WEAKLY SUPERVISED CLASSIFICATION OF MEDICAL IMAGES

G. Quellec[†] *M. Lamard*[⋆†] *G. Cazuguel*[△†] *M. D. Abràmoff*[▽] *B. Cochener*[⋆†○] *Ch. Roux*[△†]

[†] Inserm, U650, IFR 148 ScInBioS, Brest, F-29200 France;
[⋆] Univ Bretagne Occidentale, Brest, F-29200 France;
[△] INSTITUT TELECOM; TELECOM Bretagne; UEB; Dpt ITI, Brest, F-29200 France;
[▽] Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, IA 52242, USA;
[○] CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France;

ABSTRACT

A weakly supervised image classification framework is presented in this paper. Given reference images marked by clinicians as relevant or irrelevant, we learn to automatically detect relevant patterns, i.e. patterns that only appear in relevant images. After training, relevant patterns are sought in unseen images in order to classify each image as relevant or irrelevant. No manual segmentations are required. Because manual segmentation of medical images is extremely time-consuming, existing classification algorithms are usually trained on limited reference datasets. With the proposed framework, much larger medical datasets are now available for training. The proposed approach has been successfully applied to diabetic retinopathy detection in a retinal image dataset (A_z=0.855).

Index Terms— image classification, weak supervision, diabetic retinopathy

1. INTRODUCTION

The continually increasing amount of images stored in medical databases makes data-driven knowledge acquisition 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). Therefore, extraction of a limited subset of the available data, for clinician interpretation and algorithm supervision, is common practice. A different approach is explored in this paper: we believe similar or better detection performance can be achieved through limited clinician interpretation, so long as a larger reference dataset is available. In this paper, 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 relevant pattern. Assigning Boolean interpretations to medical images is often part of the clinical protocol, so after de-identification, images may be used directly for research without additional work for clinicians.

One medical application where many images are available is Diabetic Retinopathy (DR) detection. 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]. As a consequence, large datasets of digitized eye fundus photographs have been collected. One of these datasets, MESSIDOR[1] (1,200 interpreted images), is freely available. However, MESSIDOR images have not been manually segmented, so they cannot be used to supervise state-of-the-art DR detection methods [4], which are based on strongly supervised lesion detection. Only smaller image datasets come with manual segmentations: ROC (100 images), DIARETDB1 (89 images), etc. Weakly supervised frameworks are therefore needed to make the most of larger datasets such as MESSIDOR.

A general framework for automatically finding relevant patterns in images is presented. Its purpose is to automatically define 1) a local *relevance score* for sub-images of varying sizes in the reference image dataset and, at the end of the training phase, 2) a local relevance score for sub-images in unseen images. Once a local relevance score has been defined for each sub-image in an unseen image, a global relevance score is defined for the entire image. The proposed framework fits into the Multiple-Instance Learning (MIL) paradigm. In MIL, the task is to learn a *concept* given relevant and irrelevant sets (or bags) of feature vectors. A bag is labeled *relevant* if and only if at least one of its elements falls into that concept [5].

2. METHOD OUTLINE

Let I be an image of size $\boldsymbol{l} = \{l_1, l_2\} \in (\mathbb{N}^*)^2$. Sub-images of I are extracted at each resolution $j \in \{1, 2, ..., J\}$, $J \in \mathbb{N}^*$. The size of each sub-image of I, at resolution j, is given by

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

$\frac{1}{j}\boldsymbol{l}$. At a given resolution, rectangular sub-images are organized in g^2 overlapping grids, $g \in \mathbb{N}^*$ (see figure 1). Each grid is shifted from the origin by an offset vector $\boldsymbol{\delta}^{(j)}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma} \in \{1, ..., g\}^2$:

$$\boldsymbol{\delta}^{(j)}(\boldsymbol{\gamma}) = \left\{ \frac{(\boldsymbol{\gamma}_1 - 1)\boldsymbol{l}_1}{jg}, \frac{(\boldsymbol{\gamma}_2 - 1)\boldsymbol{l}_2}{jg} \right\} \tag{1}$$

g controls the number of analyzed sub-images per resolution. For $g = 1$, there is one grid per resolution: sub-images form a partition of I. For $g > 1$, sub-images at each resolution overlap.

Before defining a global relevance score for I, a local relevance score is first defined for each of its sub-images (see section 3). In that purpose, each sub-image is described by a feature vector $\boldsymbol{s}$. Let $I(\boldsymbol{s}, j)$ denote the sub-image of I (at resolution j) described by $\boldsymbol{s}$. To compute the local relevance score for $I(\boldsymbol{s}, j)$, we search for the nearest neighbors of $\boldsymbol{s}$ in a reference dataset.

