A Discriminatively Trained Fully Connected Conditional Random Field Model for Blood Vessel Segmentation in Fundus Images

José Ignacio Orlando*, Elena Prokofyeva, and Matthew B. Blaschko

***Abstract*—*Objective:* In this work, we present an extensive description and evaluation of our method for blood vessel segmentation in fundus images based on a discriminatively trained, fully connected conditional random field model. *Methods:* Standard segmentation priors such as a Potts model or total variation usually fail when dealing with thin and elongated structures. We overcome this difficulty by using a conditional random field model with more expressive potentials, taking advantage of recent results enabling inference of fully connected models almost in real-time. Parameters of the method are learned automatically using a structured output support vector machine, a supervised technique widely used for structured prediction in a number of machine learning applications. *Results:* Our method, trained with state of the art features, is evaluated both quantitatively and qualitatively on four publicly available data sets: DRIVE, STARE, CHASEDB1 and HRF. Additionally, a quantitative comparison with respect to other strategies is included. *Conclusion:* The experimental results show that this approach outperforms other techniques when evaluated in terms of sensitivity, F1-score, G-mean and Matthews correlation coefficient. Additionally, it was observed that the fully connected model is able to better distinguish the desired structures than the local neighborhood based approach. *Significance:* Results suggest that this method is suitable for the task of segmenting elongated structures, a feature that can be exploited to contribute with other medical and biological applications.**

***Index Terms*—Blood vessel segmentation, Fundus imaging, Conditional Random Fields, Structured Output SVM.**

I. Introduction

AGE-related macular degeneration (AMD) (26%), glaucoma (20.5%) and diabetic retinopathy (8.9%) are the most frequent causes of preventable blindness in Europe [1]. These diseases are linked to such changes in fundus as change of shape and structure of vessels and lesions that are often easy to detect using fundus images.

Fundus photographs (Fig. 1a) are projective color images of the inner surface of the human eye. Such images are widely used as they allow physicians to examine in a non-invasive way the retina and its anatomical components, including the vascular tree, the optic disc and the fovea [2].

The development of automatic tools for the early detection of retinal diseases is valuable since they can be easily integrated in screening programs, where large numbers of images are taken from patient populations, and careful evaluation by physicians is not feasible in a reasonable time [3]. These tools are usually aided by the analysis of morphological attributes of retinal blood vessels, which provide valuable information for the diagnosis, screening, treatment and evaluation of the previously mentioned diseases [3]. In other cases, vessels need to be previously detected in order to facilitate the automation of the detection of lesions with similar intensities [4].

However, any automated analysis of the retinal vasculature requires its accurate segmentation first. In current best practice, this task is performed manually by trained experts, although this is particularly tedious and time-consuming. Furthermore, difficulties in the imaging process–such as inadequate contrast between vessels and background, and uneven background illumination–and the variability of vessel width, brightness and shape, reduce significantly the coincidence among segmentations performed by different human observers [5]. These facts motivate the development of automatic strategies for blood vessel segmentation without human intervention [3].

Although numerous attempts have been made in the field of automated retinal vessel segmentation, this task is still an active area of research due to the potential impact of having more accurate results [2]. In general, existing approaches can be classified into two main categories, supervised and unsupervised. Supervised methods require a set of training samples–typically composed of pixels features and their known annotations–to learn a model or a classifier. Several classifiers have been considered in the literature, including k-nearest neighbors [6], Bayesian [7], support vector machines [8], [9], neural networks [10], [11], decision trees [12], [13], Gaussian mixture models [5], AdaBoost [14], among others. A trainable filter, named B-COSFIRE, was recently introduced in [15] to highlight the retinal vasculature. Though the method is not supervised in the sense of training a classifier, the strategy they follow to adjust its parameters is based on training data. By contrast, unsupervised methods are systems that are able to segment the vasculature without requiring any manual annotations, although typically at the cost of lower accuracy. In general, most of these strategies are based on applying thresholding, vessel tracking techniques [16] or region-oriented approaches–such as region growing [17]–[19]

*José Ignacio Orlando is with National Council for Scientific and Technological Research (CONICET) and Pladema Institute, Tandil, Buenos Aires, 7000 Argentina. E-mail: jiorlando@conicet.gov.ar.

Elena Prokofyeva is with Inserm, U1018, University of Versailles Saint-Quentin, Villejuif, France, and Northern State Medical University, Troitsky av. 51, Arkhangelsk, Russia. E-mail: elena.prokofyeva@inserm.fr.

Matthew Blaschko is with ESAT-PSI-Visics, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. E-mail: Matthew.Blaschko@kuleuven.be. Work was done in part at Inria Scalay & CentraleSupélec, Grande Voie des Vignes, 92295 Châtenay-Malabry, France.

Manuscript received Month Day, Year; revised Month Day, Year.

or active contours [20], [21]–after vessel enhancement. This task is performed by means of morphological operations [22], matched filter responses [23]–[25], the complex continuous wavelet transform [26], among others [27]. The method we propose in this paper belongs to the supervised category.

Conditional Random Fields (CRFs) are extensively used for image segmentation in several applications [28]–[30]. To the best of our knowledge, however, they were never applied before to blood vessel segmentation in fundus images. This is likely due to that the standard pairwise potentials, such as in a Potts model, assign a low prior to the elongated structures that comprise a vessel segmentation. This fact motivated us to introduce a novel method for blood vessel segmentation based on fully connected CRFs [31], which we extend in this article. Fully connected CRFs were previously applied in [32] and [33] for liver and brain tumor segmentation in CT and MRI, but their implications on the segmentation of two dimensional, thin structures was not previously studied. In this work we demonstrate that the dense connectivity augments the capability of the method to detect elongated structures, overcoming the original difficulty of local neighborhood based CRFs and improving results significantly. This property can potentially contribute to a number of different biological and medical applications where the segmentation of such structures is required, including automatic plant root phenotyping [34] or neuron analysis [35].

As is shown in [36], local classification leads to misclassification issues that might arise while incorporating prior knowledge about the shape of the desired structures on the learning process. CRFs are able to provide such information through the pairwise potentials. Structured Output SVM (SOSVM) has been used before to learn local neighborhood based CRFs [37], [38]. However, learning dense CRFs using SOSVMs was avoided before due to its computational intractability, since the learning method requires multiple calls to the inference algorithm during training, and the inference in dense CRFs is usually slow. We overcome this problem by making use of recent advances in efficient inference in fully connected CRFs [36].

In this paper, we complement our previous work [31] with further information and implementation details. We also modify the strategy to estimate additional parameters of the method in order to optimize its performance during training. Additionally, we extend the validation of our results with an evaluation performed both quantitatively and qualitatively on four standard and publicly available data sets (DRIVE, STARE, CHASEDB1 and HRF) to study the behavior of the algorithm under different contexts, including images of healthy patients, containing pathologies and taken at different resolutions. According to our experiments, this method outperforms current strategies when evaluating in terms of several different quality measures.

The remainder of this paper is organized as follows: Section II explains in detail our method. In Section III we provide information about the data sets and the quantitative measures used in our experiments. Section IV presents our results, including a comparison to other recently published approaches. Section V discusses the advantages of the proposed method and further lines of research. Finally, Section VI concludes the paper. Supplementary materials provide extensive additional evaluation.

II. Methods

This section explains in detail our method. First, both the local neighborhood based and the fully connected CRF formulations are described (Section II-A). Afterwards, we summarize the strategy to learn such models by means of a SOSVM (Section II-B). The features used to evaluate our method are explained in Section II-C. Finally, Section II-D describes a compensation factor that can be used to segment images with different resolutions without needing to recalibrate feature parameters.

