a popular single-instance learner, using the same image features ($A_z = 0.757$). This comparison shows the interest of multiple-instance learning in this application. Using the multi-resolution approach, the proposed framework outperformed previously published results (Agurto et al., 2010; Antal et al., 2011). Of course, classification performance and computation times are a trade-off: with the multi-resolution approach, the average processing time was 13.2s per image. Much faster results were obtained with the single-resolution approach: computation times drop to 1.04 s using a scale factor $j = 4$. With this setup, a performance of $A_z = 0.852$ ($A_z > 0.84$) was obtained. The proposed framework was also applied to e-ophtha. This dataset is more challenging in that relevance labels are not assigned to individual images but to examination records as a whole: each record contains four images on average. Besides, about 10% of examination records were regarded by experts as ungradable due to poor image quality. In that respect, a performance of $A_z = 0.761$ is satisfactory. We observed, in a subset of 273 manually segmented images, that all eight types of diabetic retinopathy lesions are locally detected with an area under the ROC curve ranging from $A_z = 0.671$ (exudates) to $A_z = 0.721$ (neovascularizations). It should be noted that the system also detects mild textural abnormalities that were not manually detected and were therefore erroneously counted as false alarms. Among other factors, these textural abnormalities can be due to cataract or retinal epithelial pigment variations. Besides, even if the detections are not perfect locally, the integration of all local detections over the entire examination record proved efficient.

The proposed framework has several advantages.

1. Previous DR screening systems target individual lesions. However, what characterizes the appearance of a lesion or an image as abnormal, is sometimes a complex set of interconnected elements, at different scales. By taking this factor into account, the proposed framework has the ability to push classification performance further. In the proposed framework, unlike other MIL frameworks, relevant image patterns are identified by the image features themselves, and not necessarily by salient

points. This property is particularly well suited to medical images, in which relevant patterns may only be visible through subtle textural or color changes (that would be missed by salient point detectors).
2. Previous systems rely on manual lesion segmentations for supervision. However, clinicians often do not work in such a manner: they do not necessarily count lesions. Moreover, they may not be able to explicitly indicate which image patterns incite them to classify an image as relevant or irrelevant, though they know at a global level that the image is relevant. This framework is more in accordance with clinicians' reasoning.
3. The proposed framework is trained and validated using existing clinical data and records as they are. Clinicians are not asked additional work: the refer/do not refer decision is the standard one taken in DR screening. So, the proposed approach is well suited to clinician's practice.
4. This framework could be applied to a variety of medical applications: it is useful whenever obtaining accurate ground truth in large datasets is difficult, a recurrent problem in medical applications.

The framework also has limitations.

1. Compared to previous systems, many parameters need to be trained. As a consequence, the training procedure is complex and takes a lot of time.
2. Performance likely depends heavily on which image features were used to characterize image patches. Therefore, it is assumed that users know which features may be relevant in their application.
3. The system relies on reference datasets in which several clinicians labeled non-overlapping sets of images independently. Agreement between clinicians can be as low as 92%. These disagreements were not taken into account in the system.

These limitations will be addressed in future works. First, in order to reduce training times, we will study which parameters are application-independent and which parameters need to be re-trained. Second, a procedure to generate an optimal set of independent features from data will be proposed. So far, feature generation requires manual segmentations (Quellec et al., 2011b): a multiple-instance learning extension will be proposed. Third, the proposed approach will be extended to relevance regression. This extension will have at least two implications: (1) it will allow clinicians to indicate a relevance degree and (2) it will allow multiple-clinician labeling: + (resp. −) will be replaced by the percentage of clinicians that labeled each image as relevant (resp. irrelevant). Finally, regarding DR screening, we have started including demographic and biological data into the system, as well as other dedicated image processing algorithms, to push performance further.

In summary, a novel multiple-instance learner for automatic medical image classification was presented and its relevance was shown in large retinal image datasets.