During the training phase, the importance of each feature vector in the reference dataset is weighted (see section 4) in order to adapt the local relevance scores to our definition for *relevance*.

Finally, the global relevance score for I is defined as the L^m norm ($m \in \mathbb{N}^*$) of the vector consisting of all local relevance scores computed for sub-images of I.

3. LOCAL RELEVANCE SCORE

Given a definition for relevance, $-$ and $+$ superscripts indicate *irrelevance* and *relevance*, respectively. In particular, $\mathcal{I}^-$ (resp. $\mathcal{I}^+$) denotes the set of all irrelevant (resp. relevant) images in the reference dataset.

Let $I \in \mathcal{I}^+ \cup \mathcal{I}^-$ be an image from the reference dataset. Let $\boldsymbol{s}$ be a feature vector describing a sub-image of I at resolution j. Given that the class label ($+$ or $-$) is global (it is assigned to I as a whole), no local class information is available for $\boldsymbol{s}$. However, we can take advantage of the class labels assigned to all other images in the reference dataset ($(\mathcal{I}^+ \cup \mathcal{I}^-) \setminus \{I\}$) to define a (local) relevance score for $I(\boldsymbol{s}, j)$. This relevance score derives from the distance between $\boldsymbol{s}$ and its $k \in \mathbb{N}^*$ nearest neighbors in $(\mathcal{I}^+ \cup \mathcal{I}^-) \setminus \{I\}$, according to a Minkowski metric $\|.\|_m$ ($m \in \mathbb{N}^*$). The number of neighbors originating from the same reference image $J \in (\mathcal{I}^+ \cup \mathcal{I}^-) \setminus \{I\}$ is limited to $\widehat{k} \in \mathbb{N}^*$. Let $\mathcal{N}^-(\boldsymbol{s})$ (resp. $\mathcal{N}^+(\boldsymbol{s})$) denote the set of all neighbors originating from irrelevant (resp. relevant) images. Let $R(\boldsymbol{s})$ denote the radius of $\boldsymbol{s}$'s neighborhood:

$$R(\boldsymbol{s}) \triangleq \max_{\boldsymbol{t} \in \mathcal{N}^+(\boldsymbol{s}) \bigcup \mathcal{N}^-(\boldsymbol{s})} \|\boldsymbol{s} - \boldsymbol{t}\|_m \tag{2}$$

Let $p(\boldsymbol{s}) \in [0; 1]$ denote the membership degree of $\boldsymbol{s}$ to $\mathcal{I}^+$:

$$\begin{cases} N^+(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t} \in \mathcal{N}^+(\boldsymbol{s})} \lambda(\boldsymbol{t}) \left[R(\boldsymbol{s})^m - \|\boldsymbol{s} - \boldsymbol{t}\|_m^m \right] \\ N^-(\boldsymbol{s}) \triangleq \sum\limits_{\boldsymbol{t} \in \mathcal{N}^-(\boldsymbol{s})} \lambda(\boldsymbol{t}) \left[R(\boldsymbol{s})^m - \|\boldsymbol{s} - \boldsymbol{t}\|_m^m \right] \\ N(\boldsymbol{s}) \triangleq N^+(\boldsymbol{s}) + N^-(\boldsymbol{s}) \\ p(\boldsymbol{s}) \triangleq \dfrac{N^+(\boldsymbol{s})}{N(\boldsymbol{s})} \end{cases} \tag{3}$$

where $\lambda(\boldsymbol{t})$ is a weight associated with each reference feature vector $\boldsymbol{t}$ ($\lambda(\boldsymbol{t}) = 1$, initially). $p(\boldsymbol{s})$ increases with the percentage of feature vectors originating from relevant images in the neighborhood of $\boldsymbol{s}$, and also with the similarity of these feature vectors to $\boldsymbol{s}$.

The relevance score for $I(\boldsymbol{s}, j)$, denoted by $\Delta_p(\boldsymbol{s})$, is defined as follows:

$$\Delta_p(\boldsymbol{s}) \triangleq p(\boldsymbol{s}) - p_0 \tag{4}$$

where p_0 is the percentage of feature vectors in the reference dataset originating from relevant images. If $I(\boldsymbol{s}, j)$ is relevant, then $\Delta_p(\boldsymbol{s})$ is expected to be greater than zero. Otherwise, $\Delta_p(\boldsymbol{s})$ is expected to be close to zero.