A. Conditional Random Fields for vessel segmentation

The segmentation task can be posed as an energy minimization problem in a conditional random field (CRF). In the original definition of CRFs, images are mapped to graphs, where each pixel represents a node, and every node is connected with an edge to their neighbors according to a certain connectivity rule [29], [36], [39]. In local neighborhood based CRFs, nodes are connected following a 4 pixel neighborhood connectivity [40], while in the fully connected definition each node is assumed to be linked to every other pixel of the image [36].

We denote by $\mathbf{y} = \{y_i\}$ a labeling over all pixels of the image I in the label space $\mathcal{L} = \{-1, 1\}$, where 1 is associated to blood vessels and -1 to any other class. A conditional random field $(I, \mathbf{y})$ is characterized by the Gibbs distribution:

$$p\left(\mathbf{y}|I\right) = \frac{1}{Z(I)} \exp\left(-\sum_{c \in \mathcal{C}_\mathcal{G}} \Phi_c(\mathbf{y}_c|I)\right) \tag{1}$$

where $Z(I)$ is a normalization constant, $\mathcal{G}$ is the graph associated to I and $\mathcal{C}_\mathcal{G}$ is a set of cliques in $\mathcal{G}$, each inducing a potential Φ_c [30]. This distribution states the conditional probability of a labeling $\mathbf{y}$ given the image I. The Gibbs energy function can be derived from this likelihood:

$$E(\mathbf{y}|I) = \sum_{c \in \mathcal{C}_\mathcal{G}} \Phi_c(\mathbf{y}_c|I) \tag{2}$$

Thus, the maximum a posteriori (MAP) labeling can be obtained by minimizing the corresponding energy:

$$\mathbf{y}^* = \arg\min_{\mathbf{y} \in \mathcal{L}} E(\mathbf{y}|I) \tag{3}$$

After minimizing $E(\mathbf{y}|I)$, a binary segmentation of the vasculature is obtained. For notational convenience, we will omit the conditioning in the rest of the paper, and we will use $\psi_c(\mathbf{y}_c)$ to denote $\Phi_c(\mathbf{y}_c|I)$. Additionally, we will consider energies that decompose as summations over unary and pairwise potentials, in contrast to more general higher order potentials [41].

Given a graph $\mathcal{G}$ on $\mathbf{y}$, its energy is obtained by summing its unary and pairwise potentials (ψ_u and ψ_p, respectively):

$$E(\mathbf{y}) = \sum_i \psi_u\left(y_i, \mathbf{x}_i\right) + \sum_{(i,j) \in \mathcal{C}_\mathcal{G}} \psi_p\left(y_i, y_j, \mathbf{f}_i, \mathbf{f}_j\right) \tag{4}$$

where $\mathbf{x}_i$ and $\mathbf{f}_i$ are the unary and pairwise features, respectively. Unary potentials define a log-likelihood over the label assignment $\mathbf{y}$, and they are traditionally computed by a classifier [36]. Pairwise potentials define a similar distribution but considering only the interactions between pixels features and their labels, according to $\mathcal{C}_\mathcal{G}$, which is determined by the graph connectivity.

Unary potentials are common to both the local neighborhood based and the fully connected CRF, and they are obtained as follows:

$$\psi_u\left(y_i, \mathbf{x}_i\right) = -\langle \mathbf{w}_{u_{y_i}}, \mathbf{x}_i \rangle - \mathbf{w}_{\beta_{y_i}} \beta \tag{5}$$

where β is a bias constant, and $\mathbf{w}_{u_{y_i}}$ and $\mathbf{w}_{\beta_{y_i}}$ represents the weight vectors for the features and the bias term, respectively, both associated to the label y_i. The unary vector $\mathbf{x}_i$ is given by an arbitrary combination of features extracted from the image.

Pairwise potentials are defined as a linear combination of functions. Thus, our pairwise energy is obtained according to:

$$\psi_p\left(y_i, y_j, \mathbf{f}_i, \mathbf{f}_j\right) = \mu(y_i, y_j) \sum_{m=1}^{M} {w_p}^{(m)} k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) \tag{6}$$

where each $k^{(m)}$ is a fixed function over an arbitrary feature $f^{(m)}$, ${w_p}^{(m)}$ is a linear combination weight, and $\mu(y_i, y_j)$ represents a label compatibility function. The Gaussian kernels determine the similarity between connected pixels by means of $f^{(m)}$. Since the neighboring information is provided by the connectivity rule followed by the model, these kernels depend on the CRF formulation, so they are described afterwards. The remaining terms are detailed in the sequel.

The compatibility function μ penalizes similar pixels that are assigned to different labels, and it is given by the Potts model $\mu(y_i, y_j) = [y_i \neq y_j]$, where Iverson bracket notation $[\cdot]$ indicates one if the statement is true and zero otherwise.

Parameters $\mathbf{w}_u$, ${\mathbf{w}_p}^{(m)}$ control the relevance of the unary features and the pairwise kernels on the energy function, respectively. Additionally, $\mathbf{w}_\beta$ is used to learn the bias term. The adjustment of these parameters is not feasible to be done manually due to their high dimensionality, so we propose to learn them using a Structured Output SVM, as is explained in detail in Section II-B.

1) Local neighborhood based CRFs: Local neighborhood based CRFs (LNB-CRFs) are defined over grid graphs. Thus, in this type of model each node (pixel) is assumed to be connected by an edge to its 4-connected neighbors. The function for the pairwise potentials given the m-th pairwise feature is obtained as follows:

$$k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) = \frac{|f_i^{(m)} - f_j^{(m)}|}{2\theta_{(m)}^2} \tag{7}$$

where $\theta_{(m)}$ is a bandwidth that controls the relevance of the dissimilarities between pixel features. The energy of the grid based model is minimized using the min-cut/max-flow approach proposed by [40].

2) Fully connected CRFs: In a fully connected CRF model (FC-CRF), each node of the graph is assumed to be linked to every other pixel of the image. Using these higher order potentials, the method is able to take into account not only neighboring information but also long-range interactions between pixels. This property improves the segmentation accuracy, but makes implementation of the inference process computationally expensive in general. Recently, however, Krähenbühl and Koltun [36] have introduced an efficient inference approach under the restriction that the pairwise potentials are a linear combination of Gaussian kernels over an Euclidean feature space. This approach, which is based on taking a mean field approximation of the original CRF, is able to produce accurate segmentations in a few seconds.

Pairwise kernels for the fully connected model have the following form:

$$k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) = \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\theta_p^2} - \frac{|f_i^{(m)} - f_j^{(m)}|^2}{2\theta_{(m)}^2}\right) \tag{8}$$

where $\mathbf{p_i}$ and $\mathbf{p_j}$ are the coordinate vectors of pixels i and j. Positions are included in the pairwise terms to increase the effect of close pixels over distant ones. Kernel widths θ_p and $\theta_{(m)}$ control the degree of relevance of the two parts of the kernels in the expression. For instance, when θ_p increases, much longer interactions are taking into account. On the contrary, when θ_p decreases, only local neighborhoods significantly affect the result. Similarly, when $\theta_{(m)}$ increases or decreases, higher or lower differences on the m-th feature are tolerated, respectively.