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). M.D. Abràmoff was supported by the National Institutes of Health (EY017066) and the US Department of Veterans Affairs. The Messidor dataset was kindly provided by the Messidor program partners (see http://messidor.crihan.fr).

Appendix A. Multiscale extension

In the basic method above, a single scale factor is used to define patches. In the multiscale extension, scale factors are in $\{1,2,4,8,\dots,J\}$, $J \in \mathbb{N}_1$. A multiscale local relevance score is defined in Section A.1. This multiscale relevance score is associated with a multiscale weight updating strategy (see Section A.2). The global multiscale relevance score is defined in Section A.3.

A.1. Multiscale local relevance score

Let I be an image. Let $I_{J,x,y}$ be a patch of I, at the highest scale J, and let $\boldsymbol{s}$ be its signature.

For the multiscale relevance score definition, we define set $\mathcal{S}_j(\boldsymbol{s})$ and function ω. Set $\mathcal{S}_j(\boldsymbol{s})$ contains the signature of all patches of I, at scale j, that overlap with $I_{J,x,y}$ (see Fig. 8). $\omega(\boldsymbol{s},\boldsymbol{u}) \in [0;1]$ denotes the fraction of a patch described by $\boldsymbol{u} \in \mathcal{S}_j(\boldsymbol{s})$ that overlaps with $I_{J,x,y}$. The multiscale local relevance score for $I_{J,x,y}$ combines one-scale relevance probabilities computed for signatures $\boldsymbol{u} \in \mathcal{S}_j(\boldsymbol{s}), j \in \{1,2,4,8,\dots,J\}$, with respect to $\omega(\boldsymbol{s},\boldsymbol{u})$.

Image feature weights are now scale-specific, i.e. $\boldsymbol{v}_d$ becomes $\boldsymbol{v}_{d,j}$. $\phi_m^+(\boldsymbol{s})$, originally defined in Eq. (7), now becomes:

$$\phi_m^+(\boldsymbol{s}) \triangleq \sum_{j\in\{1,2,4,8,\dots,J\}} \sum_{\boldsymbol{u}\in\mathcal{S}_j(\boldsymbol{s})} \omega(\boldsymbol{s},\boldsymbol{u}) \times \left\{ \sum_{\boldsymbol{t}\in\mathcal{N}^+(\boldsymbol{u})} \lambda(\boldsymbol{t}) \sum_{d=1}^{D} \boldsymbol{v}_{d,j}\phi_{m,d}(\boldsymbol{u},\boldsymbol{t}) \right\} \quad \text{(A.1)}$$

$\phi_m^-(\boldsymbol{s})$ is modified similarly. That being said, the relevance score for $I_{J,x,y}$ remains unchanged: it is given by Eq. (5). Weight updating, on the other hand, was slightly modified as explained below.

A.2. Multiscale weight updating

Let $\boldsymbol{s}^{(e)}, e = 1\dots E$, be the irrelevant signatures describing a patch at scale J such that $p(\boldsymbol{s}^{(e)}) > p_0$.

A.2.1. Image feature weight updating

For image feature weight updating, we would like to multiply the weight $\boldsymbol{v}_{d,j}$ of each feature d, at each scale j, by a coefficient $\boldsymbol{x}_{d,j} \in \mathbb{R}$, such that $p(\boldsymbol{s}^{(e)}) = p_0$ after multiplication.

In the multiscale framework, Eq. (10) becomes:

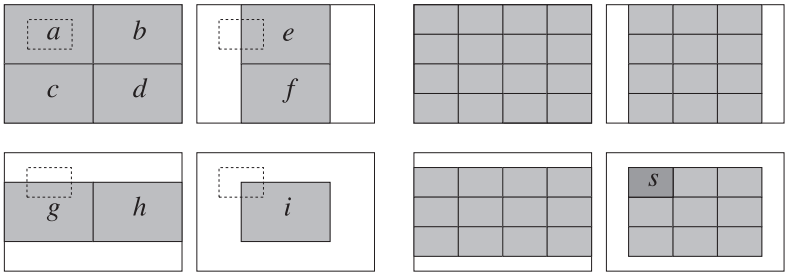
$$0 = \sum_{d=1}^{D} \boldsymbol{x}_{d,j}\boldsymbol{v}_{d,j} \sum_{j\in\{1,2,4,8,\dots,J\}} \sum_{\boldsymbol{u}\in\mathcal{S}_j(\boldsymbol{s}^{(e)})} \omega(\boldsymbol{s}^{(e)},\boldsymbol{u}) \times \left\{ \Big(\big(p(\boldsymbol{s}^{(e)}) - p_0\big)\phi_{m,d}(\boldsymbol{u}) + \phi_{m,d}^+(\boldsymbol{u}) \Big) S(\boldsymbol{u}) - \phi_{m,d}(\boldsymbol{u}) S^+(\boldsymbol{u}) \right\} \quad \text{(A.2)}$$

The system to solve is similar to (11) (and therefore (14)), except that there are now $|\{1,2,4,8,\dots,J\}|D$ unknowns ($\boldsymbol{x}_{d,j}, d = 1\dots D, j \in \{1,2,4,8,\dots,J\}$) instead of D.

A.2.2. Reference signature weight updating

For reference signature weight updating, we would like to multiply the weight $\lambda(\boldsymbol{t})$ of each possibly-relevant signature $\boldsymbol{t}$ in the neighborhood of $\boldsymbol{s}^{(e)}$ by a coefficient $x(\boldsymbol{s}^{(e)}) \in \mathbb{R}$, such that $p(\boldsymbol{s}^{(e)}) = p_0$ after multiplication. In the multiscale framework, $\boldsymbol{t}$ does not necessarily belongs to $\mathcal{N}^+(\boldsymbol{s}^{(e)})$, but rather to $\mathcal{N}^+(\boldsymbol{u}), \boldsymbol{u} \in \mathcal{S}_j(\boldsymbol{s}^{(e)})$ (see Eq. (A.1)).

Eq. (16), that provides the best $x(\boldsymbol{s}^{(e)})$ coefficient, still holds in the multiscale framework. The only thing that changes is the way $\lambda(\boldsymbol{t})$ is globally updated, using all $x(\boldsymbol{s}^{(e)})$ local estimates: each equation in system (17) is now weighted by $\omega(\boldsymbol{s},\boldsymbol{u})\big(p(\boldsymbol{s}^{(e)}) - p_0\big)$, not simply $p(\boldsymbol{s}^{(e)}) - p_0$.



(a) Four patch grids at scale $j = 2$

(b) Four patch grids at the highest scale ($j = J = 4$)

Fig. 8. Multiscale relevance score geometry. Let I be an image and $\boldsymbol{s}$ be a signature describing a patch of I at scale $J = 4$ (see (b)). As seen on (a), $\mathcal{S}_2(\boldsymbol{s}) = \{\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{g}, \boldsymbol{i}\}$, $\omega(\boldsymbol{s}, \boldsymbol{a}) = \frac{1}{4}$, $\omega(\boldsymbol{s}, \boldsymbol{e}) = \frac{1}{8}$, $\omega(\boldsymbol{s}, \boldsymbol{g}) = \frac{1}{8}$ and $\omega(\boldsymbol{s}, \boldsymbol{i}) = \frac{1}{16}$.

A.3. Multiscale global relevance score

In the multiscale extension, local relevance scores computed for patches at the highest scale integrate local relevance scores computed at lower scales (see Eq. (A.1)). As a consequence, the global relevance score simply is:

$$r(I) = \sqrt[m]{\sum_{x,y} \left| r(I_{J,x,y}) \right|^m} \tag{A.3}$$

Appendix B. Generalization to higher-dimensional images

Should the proposed framework be applied to n-dimensional images, $n \neq 2$, the only things that need to be changed is the shape and the spatial distribution of patches (see Section 5). patches are no longer rectangles but, more generally, parallelotopes. patches are no longer organized in K^2 patch grids per scale (see Fig. 2) but in K^n patch grids. Sections 6 and 7 are unchanged. In A, function ω needs to be adapted to the new shape and the new spatial distribution of patches. Note that time can be one of the n dimensions.