4. LOCALIZED WEIGHT UPDATING

Let $I \in \mathcal{I}^-$ be a reference image labeled as irrelevant. Let $\boldsymbol{s}$ be a feature vector describing a sub-image of I. By definition of $\mathcal{I}^-$, we know that $\boldsymbol{s}$ is an irrelevant pattern. As a consequence, the relevance score for $I(\boldsymbol{s}, j)$ should be non-positive: $\Delta_p(\boldsymbol{s}) \leq 0$ (i.e. $p(\boldsymbol{s}) \leq p_0$). If $\Delta_p(\boldsymbol{s}) > 0$, then the relevance score definition should be updated. To do so, we modify the weight $\lambda(\boldsymbol{t})$ of all feature vectors $\boldsymbol{t}$ in the neighborhood of $\boldsymbol{s}$. Note that $\lambda(\boldsymbol{t})$ is global. Local optimal solutions (obtained as described in section 4.1) are combined into a global solution as described in section 4.2.

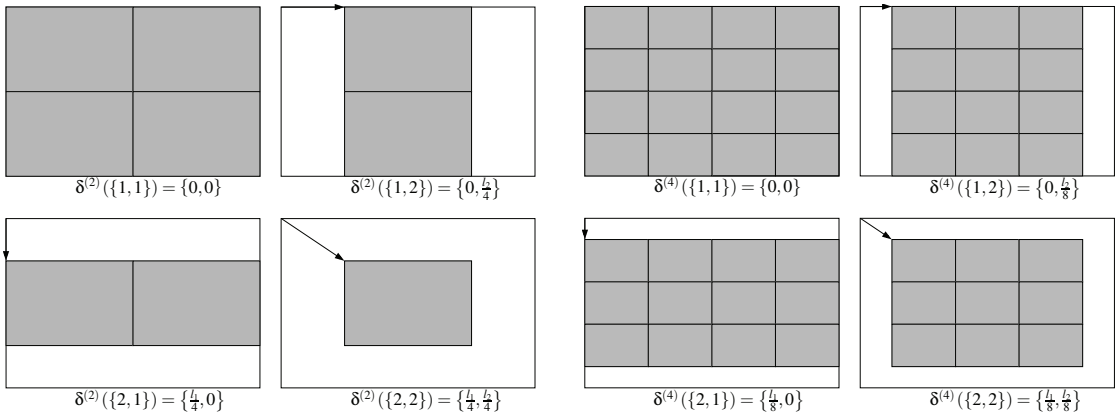
4.1. Locally Optimal Weight Updates

If $\Delta_p(\boldsymbol{s}) > 0$, then the weight of each feature vector $\boldsymbol{t} \in \mathcal{N}^+(\boldsymbol{s})$ (i.e. each feature vector, originating from an relevant image, in the neighborhood of $\boldsymbol{s}$) is multiplied by a coefficient $x(\boldsymbol{s}) \in \mathbb{R}$ such that $\Delta_p(\boldsymbol{s}) = 0$ after multiplication. This is equivalent to multiplying $N^+(\boldsymbol{s})$ by $x(\boldsymbol{s})$ (see equation 3). In order to find the optimal $x(\boldsymbol{s})$ coefficient, the following equation must be solved (see equations 3 and 4):

$$\frac{N^+(\boldsymbol{s})x(\boldsymbol{s})}{N^+(\boldsymbol{s})x(\boldsymbol{s}) + N^-(\boldsymbol{s})} - 0 = p_0 = \frac{N^+(\boldsymbol{s})}{N(\boldsymbol{s})} - \Delta_p(\boldsymbol{s}) \tag{5}$$

or, equivalently:

$$x(\boldsymbol{s}) = \frac{N^+(\boldsymbol{s})N^-(\boldsymbol{s}) - \Delta_p(\boldsymbol{s})\, N^-(\boldsymbol{s})N(\boldsymbol{s})}{N^+(\boldsymbol{s})N^-(\boldsymbol{s}) + \Delta_p(\boldsymbol{s})\, N^+(\boldsymbol{s})N(\boldsymbol{s})} \tag{6}$$



(a) Four sub-image grids at resolution $j = 2$

(b) Four sub-image grids at resolution $j = 4$

Fig. 1: Sub-image geometry. Let I be an image of size $l_1 \times l_2$. Figures (a) and (b) show the location of the sub-images of I at resolution $j = 2$ and $j = 4$, respectively, for $g = 2$. Four sub-image grids, shifted from the origin by an offset vector $\boldsymbol{\delta}^{(j)}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma} \in \{1, 2\}^2$, are created at each resolution. Offset vectors are represented by arrows.