B. Learning CRFs with Structured Output SVM

Our goal is to learn a vector $\mathbf{w} = (\mathbf{w}_u, \mathbf{w}_\beta, \mathbf{w}_p)$, where $\mathbf{w}_u$, $\mathbf{w}_\beta$ and $\mathbf{w}_p$ are the weights for the unary features, for the bias term and for the pairwise kernels, respectively. The vector $\mathbf{w}$ can be high-dimensional if multiple features are considered, so manual or automated adjustment using techniques such as grid search is not feasible in a reasonable time. Supervised learning of the unary potentials separately from the pairwise potentials might be an alternative, but this approach ignores the influence of the pairwise potentials on the general energy formulation, and can lead to worse results than joint learning of the weights. We therefore propose to obtain $\mathbf{w}$ in a supervised way, using the 1-slack formulation of the SOSVM with margin-rescaling presented in [38]. Such a discriminative training approach has shown promising results for building highly complex and accurate models in several areas, including object detection, image segmentation and computer vision applications, even for large datasets. To the best of our knowledge, however, it was never used before for the task of learning FC-CRFs.

Let the training set $S = \{(s^{(1)}, y^{(1)}), ..., (s^{(n)}, y^{(n)})\}$, where n is the number of training images. Each $y^{(i)}$ corresponds to the ground truth of the i-th image in the training set. Each set $s^{(i)} = \{x^{(i)}, \beta, f^{(i)}\}$ contains the set $x^{(i)}$ of unary feature vectors, a bias constant $\beta = 1$, and the set $f^{(i)}$ of pairwise features for every pixel in the image.

The weights $\mathbf{w}$ are obtained by solving:

$$\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \tag{9}$$

subject to

$$\forall \left(\bar{y}^{(1)}, ..., \bar{y}^{(n)}\right) :$$
$$\sum_{i=1}^{n} \langle \mathbf{w}, \varphi(s^{(i)}, y^{(i)}) - \varphi(s^{(i)}, \bar{y}^{(i)})\rangle \geq \sum_{i=1}^{n} \Delta(y^{(i)}, \bar{y}^{(i)}) - \xi$$

where C is a regularization constant; ξ is a slack variable shared across all the constraints $\bar{y}^{(i)}$; $\varphi(s, y)$ is a feature map function that relates a given set s with a given labelling y; and $\Delta(y, \bar{y})$ is a loss function that evaluates the difference between a ground truth y and a constraint $\bar{y}$. In this work, we define Δ as the Hamming loss:

$$\Delta(y, \bar{y}) = \sum_{i} [y_i \neq \bar{y}_i] \tag{10}$$

where we used Iverson bracket notation. This function penalizes all the differences between the predicted labelling and the gold standard segmentation.

Our feature map is defined as follows:

$$\varphi(s, y) = \left(\sum_{k} \varphi_u(\mathbf{x}_k, y_k), \sum_{k} \varphi_\beta(\beta, y_k), \sum_{k}\sum_{j<k} \varphi_p\left(y_k, y_j, \mathbf{f}_k, \mathbf{f}_j\right) \right)$$

where the components represent the sum of the unary feature map, the bias feature map and the pairwise feature map, respectively, for all the pixels in the image. We give precise definitions of φ_u, φ_β, and φ_p in the sequel.

We define a binary vector $\varphi_y(y_i) \in \{0, 1\}^{|\mathcal{L}|}$ such that:

$$\varphi_y(y_i) = \begin{cases} (1, 0) & if \quad y_i = -1 \\ (0, 1) & if \quad y_i = 1 \end{cases} \tag{11}$$

The individual feature maps are obtained as follows:

$$\varphi_u(\mathbf{x}_k, y_k) = \mathbf{x}_k \otimes \varphi_y(y_k) \tag{12}$$

$$\varphi_\beta(\beta, y_i) = \beta\varphi_y(y_i) \tag{13}$$

$$\forall m : \left[\varphi_p\left(y_k, y_j, \mathbf{f}_k, \mathbf{f}_j\right)\right]_m = \mu(y_i, y_j) k^{(m)}(f_i^{(m)}, f_j^{(m)}) \tag{14}$$

where $\otimes$ is the Kronecker product. We solve Eq. (9) efficiently using the cutting-plane approach proposed in [38].

C. Features

We evaluated our method using features that are widely used in the field of blood vessel segmentation in fundus images: responses to the multiscale line detectors presented by Nguyen *et al.* [42] and responses to 2D Gabor wavelets [7] are used to compute the unary potentials, and a vessel-enhanced image processed with the method by Zana and Klein [22] for the pairwise potentials.

All features are extracted from grey scale images, obtained by taking the inverted green band of the original, RGB color image, as reported by other works [10], [15]. Additionally, due to false detections introduced by the selected features on the border of the FOV, we replicate the strategy proposed in [7] to simulate a wider aperture of the capture device. By means of this technique, false detections occurring outside the original FOV can be easily removed by multiplying the resulting image with the original FOV mask. An example of the resulting preprocessed image is shown in Fig. 1b.

Nguyen *et al.* line detectors exploit the property that blood vessels appear as elongated structures. The average intensity is calculated along a line of length l passing through each target pixel $\mathbf{P}$ at different orientation angles α. The line with the largest mean intensity $L_l(\mathbf{P})$ is selected from all the considered orientations, and the line strength of the pixel is computed by taking the difference $S_l(\mathbf{P}) = L_l(\mathbf{P}) - N_s(\mathbf{P})$, with $N_s(\mathbf{P})$ being the average intensity in a square window centered on $\mathbf{P}$ with edge length s. An example of the responses obtained with $l = 15$ is shown in Fig. 1c. The original version of this feature combines responses at different scales and the inverted green channel into a single feature, which is then thresholded. Here we take each S_l and the inverted image separately, since our method is able to learn the best weights to combine the features. Thus, instead of having a single value per pixel, we have a feature vector composed of the responses to each value of l and the image I.

2D Gabor wavelets have the capability to detect oriented features and can be tuned to specific frequencies. This property is especially useful to enhance the vasculature, since blood vessels appear at different sizes and orientations. We compute this feature exactly as reported by Soares *et al.* [7] at different scales a. Responses of the image to this wavelet, taken at different values of a, are included as features. Fig. 1d depicts an example obtained with $a = 3$.

Zana and Klein's technique for vessel enhancement takes advantage of the fact that the vessels are linear, connected, and their curvature varies smoothly along the crest line [22]. Noise of the image is first reduced by applying an opening by reconstruction operation, using linear structuring elements of length l at different angles. Afterwards, multiple top-hat morphological operations are applied using the same structuring elements, and the sum of the corresponding responses for each given angle is taken. This transformation reduces small bright noise and improves the contrast of all linear components. Structures whose curvature is linearly coherent are then detected by means of a cross-curvature evaluation, performed by applying a Laplacian of Gaussian with windows of size 7×7 pixels and standard deviation $7/4$. Finally, an alternating filter composed by successive application of a morphological opening, a closing and an opening is applied to remove false detections of non linear patterns on bright or dark thin irregular zones and background linear features. In the three last operations, the same linear structuring element of length l is used. We have observed that this feature is highly sensitive to uneven illumination of the fundus, degrading its ability to characterize the blood vessels effectively. In order to improve its quality we incorporated an additional prepro-

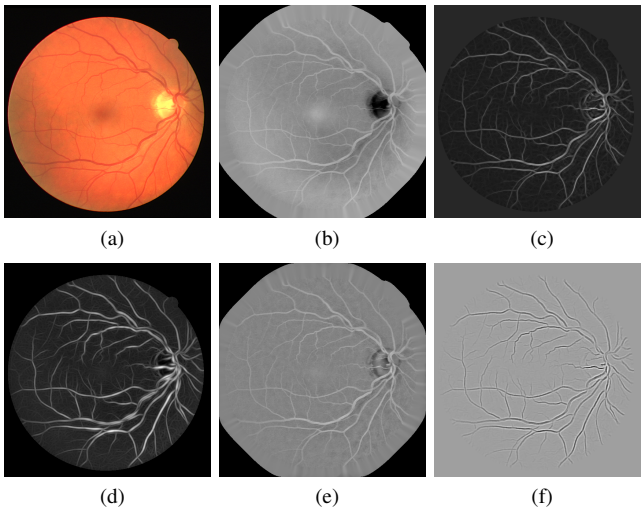


Fig. 1. Image preprocessing and unary and pairwise features examples. (a) Original color image. (b) Inverted green band after border expansion. (c) Response to Nguyen *et al.* line detector ($l = 15$). (d) Response to Soares *et al.* 2D Gabor wavelet at the scale $a = 3$. (e) Inverted image after bias correction. (f) Image enhanced using Zana and Klein method ($l = 9$).

cessing, only for this feature, where an estimated background is subtracted from the green band of the original color image. The background is estimated by convolving the green band with a median filter, where the size of the filter kernel is large enough to ensure that the blurred image contains no visible structures such as vessels. This approach has been applied several times in the literature [10], [43], and an example of the resulting image is illustrated in Fig. 1e. The Zana and Klein feature is illustrated in Fig. 1f.