References

Abràmoff, M.D., Suttorp-Schulten, M.S. A., 2005. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemed. J. E. Health 11 (6), 668–674.

Agurto, C., Murray, V., Barriga, E., Murillo, S., Pattichis, M., Davis, H., Russell, S.R., Abràmoff, M.D., Soliz, P., 2010. Multiscale AM-FM methods for diabetic retinopathy lesion detection. IEEE Trans. Med. Imag. 29 (2), 502–512.

Andrews, S., Tsochantaridis, I., Hofmann, T., 2003. Support vector machines for multiple-instance learning. Advances, in Neural Information Processing Systems 15, pp. 561–568.

Antal, B., Lázár, I., Hajdu, A., Török, Z., Csutak, A., Peto, T., 2011. Evaluation of the grading performance of an ensemble-based microaneurysm detector. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 5943–5946.

Azimi-Sadjadi, M.R., Salazar, J., Srinivasan, S., 2009. An adaptable image retrieval system with relevance feedback using kernel machines and selective sampling. IEEE Trans. Image Process. 18 (7), 1645–1659.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chaum, E., Karnowski, T.P., Govindasamy, V.P., Abdelrahman, M., Tobin, K.W., 2008. Automated diagnosis of retinopathy by content-based image retrieval. Retina 28 (10), 1463–1477.

Chen, Y., Wang, J., 2004. Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. 5, 913–939.

Datta, R., Joshi, D., Li, J., Wang, J.Z., 2008. Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. 40 (2), 1–60.

Donald, K., Smeaton, A., July 2005. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In: Proc. Int. Conf. Image Video Retrieval, pp. 61–70.

Erginay, A., Chabouis, A., Viens-Bitker, C., Robert, N., Lecleire-Collet, A., Massin, P., 2008. OPHDIAT: quality-assurance programme plan and performance of the network. Diabetes Metab. 34 (3), 235–242.

Giancardo, L., Meriaudeau, F., Karnowski, T.P., Li, Y., Garg, S., Tobin, K.W., Chaum, E., 2012. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med. Image Anal. 16 (1), 216–226.

Giancardo, L., Meriaudeau, F., Karnowski, T.P., Li, Y., Tobin, K.W., Chaum, E., 2011. Microaneurysm detection with radon transform-based classification on retina images. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 5939–5942.

Hunter, R.S., 1958. Photoelectric color difference meter. J. Opt. Soc. Am. 48 (12), 985–993.

Joblove, G.H., Greenberg, D., 1978. Color spaces for computer graphics. SIGGRAPH Comput. Graph. 12 (3), 20–25.

Kauppi, T., Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Voutilainen, A.R. R., Pietilä, J., Kälviäinen, H., Uusitalo, H., 2007. DIARETDB1 diabetic retinopathy database and evaluation protocol. In: Proc. of the 11th Conf. on Medical Image Understanding and Analysis (Aberystwyth, Wales, 2007).

Klonoff, D., Schwartz, D., 2000. An economic analysis of interventions for diabetes. Diabetes Care 23, 390–404.

Ko, B., Byun, H., 2002. Integrated region-based image retrieval using region's spatial relationships. In: Proc. IEEE Int. Conf. Pattern Recognition, vol. 1, pp. 196–199.

Li, Y., Karnowski, T.P., Tobin, K.W., Giancardo, L., Morris, S., Sparrow, S.E., Garg, S., Fox, K., Chaum, E., 2011. A health insurance portability and accountability act-compliant ocular telehealth network for the remote diagnosis and management of diabetic retinopathy. Telemed. J. E. Health 17 (8), 627–634.

Maron, O., Lozano-Pérez, T., 1998. A framework for multiple-instance learning. In: Proc. Conf. Advances in Neural Information Processing Systems (NIPS 98), pp. 570–576.