4.2. Global Weight Updating

A reference feature vector $\boldsymbol{t}$ may appear in the neighborhood of several feature vectors: this is quite likely since, on average, each reference feature vector appears in the neighborhood of k feature vectors during the training phase ($k \gg 1$). Let $\boldsymbol{s}^{(e)}$, $e = 1..E$, be the feature vectors originating from irrelevant images such that $\Delta_p\left(\boldsymbol{s}^{(e)}\right) > 0$ and such that $\boldsymbol{t}$ appears in the neighborhood of $\boldsymbol{s}^{(e)}$. $\lambda(\boldsymbol{t})$, the weight of each feature vector $\boldsymbol{t}$, is multiplied by a coefficient $y(\boldsymbol{t}) \in \mathbb{R}$; $y(\boldsymbol{t})$ is chosen in order to try and satisfy equation 6 for each feature vector $\boldsymbol{s}^{(e)}$, $e = 1..E$. The following system should therefore be solved:

$$\begin{cases} \boldsymbol{a} y(\boldsymbol{t}) & = \boldsymbol{b} \\ \boldsymbol{a}_e & \triangleq 1 \\ \boldsymbol{b}_e & \triangleq x(\boldsymbol{s}^{(e)}) \end{cases} \tag{7}$$

$e = 1..E$. This system has one unknown, $y(\boldsymbol{t})$, and E equations. System 7 is solved in the least mean squared sense, subject to $y(\boldsymbol{t}) \geq 0$, after weighting each equation e by $\Delta_p\left(\boldsymbol{s}^{(e)}\right)$. For equation weighting, $\boldsymbol{a}$ and $\boldsymbol{b}$ are replaced by $\tilde{\boldsymbol{a}}$ and $\tilde{\boldsymbol{b}}$, respectively:

$$\begin{cases} \tilde{\boldsymbol{a}}_e & \triangleq \Delta_p\left(\boldsymbol{s}^{(e)}\right) \\ \tilde{\boldsymbol{b}}_e & \triangleq \Delta_p\left(\boldsymbol{s}^{(e)}\right) x(\boldsymbol{s}^{(e)}) \end{cases} \tag{8}$$

Let G be the function to minimize:

$$\begin{aligned} G\left(y(\boldsymbol{t})\right) &\triangleq \left(\tilde{\boldsymbol{a}} y(\boldsymbol{t}) - \tilde{\boldsymbol{b}}\right)^T \left(\tilde{\boldsymbol{a}} y(\boldsymbol{t}) - \tilde{\boldsymbol{b}}\right) \\ &= \tilde{\boldsymbol{a}}^T \tilde{\boldsymbol{a}} y(\boldsymbol{t})^2 - 2\tilde{\boldsymbol{a}}^T \tilde{\boldsymbol{b}} y(\boldsymbol{t}) + \tilde{\boldsymbol{b}}^T \tilde{\boldsymbol{b}} \end{aligned} \tag{9}$$

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

$$G'\left(y(\boldsymbol{t})\right) = 2\tilde{\boldsymbol{a}}^T \tilde{\boldsymbol{a}} y(\boldsymbol{t}) - 2\tilde{\boldsymbol{a}}^T \boldsymbol{b} = 0 \tag{10}$$

or, equivalently, when:

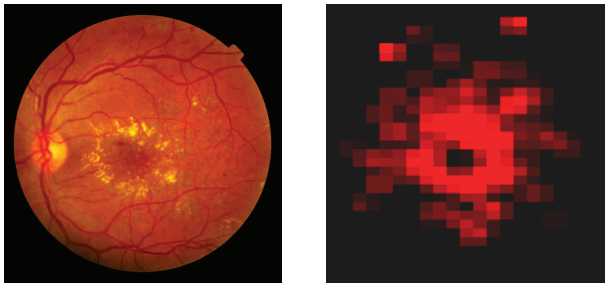
$$\widehat{y(\boldsymbol{t})} \triangleq \frac{\tilde{\boldsymbol{a}}^T \tilde{\boldsymbol{b}}}{\tilde{\boldsymbol{a}}^T \tilde{\boldsymbol{a}}} = \frac{\sum_{e=1}^{E} \Delta_p\left(\boldsymbol{s}^{(e)}\right)^2 x(\boldsymbol{s}^{(e)})}{\sum_{e=1}^{E} \Delta_p\left(\boldsymbol{s}^{(e)}\right)^2} \tag{11}$$