All features are normalized independently to zero mean and unit variance, using the mean and the standard deviation of each feature calculated on each image [7].

D. Scaling Models to Images of Different Resolution

Although the weights for the unary features and the pairwise kernels are adjusted during the learning process, system performance is still related to the capability of the features to effectively characterize vascular structure. In general, features are sensitive to their parameters, which are usually related to vessel properties such as their calibre, which is at the same time related to the image resolution. Responses to the 2D Gabor wavelet, for example, depend on the scale a. Similarly, Nguyen *et al.* line detectors and the Zana and Klein enhancement strategy depend on the length l of the detectors or the linear structuring element, respectively. Most of these feature parameters were originally set using low resolution images, such as those in the DRIVE dataset [44]. When applying such features to higher resolution images, performance is significantly reduced if the feature extraction procedure is not proportionately scaled. Other parameters such as the angles for computing feature responses at different orientations are not influenced by changes in the resolution of the images.

A similar behavior can be expected for preprocessing parameters, e.g. the size of the median filter used to estimate the background, or the size of the aperture simulated by the border expansion. The parameter θ_p used on the pairwise potentials of the FC-CRF is also influenced by image resolution, since it weighs the pairwise interactions according to the relative distance of each pixel.

The proper adjustment of such parameters is relevant when applying the framework to images of different resolution, and having an automatic strategy for their calibration is valuable. Grid search using labelled images in the training set is computationally prohibitive due to the high dimensional space that comprises the parameters and their combinations. As an alternative to manual adjustment, some authors propose to derive parameters related to the vessel calibre from the width of the vessel of interest [15] or the size of the optic nerve head [45]. However, both methods require prior knowledge about structures that are difficult to measure and vary from one image to another. We introduce in the sequel a different strategy to automatically adapt features and model parameters to images of different resolution.

Instead of adjusting the configuration for each single data set resolution, we propose a simple approach based on estimating the best configuration of feature parameters on a single data set, and then adapting such parameters by multiplying them with a compensation factor $\rho = \frac{\mathcal{X}_{\text{new}}}{\mathcal{X}_{\text{training}}}$, where $\mathcal{X}_{\text{training}}$ represents the average width of the FOV in the images used to configure the scales, and $\mathcal{X}_{\text{new}}$ is the average width of the FOV in the new images. As changes in the resolution are expected to be related to changes in the number of pixels associated to the FOV region, this simple approach approximates invariance of feature computation with respect to scaling.

III. Materials and Evaluation

This section describe the data sets and the metrics used to evaluate our method. Additionally, we provide further details about the strategy followed to estimate the parameters θ_p and $\theta_{(m)}$ of the CRF, and the C parameter of the SOSVM (Eq. (9)).

A. Datasets

Our experiments were carried out on DRIVE [44], STARE [46], CHASEDB1 [12] and HRF [24], [47], four standard, publicly available data sets of fundus images used for the evaluation of blood vessel segmentation algorithms.

DRIVE[1] includes 40 color fundus photographs (7 of them with pathologies) with a 45° FOV, with 8 bits per color channel at 565×584 pixel resolution. The set is divided into a training and a test set, both containing 20 images. Two different manual annotations are provided for the test set, and only one annotation per image is available on the training set.

STARE[2] comprises 20 images, 10 of them containing pathologies, captured at 35° FOV, with 8 bits per color channel and a resolution of 700×605 pixels. Two observers manually segmented all images, with the second observer marking vessels with thinner annotations than the first one [5]. Despite this variability in annotation methodology between the two

[1] http://www.isi.uu.nl/Research/Databases/DRIVE/

[2] http://www.ces.clemson.edu/~ahoover/stare/probing/index.html

observers, performance is normally evaluated using the first observer's segmentation as the ground truth [10], [15]. FOV masks are not provided in the original set, so the masks built by Marin *et al.*[3] [10] were used. The set is not divided into training and test, and no consistent evaluation methodology has emerged from the reviewed literature. In order to be able to compare our method with respect to the largest amount of the available literature we performed our evaluation on STARE using leave-one-out cross-validation.

CHASEDB1[4] contains images of each eye of 14 children, comprising a total of 28 images. Pictures were captured with 30° FOV, using 8 bits per color channel at 1280×960 pixels, centered on the optic disc. Two expert labelings per image are provided. FOV masks were obtained using the approach proposed in [15], as they are not provided in the original dataset. The 28 images are divided into training and test, with 8 and 20 images in each set, respectively [12]. The first 20 images are used for testing, and the last 8 images for training.

The HRF[5] data set contains 15 images of healthy patients, 15 images of patients with diabetic retinopathy and 15 images of glaucomatous patients. Images were captured with 60° FOV and 3304×2336 pixel resolution. Only one ground truth segmentation per image is available, generated by a group of experts. To the best of our knowledge, this data set was not previously employed to evaluate supervised vessel segmentation algorithms, so we constructed a training set comprising the first 5 images of each subset, and tested on all remaining images. To reduce the computational cost of our experiments, images and labels on the training set and images on the test set were downsampled by a factor of 2, and results were afterwards upsampled so they can be compared with respect to the original manual annotations.

B. Evaluation Metrics

Results were analyzed quantitatively by comparing our segmentations with the gold standard labelings provided on each data set. Seven different measurements were obtained, all of them in terms of the number of true positives TP, true negatives TN, false positives FP and false negatives FN, and considering only the pixels inside the FOV:

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

$$F1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}, \quad G = \sqrt{Se \times Sp}$$

$$MCC = \frac{TP/N - S \times P}{\sqrt{P \times S \times (1-S) \times (1-P)}}$$

where $N = TP+TN+FP+FN$ is the total number of pixels of the image, $S = (TP+FN)/N$ and $P = (TP+FP)/N$. Sensitivity (Se, also known as Recall Re) measures the capability of the method to properly detect blood vessels, while specificity (Sp) is an indicator of the capability of distinguishing all other non-vessel structures. Sp suffers in the presence of imbalanced classes due to the low influence of the false positives term in the denominator of the fraction. By contrast, Se does not present this issue as it involves a fraction of pixels corresponding to the vessel class. However, though a higher Se value is desired, it must be analyzed in combination with Sp, as Se can be trivially maximized by labeling all pixels as vascular. Precision (Pr) quantifies the ratio of pixels classified as vessel that are correctly identified. The Accuracy (Acc) is not included as it is sensitive to unbalanced distributions in the number of pixels belonging to the positive and negative classes [15].[6] We include the Matthews Correlation Coefficient (MCC), the F1-score (F1) and the G-mean (G), which are overall performance measures that are more suitable to imbalanced class ratios. The MCC is a correlation coefficient between the manual and predicted binary segmentations, and it has been previously used for the evaluation of retinal vessel segmentation methods [11], [15], [17]. It returns a value between 1 and $+1$, with $+1$ indicating a perfect prediction, 0 no better than random, and 1 a total disagreement between prediction and ground truth. The F1-score is the harmonic mean of precision and recall, and it also has the property of better characterizing quality when data are imbalanced. It achieves its maximum value of 1 when the segmentation of the positive class is perfect, and its lowest value of 0 when the segmentation is completely wrong. Similarly, the G-mean is a metric that measures the balanced between Se and Sp by taking their geometric mean, returning a value between 0 and 1 [48]. Finally, receiver operating characteristic (ROC) curves were also generated from the unary potentials and the energy of the FC-CRF, and the area under each curve (AUC) was computed. The AUC on STARE was obtained by computing the ROC curve on each image and taking the average value.