Maron, O., Ratan, A., 1998. Multiple-instance learning for natural scene classification. In: Proc. 15th Int'l Conf. Machine Learning (ICML'98), pp. 341–349.

Massin, P., Chabouis, A., Erginay, A., Viens-Bitker, C., Lecleire-Collet, A., Meas, T., Guillausseau, P.J., Choupot, G., André, B., Denormandie, P., 2008. OPHDIAT: a telemedical network screening system for diabetic retinopathy in the Ile-de-France. Diabetes Metab. 34 (3), 227–234.

Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. Int. J. Med. Inform. 73 (1), 1–23.

Niemeijer, M., Abràmoff, M.D., van Ginneken, B., 2009. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. IEEE Trans. Med. Imag. 28 (5), 775–785.

Niemeijer, M., van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., Abràmoff, M.D., 2010. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imag. 29 (1), 185–195.

Ohta, Y.-I., Kanade, T., Sakai, T., 1980. Color information for region segmentation. Comput. Graph. Image Process. 13 (3), 222–241.

Oliveira, C.M., Cristóvao, L.M., Ribeiro, M.L., Abreu, J.R., 2011. Improved automated screening of diabetic retinopathy. Ophthalmologica 226 (4), 191–197.

Philip, S., Fleming, A.D., Goatman, K.A., Fonseca, S., McNamee, P., Scotland, G.S., Prescott, G.J., Sharp, P.F., Olson, J.A., 2007. The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme. Br. J. Ophthalmol. 91 (11), 1512–1517.

Quellec, G., Lamard, M., Cazuguel, G., Bekri, L., Daccache, W., Roux, C., Cochener, B., 2011a. Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs. Invest. Ophthalmol. Vis. Sci. 52 (1), 8342–8348.

Quellec, G., Lamard, M., Cazuguel, G., Cochener, B., Roux, C., 2010a. Adaptive nonseparable wavelet transform via lifting and its application to content-based image retrieval. IEEE Trans. Image Process. 19 (1), 25–35.

Quellec, G., Lamard, M., Cazuguel, G., Cochener, B., Roux, C., 2010b. Wavelet optimization for content-based image retrieval in medical databases. Med. Image. Anal. 14 (2), 227–241.

Quellec, G., Russell, S.R., Abràmoff, M.D., 2011b. Optimal filter framework for automated instantaneous detection of lesions in retinal images. IEEE Trans. Med. Imag. 30 (2), 523–533.

Rahmani, R., Goldman, S.A., Zhang, H., Cholleti, S.R., Fritts, J.E., 2008. Localized content-based image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 30 (11), 1902–1912.

Ramaswamy, S., Rose, K., 2009. Towards optimal indexing for relevance feedback in large image databases. IEEE Trans. Image Process. 18 (12), 2780–2789.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover distance as a metric for image retrieval. Int. J. Comput. Vis. 40 (2), 99–121.

Sánchez, C. I., Niemeijer, M., Abràmoff, M. D., van Ginneken, B., 2010. Active learning for an efficient training strategy of computer-aided diagnosis systems: application to diabetic retinopathy screening. In: Med. Image Comput. Comput. Assist. Interv., vol. 13, pp. 603–610.

Schölkopf, B., Smola, A.J., 2002. Learning with kernels: support vector machines. Regularization. Optimization and Beyond. MIT Press.

Sivic, J., Zisserman, A., October 2003. Video google: a text retrieval approach to object matching in videos. In: Proc. IEEE Int. Conf. Comput. Vision, vol. 2, pp. 1470–1477.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22 (12), 1349–1380.

Walter, T., Massin, P., Erginay, A., Ordonez, R., Jeulin, C., Klein, J.C., 2007. Automatic detection of microaneurysms in color fundus images. Med. Image Anal. 11 (6), 555–566.

Zhang, J., Ye, L., 2009. Content based image retrieval using unclean positive examples. IEEE Trans. Image Process. 18 (10), 2370–2375.