If $\widehat{y(\boldsymbol{t})} < 0$, then constraint $y(\boldsymbol{t}) \geq 0$ is infringed when $y(\boldsymbol{t}) = \widehat{y(\boldsymbol{t})}$. However, in that case, it can be easily checked that $G_{|y(\boldsymbol{t}) \geq 0}$ is strictly increasing ($G'\left(y(\boldsymbol{t})\right) > 0, \forall y(\boldsymbol{t}) \geq 0$): therefore, if $\widehat{y(\boldsymbol{t})} < 0$, the optimal solution is $y(\boldsymbol{t}) = 0$. If $\widehat{y(\boldsymbol{t})} \geq 0$, then the optimal solution is $y(\boldsymbol{t}) = \widehat{y(\boldsymbol{t})}$.

5. APPLICATION TO DIABETIC RETINOPATHY DETECTION IN THE MESSIDOR DATASET

The MESSIDOR dataset consists of 1200 eye fundus color photographs. It was collected for the assessment of automated DR screening algorithms [2]. Images were acquired by 3 ophthalmology departments, in France, using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. Images were captured using 8 bits per color plane at 1440x960, 2240x1488 or 2304x1536 pixels. 800 images were acquired with pupil dilation (one drop of Tropicamide at 10%) and 400 without dilation.

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



(a) Original image (b) Local abnormality scores

Fig. 2: Detection Example. Local abnormality scores in each sub-image are proportional to the red intensity level.

Table 1: Classification performance on the MESSIDOR test subset (with $g = 2$, $m = 2$ and $\hat{k} = 1$)

resolution	1	2	3	4	5
A_z	0.719	0.740	0.835	**0.855**	0.687
T	0.552s	0.651s	1.02s	3.34s	12.7s

In this experiment, an image is considered relevant if and only if the patient has been labeled as having Diabetic Retinopathy (DR) by a retinal expert. 654 images out of 1200 have been annotated as relevant.

Sub-images were described using textural and shape features in the Lab color space. Zernike moments were used for local shape description [6]. Wavelet-based features were used for texture characterization [7].

The dataset was divided into a training and a test subset (of 600 images each). A detection example is given in figure 2. Classification performance was measured in terms of area under the Receiver-Operating Characteristic (ROC) curve, noted A_z. A_z values are reported in table 1, together with the total processing time per image (T). All computations were carried out on one core of an Intel Xeon E5520 processor running at 2.27GHz. One reason why performance drops at resolution 5 is that the wavelet-based texture features are less relevant in small sub-images, due to the lack of data.

6. DISCUSSION

A weakly supervised image classification framework has been presented in this paper. Unlike competing frameworks, such as Rahmani's ACCIO! [8], objects of interest are identified by user-defined, problem-specific features, 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).

High classification performances have been obtained for DR detection: an area under the ROC curve of A_z=0.855 has been achieved, which is higher than previously published results (A_z=0.84) [9].

In this paper, adapting local relevance scores involved weighting reference feature vectors. In future works, image features (texture, color, or shape features) could also be weighted, in a similar fashion, to improve performance further. This framework, which releases clinicians from the time-consuming task of manual segmentation, could well become clinicians' best friend.

7. REFERENCES

[1] D. Klonoff and D. Schwartz, "An economic analysis of interventions for diabetes," *Diabetes Care*, vol. 23, 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] 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, pp. 227–234, 2008.

[4] 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.

[5] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc Conf Advances in Neural Information Processing Systems (NIPS 98)*, 1998, pp. 570–576.

[6] 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, May 1990.

[7] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Adaptive nonseparable wavelet transform via lifting and its application to content-based image retrieval," *IEEE Trans Image Process*, vol. 19, no. 1, pp. 25–35, January 2010.

[8] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts, "Localized content-based image retrieval," *IEEE Trans Pattern Anal Mach Intell*, vol. 30, no. 11, pp. 1902–1912, November 2008.

[9] C. Agurto, V. Murray, E. Barriga, S. Murillo, M. Pattichis, H. Davis, S. R. Russell, M. D. Abràmoff, and P. Soliz, "Multiscale AM-FM methods for diabetic retinopathy lesion detection," *IEEE Trans Med Imaging*, vol. 29, no. 2, pp. 502–512, 2010.