C. Model selection

Parameters for computing the unary features were initially fixed to the values reported by the original references, which use the DRIVE training set [7], [42]. Thus, responses to the 2D Gabor wavelet were obtained at scales $a = 2, 3, 4, 5$, and responses to line detectors were analyzed from $l = 1$ to 15, with increments of $l_0 = 2$. As Zana and Klein used different data to estimate the length of the structuring element, we selected $l = 9$, which is consistent with the average calibre of DRIVE vessels as reported in [49]. The size of the windows used for preprocessing the image to compute this feature was fixed at 35 pixels, and the border of the FOV was expanded by 50 pixels. For datasets other than DRIVE, we made use of the compensation factor ρ described in Section II-D. In the case of the Nguyen line detector, the increment l_0 is also multiplied to reduce the dimensionality of the feature vector when evaluating on images with higher resolutions. The angles were set to the values reported in the original references [7], [22], [42].

To estimate the parameter C of the SOSVM we randomly divided each training set into two new subsets, *training** and *validation*, containing 70% and 30% of the *training* images,

[3]http://www.uhu.es/retinopathy/muestras2.php

[4]http://blogs.kingston.ac.uk/retinal/chasedb1/

[5]https://www5.cs.fau.de/research/data/fundus-images/

[6]Accuracies are included in the supplementary material for completeness.

TABLE I
EVOLUTION OF F1-SCORE DURING FORWARD FEATURE SELECTION ON DRIVE.

Features	Unary potentials			Pairwise potentials	
	Iter. 1	Iter. 2	Iter. 3	Iter. 1	Iter. 2
Line detector	0.6898	**0.7256**	-	0.7535	0.6985
2D Gabor wavelet	**0.6967**	-	-	0.7423	0.7437
Zana and Klein enhanced	0.6378	0.7043	0.7129	**0.7546**	-

respectively. We used *training** to train the model, and *validation* to estimate the best C value. A model selection phase was initially performed, in which the SOSVM was trained using different values of C. Performance of each trained model was evaluated on the validation set. Values of $C \in \{(10^i)/c\}$, with $i \in \{-2, ..., 0, ..., 3\}$ with c equal to the total number of FOV pixels in the training set were evaluated. The configuration was selected to maximize the average F1-score on the validation set. That learned configuration of the CRF model was then evaluated on the test set only once. This configuration prevent us from using any test data to adjust parameters, allowing us to obtain a non-biased estimate of the test error [50]. Cross-validation was used on the STARE dataset.

A similar approach was followed with the purpose of adjusting θ_p: for a fixed value of C, different θ_p values spanning from 1 to 15 were explored, and the one that maximized the F1-score on the validation set was chosen. This search, however, was performed only once, using the DRIVE validation set. At the time of evaluation on the remaining sets, the selected value $\theta_p = 5$ was multiplied by the compensation factor.

Forward feature selection using DRIVE's *training** and *validation* sets was followed to identify which feature combinations are more suitable for the unary and the pairwise potentials. Table I shows the progress in the mean F1-score obtained on the validation set for each configuration of features in each iteration. Numbers in bold indicate that the set of features was chosen in that iteration. We observed that the 2D Gabor wavelet contributes to detect thicker structures but with a large number of false positives. By adding responses to the Nguyen *et al.* line detector, false positives are reduced and narrower vessels are segmented. Pairwise potentials showed better performance when using the Zana and Klein vessel enhanced image in the kernel. This configuration was applied to all subsequent data sets.

A strategy to estimate the scale parameter of a radial basis function is to take the median of all pairwise distances of a random sample of pixels [51]. Since part of each pairwise kernel consists of a radial basis function (Eq. 6), this approach was applied to the estimation of $\theta_{(m)}$. This estimator is robust in that it has low variance when it is computed over different random samples [51]. However, small changes to $\theta_{(m)}$ can affect the results due to the exponentiation in the pairwise term and the number of interactions taking into account by the fully connected model. Therefore, we estimate $\theta_{(m)}$ as the median of medians obtained over 50 different random samples of pairs $(f_i^{(m)}, f_j^{(m)})$ extracted from the training set of each data set.

TABLE II
QUANTITATIVE EVALUATION OF THE RESULTS OBTAINED ON DRIVE, STARE, CHASEDB1 AND HRF, USING ONLY THE UNARY POTENTIALS (UP) OR THE FULLY CONNECTED CRF (FC-CRF).

Dataset	Method	Se	Sp	Pr	F1	G	MCC
DRIVE	UP	0.7079	**0.9802**	**0.8394**	0.7661	0.8324	0.7401
	FC-CRF	**0.7897**	0.9684	0.7854	**0.7857**	**0.8741**	**0.7556**
STARE	UP	**0.7692**	0.9675	0.7445	0.7517	0.8618	0.7252
	FC-CRF	0.7680	**0.9738**	**0.7740**	**0.7644**	**0.8628**	**0.7417**
CHASEDB1	UP	0.7110	0.9707	0.7386	0.7209	0.8304	0.6919
	FC-CRF	**0.7277**	**0.9712**	**0.7438**	**0.7332**	**0.8403**	**0.7046**
HRF	UP	0.7315	**0.9680**	**0.7012**	0.7127	0.8413	0.6851
	FC-CRF	**0.7874**	0.9584	0.6630	**0.7158**	**0.8686**	**0.6897**

IV. EXPERIMENTS AND RESULTS

In this section we present the results obtained in our experiments. The prototype of this method was implemented with MATLAB R2013a, using MEX functions to interface with C++ implementations of the LNB-CRF and the FC-CRF. In Section IV-A we summarize and analyze the results, while in Section IV-B we include a comparison with respect to other published methods.

A. Results

A quantitative evaluation of the results obtained on our experiments using only the unary potentials and using the FC-CRF is presented in Table II. The binary segmentations were obtained by minimizing the corresponding energies using the mean field approximation strategy proposed in [36]. Results obtained with the LNB-CRF are not included in the table as they are exactly the same than those obtained using only the unary potentials. This is due to the fact that, in this configuration, the SOSVM assigns an almost-zero value to $\mathbf{w}_p$, the parameter that weights the local pairwise potentials, and at the same time it does not modify the weights associated to the unary potentials with respect to the configuration achieved when only the unary potentials are considered. This is because the grid connectivity does not provide valuable information in the context of elongated structures such as retinal vessels. By contrast, SOSVM assigns a non-zero value to $\mathbf{w}_p$ when training the FC-CRF, and also modifies the weights associated to the unary features and the bias term, meaning that the pairwise potentials influence the other parameters and contribute substantially to the prediction function. To evaluate the statistical significance of such influence, a set of right-tailed Wilcoxon signed-ranks hypothesis tests were performed on the quality values obtained using only the unary potentials and using the FC-CRF. No hypothesis tests were performed on STARE results since segmentations on this set were obtained by leave-one-out cross validation. Taking into account that the parameter C was tuned according to a validation set subsampled from each training set, it is not possible to assume that all the results were achieved with the same configuration. In addition to the quantitative analysis, we also provide several segmentation examples to analyze qualitatively the changes introduced by the pairwise potentials.

In some of the data sets the FC-CRF contributes to a statistically significant improvement in the results when evaluating in terms of F1-score (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1:

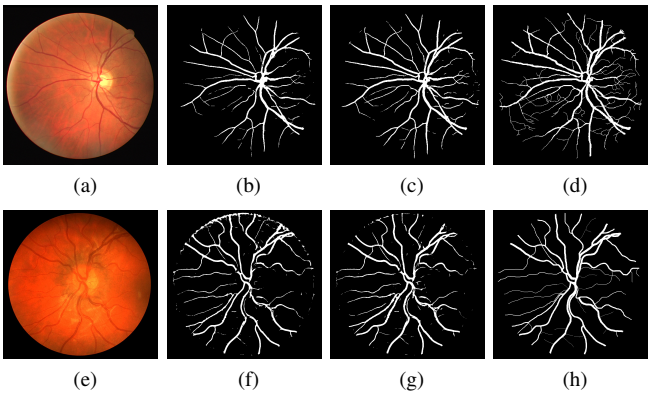


Fig. 2. Segmentation results obtained on DRIVE (first row) and CHASEDB1 (second row). First column: Image 04 of DRIVE (a) and Image_05L of CHASEDB1 (e). Second column: Segmentations obtained using only the unary potentials (b,f). Third column: Segmentations obtained using the FC-CRF model (c,g). Fourth column: ground truth labelling (d,h).

$p \approx 1 \times 10^{-3}$), G-mean (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1: $p \approx 8 \times 10^{-5}$; HRF: $p \approx 9 \times 10^{-7}$) and MCC (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1: $p \approx 4 \times 10^{-3}$; HRF: $p \approx 2 \times 10^{-2}$). F1-score and MCC are improved on average on STARE and HRF, as indicated in Table II. In the case of HRF, the improvement in the F1-score is lower than that obtained on other data sets. G-mean is also increased on average on STARE.

When evaluating on DRIVE, the pairwise potentials improve the Se value ($p \approx 4 \times 10^{-5}$) and slightly reduce the average Sp ($p \approx 4 \times 10^{-5}$). This is likely due to the FC-CRF introducing a certain fraction of false positives, as it can be observed in the reduction of the average Pr ($p \approx 4 \times 10^{-5}$). In all such cases, however, the fraction of improvement in Se is higher than the reduction on Sp and Pr, which is evidenced by the improvement in the G-mean, and also in the F1-score and the MCC values. Some examples of the segmentations obtained on DRIVE are shown in Fig. 2. It is possible to observe that the FC-CRF model incorporates a number of thinner vessels and significantly improves the connectivity of the vascular structure.

In STARE, the G-mean is slightly improved when using the dense approach. When decomposed into their terms, it is possible to see that the Se is slightly reduced in average, but with an improvement in both the average Sp and Pr, which is associated with a reduction in the number of false positive pixels. An example of this setting is observed in Fig. 3, which depicts an extreme pathological case. The FC-CRF contributed to reducing the number of false positives in the haemorrhage inside the optic disc. Yet the second human observer identified vessels within that region, the first human observer (which is assumed as the ground truth) did not mark anything there, which directly affects the Se value. Some narrow structures are integrated to the vascular tree by the FC-CRF model, and it can also be observed that the unary potentials overestimated the width of some of the major vessels.

In CHASEDB1, results are increased in terms of both Se ($p \approx 7 \times 10^{-5}$) and Sp, although the improvement in this last metric is not statistically significant. As a consequence, the G-mean is increased. The average Pr value is also improved by the FC-CRF, which is explained by the reduction in the number of false positives, as seen in Fig. 2. It is also possible to observe that in both test sets, the unary potentials overestimate the calibre of the narrower vessels, a setting that is improved when incorporating the pairwise potentials.

A different behavior can be observed on HRF (Fig. 4), where Se is significantly increased ($p \approx 9 \times 10^{-7}$) but Sp ($p \approx 9 \times 10^{-7}$) is diminished. Pr is also decreased ($p \approx 9 \times 10^{-7}$), meaning that a number of false positives in introduced. We observed qualitatively, however, that the FC-CRF detects a large number of narrow vessels that are ignored when using only the unary potentials, as can be seen in Fig. 5. It is possible to see also that some of the thinner capillaries remain ignored. Additional work on feature construction might help to incorporate such structures. It can be seen that the pairwise potentials contribute to joining isolated detections, resulting in a more general connected tree, and in an increase in the G-mean value.

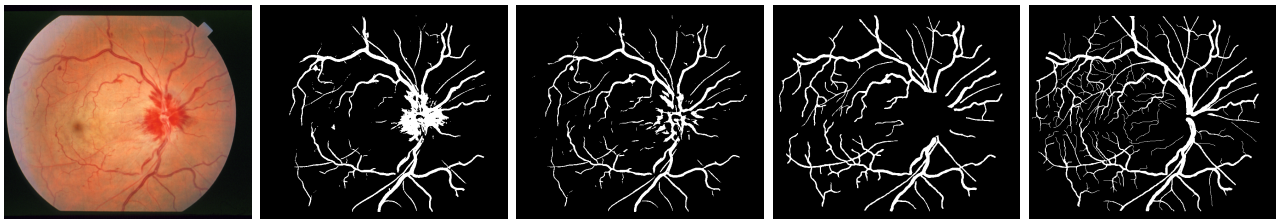
A comparison between the ROC curves obtained using only the unary potentials and using the FC-CRF can be observed in Fig. 6, where the second human observer performance on each set (if available) is plotted. The curve on STARE-C is not included since results there were obtained by cross-validation, and the segmentation of each image was made with a different model. Results on DRIVE show that the FC-CRF outperforms the unary potentials, and also that they are quantitatively tied to the second observer performance. When evaluating on HRF, however, and in line with the analysis previously performed, the unary potentials results in a better ROC curve than the FC-CRF. The areas under each of the curves are in line with these conclusions.

Finally, the computation cost of our single-thread, single-core implementation of the inference on the FC-CRF model was evaluated on an Intel(R) Xeon(R) CPU E5-2690 0 platform at 2.90GHz with 64 GB of RAM. For this purpose, the average time of applying the FC-CRF on each test set was measured. As seen in Fig 7, although the computational cost grows with the resolution of the images, it is still fast enough to be feasibly applied in a clinical setting.

B. Comparison with other methods

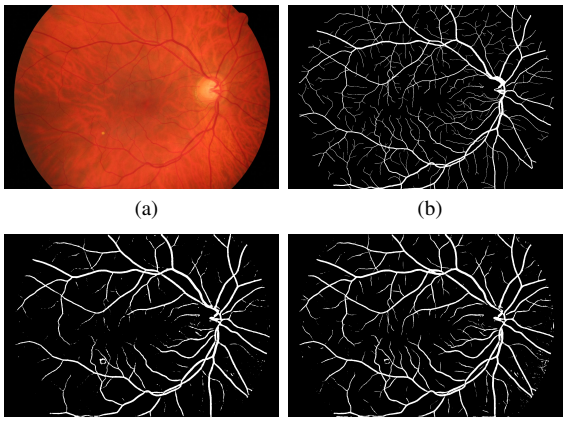
We also include a comparison of our results with respect to those reported by other state-of-the-art methods evaluated on DRIVE and STARE (Table III), CHASEDB1 and HRF (Table IV). Although our method is supervised, we also compare with unsupervised approaches in the tables. Methods that obtained the final binary segmentations using parameters that were estimated on the test data were not included on the tables of DRIVE and STARE, since that setting might underestimate the actual test error [52, Section 6].[7] Despite

[7]In [15], the results reported on DRIVE test set, STARE and CHASEDB1 correspond to the binary segmentations obtained by thresholding the responses to the B-COSFIRE filter. However, the threshold is selected by maximizing the average MCC on each corresponding test set. Similarly, in [10], [12], [13], [53] the scores provided by different classifiers are thresholded using the parameter that maximizes the average Acc on the test data (DRIVE test set, STARE and CHASEDB1 in the case of [12], [13], [53], and DRIVE test set and STARE in [10]).



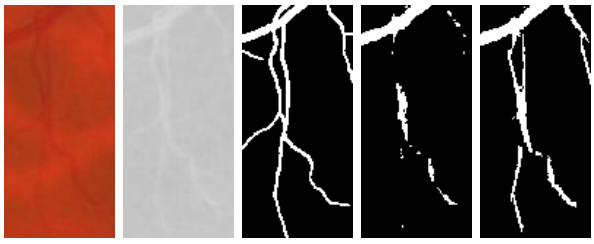
(a) (b) (c) (d) (e)

Fig. 3. Segmentation results obtained on a serious pathological case on STARE. (a) Image im0005. (b) Segmentation obtained using only the unary potentials. (c) Segmentations obtained using the FC-CRF model. (d) First human observer annotations. (e) Second human observer annotations.



(a) (b) (c) (d)

Fig. 4. Segmentation results obtained on HRF test set. (a) Image 11_g. (b) Manual annotation. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF.



(a) (b) (c) (d) (e)

Fig. 5. Example of narrow vessel detection under low contrast conditions. (a) Detail of Image 11_g. (b) Preprocessed image. (c) Manual annotation. (d) Segmentation obtained using only the unary potentials. (e) Segmentation obtained using the FC-CRF.

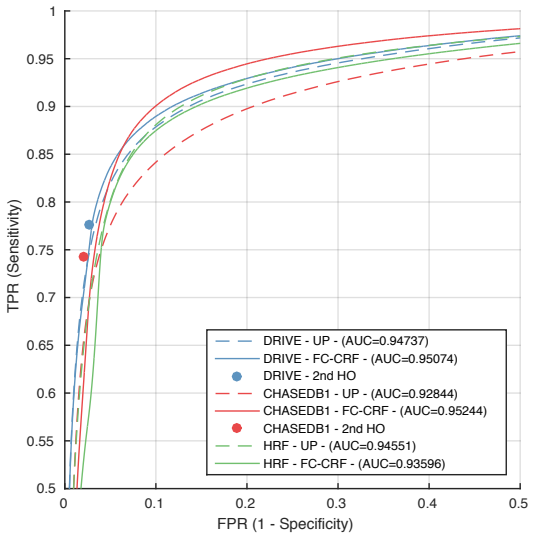


Fig. 6. ROC curves on DRIVE, CHASEDB1 and HRF, using only the unary potentials (UP, slashed line) or the FC-CRF (solid line), and second human observer (HO) performance.

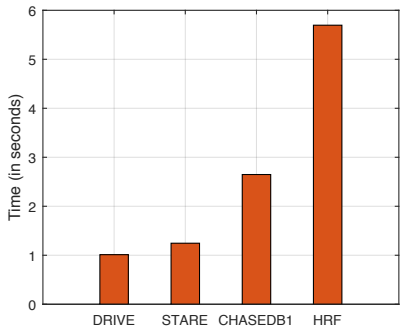


Fig. 7. Computational cost of the FC-CRF inference in all the data sets used for evaluation.

that those other papers have evaluated a different setting, we included their results for completeness in the supplementary materials. Results obtained similarly but on CHASEDB1 were included in Table IV as they are the only works evaluated on those sets. However, they are marked with an asterisk.

Our method reports the highest F1-score and MCC on DRIVE when compared with other supervised and unsupervised strategies. Se is also higher, for a relatively acceptable Sp value. As mentioned before, the Sp measure describes the capability of the method to distinguish the non-vessel class, and it usually suffers when the segmentations have a high number of false positives. However, the Pr value is higher than the method by Fathi *et al.* [26], which has reported a higher Sp value but a lower average Se.

Comparison on STARE is difficult as most of the state-of-the-art methods performed their analysis using their own strategies to train and test.[8] It must be taken into account, then, that the supervised methods listed in the comparison illustrated

[8]Some methods were trained on the first half and tested on the second half of STARE [54] or even on the entire set [15], or were trained on a random sample of pixels extracted from STARE [7], etc.

TABLE III
COMPARISON OF AVERAGE Se, Sp, Pr, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS AND THE 2ND HUMAN OBSERVER, WHEN EVALUATING ON DRIVE AND STARE.

	DRIVE						**STARE**					
Methods	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	**0.7897**	0.9684	0.7854	**0.7857**	**0.8741**	**0.7556**	0.7680	0.9738	**0.7740**	**0.7644**	0.8628	**0.7417**
2nd human observer	0.7760	0.9730	0.8066	0.7881	0.8689	0.7601	0.8951	0.9387	0.6424	0.7401	0.9166	0.7225
Supervised												
Dai *et al.* [5]	0.7359	0.9720	-	-	0.8458	-	0.7769	0.9550	-	-	0.8614	-
Niemeijer *et al.* [6]	0.6793	0.9725	-	-	0.8128	-	-	-	-	-	-	-
Lupascu *et al.* [14]	0.6728	**0.9874**	-	-	0.8151	-	-	-	-	-	-	-
Orlando and Blaschko [31]	0.7850	0.9670	0.7770	0.7810	0.8713	0.7482	-	-	-	-	-	-
Soares *et al.* [7]	0.7283	0.9788	-	-	0.8443	-	0.7200	0.9750	-	-	0.8379	-
Xu and Luo [8]	0.7760	-	-	-	-	-	-	-	-	-	-	-
You *et al.* [9]	0.7410	0.9751	-	-	0.8500	-	0.7260	0.9751	-	-	0.8414	-
Vega *et al.* [11]	0.7444	0.9600	-	0.6884	0.8454	0.6617	0.7019	0.9671	-	0.6082	0.8239	0.5927
Unsupervised												
Al-Diri *et al.* [20]	0.7282	0.9551	-	-	0.8340	-	0.7521	0.9681	-	-	0.8533	-
Chakraborti *et al.* [25]	0.7205	0.9579	-	-	0.8308	-	0.6786	0.9586	-	-	0.8065	-
Fathi and Naghsh-Nilchi [26]	0.7768	0.9759	0.7559	0.7669	0.8707	-	**0.8061**	0.9717	0.7027	0.7509	**0.8850**	-
Fraz *et al.* [17]	0.7152	0.9768	**0.8205**	0.7642	0.8358	0.7333	0.7409	0.9665	0.7363	0.7386	0.8462	0.7003
Fraz *et al.* [18]	0.7302	0.9742	0.8112	0.7686	0.8434	0.7359	0.7318	0.9660	0.7294	0.7306	0.8408	0.6908
Odstrcilik *et al.* [24]	0.7060	0.9693	-	-	0.8272	-	0.7847	0.9512	-	-	0.8639	-
Roychowdhury *et al.* [19]	0.7390	0.9780	-	-	0.8501	-	0.7320	0.9840	-	-	0.8487	-
Yin *et al.* [16]	0.6522	0.9710	-	-	0.7958	-	0.7248	0.9666	-	-	0.8370	-
Zhao *et al.* [21]	0.7420	0.9820	-	-	0.8536	-	0.7800	**0.9780**	-	-	0.8734	-

TABLE IV
COMPARISON OF AVERAGE Se, Sp, Pr, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS AND THE 2ND HUMAN OBSERVER, WHEN EVALUATING ON THE CHASEDB1 AND HRF.

CHASEDB1	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	0.7277	0.9712	0.7438	0.7332	0.8407	0.7046
2nd human observer	0.7425	0.9793	0.8090	0.7686	0.8527	0.7475
Fraz *et al.* [13]*	0.7259	0.9770	0.7732	0.7488	0.8421	-
HRF	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	**0.7874**	0.9584	0.6630	0.7158	**0.8687**	0.6897
Odstrcilik *et al.* [24]	0.7794	**0.9650**	**0.6950**	**0.7324**	0.8672	**0.7065**

in Table III were trained miscellaneously. When evaluated on this set, our method reports the highest average F1-score and MCC values, indicating a better overall performance. Additionally, the FC-CRF outperforms the other strategies in terms of the Pr measure, meaning that the number of false positive detections is lower than in the other cases.

When evaluating on CHASEDB1 it is possible to see that our method achieved a better Se value than the other strategies. The F1-score value is outperformed, yet the results reported in the reference were obtained using parameters estimated from the test labels, and the parameters of the features it used were adjusted to this specific data set. In contrast, our FC-CRF model was trained using feature parameters that were scaled using the compensation factor.

To the best of our knowledge, only unsupervised methods were previously tested on HRF [24], [55]. We include the results of [24] on the test set calculated from the binary segmentations provided by the corresponding authors. In general, it is possible to see that the FC-CRF gives higher Se values than the method proposed in [24], but with lower Sp and Pr values. This means that the FC-CRF obtains a larger number of false positive detections than the other strategy. However, global metrics including F1-score and MCC are competitive.

V. DISCUSSION

The FC-CRF model and the learning strategy better exploited the interaction between pixels features than the local neighborhood based approach. The local neighborhood approach was not able to improve results with respect to the unary potentials, as a zero weight is assigned to the pairwise term by the structured output SVM. The hypothesis tests performed on the results obtained by both the FC-CRF and the unary potentials on a number of different data sets, as explained in Section IV-A, demonstrated that the dense pairwise potentials introduced statistically significant improvements in several metrics. Additionally, as evidenced in ROC curves in Fig 6, the FC-CRF model also yields results that are tied to the second human observer performance. Such properties are due to the contribution of the high order pairwise potentials, which are able to better reconstruct the vessels even under low contrast conditions (Fig. 5) with a negligible time overhead (Fig 7). Although no directional prior is explicitly learned by the model, the combination of the distance and the feature dissimilarity terms within the pairwise kernel (Eq. (8)) provides a way to penalize too long or dissimilar interactions, respectively. Thus, if the pairwise features are robust enough, then the model will assign low energies to the labellings of filamentary structures, and will penalize other non elongated shapes. The features analyzed in this work consistently achieved better results than using only unary potentials, as observed in Table I. By contrast, the LNB-CRF model is not able to take advantage of the pairwise features, as evidenced by the absense of improvement with respect to the unary potentials. Based on this property, it is possible to conclude that dense potentials are able to better characterize the vessels. Other medical and biological applications might

benefit by using this approach for segmenting other tubular and elongated structures such as vessels, neurons or plant roots.

Extensive comparison with state of the art methods has also shown that our approach consistently performed well on several metrics, and is a fully-automated segmentation algorithm that achieves better results when evaluated in terms of global binary classification measures such as F1-score, G-mean and MCC. This is in part due to previous studies focusing on raw pixel accuracy, which ignores the fact that the number of pixels occupied by blood vessels is a relatively small fraction of the image. As a result, competing methods suffer as measured by F1-score, G-mean and MCC, which are particularly important as they reflect an accurate estimation of the vessel pixels, the primary goal in vessel segmentation for fundus image analysis.

As in the case of other supervised techniques–such as Gaussian mixture models [7] or SVMs [8]–the performance of our method is affected by the general ability of the features to characterize the retinal vasculature. State of the art features were used in our experiments in order to evaluate the contribution of the fully connected model in the improvement of the original results. Most of the features presented in the literature were designed using low resolution images such as those in DRIVE and STARE. Since the design of features involves the adjustment of different parameters that are effected by image scale, a decrease in performance is expected when the resolution of the images differs from the original setting. An alternative to reduce this effect would be to design the features for each specific resolution, though this is particularly time consuming. Instead, we proposed to use a simple technique based on applying a compensation factor ρ that is multiplied to each scale parameter before feature extraction. Using this basic approach our method is able to partially compensate for changes in resolution, outperforming other well-known segmentation strategies.

Although our approach achieves overall good performance, we have observed some misclassifications in the bright central reflex of the major arteries in the high resolution images in CHASEDB1 and HRF. This is likely due to the limited capability of both the unary and the pairwise features to deal with this property, as it was not taken into account when they were originally designed. Additional feature development or learning would be valuable in order to improve performance under this or other challenging contexts, such as in the presence of serious pathological changes. Currently, most features work well on lower resolution images, as fewer features have been developed for high resolution images such as HRF. It is a promising avenue of research to consider image structures that become apparent at higher resolutions, such as the central reflex in arteries. We encourage further research in this direction. Furthermore, evaluating other types of higher order potentials, combined with this learning approach, could potentially improve the results by capturing other types of pixel interactions. However, it must be taken into account that those approaches always involve a trade-off between performance and computational tractability.

VI. Conclusions

In this work, we have presented a detailed description and evaluation of our discriminatively trained segmentation model based on a fully connected CRF for the purpose of blood vessel segmentation in fundus images. By means of features extracted from the images and fully connected pairwise potentials, this approach is able to reconstruct the retinal vasculature much more precisely than using only the unary potentials or a local neighborhood based CRF. The effectiveness of the approach is evidenced by the general improvement in the values of F1-score, G-mean and Matthews correlation coefficient–three quantitative measures that are suitable in binary classification problems where the number of true positive and true negative pixels are unbalanced–obtained on a number of benchmark data sets. ROC curves also show that results achieved with the FC-CRF are comparable with those obtained by a second human observer. The capability of the dense potentials to reconstruct elongated structures can potentially benefit other biological and medical applications. Segmentation masks and additional implementation details are available at http://homes.esat.kuleuven.be/~mblaschk/projects/retina/.

Acknowledgments

This work is partially funded by Internal Funds KU Leuven, ERC Grant 259112, FP7-MC-CIG 334380, ANPCyT PICT 2010-1287 and ANPCyT PICT 2014-1730. J.I.O. is funded by a doctoral scholarship granted by CONICET. We thank authors of [7], [36], [40], [42] for providing us with their code, and [15], [24] for providing us with their segmentations.

References

[1] E. Prokofyeva and E. Zrenner, "Epidemiology of major eye diseases leading to blindness in Europe: A literature review," *Ophthalmic Research*, vol. 47, no. 4, pp. 171–188, 2012.

[2] M. M. Fraz *et al.*, "Blood vessel segmentation methodologies in retinal images–a survey," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 407–433, 2012.

[3] M. D. Abràmoff *et al.*, "Retinal imaging and image analysis," *Biomedical Engineering, IEEE Reviews in*, vol. 3, pp. 169–208, 2010.

[4] M. Esmaeili *et al.*, "A new curvelet transform based method for extraction of red lesions in digital color retinal images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 4093–4096.

[5] P. Dai *et al.*, "A new approach to segment both main and peripheral retinal vessels based on gray-voting and gaussian mixture model," *PloS one*, vol. 10, no. 6, p. e0127748, 2015.

[6] M. Niemeijer *et al.*, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 648–656.

[7] J. V. Soares *et al.*, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *Medical Imaging, IEEE Transactions on*, vol. 25, no. 9, 2006.

[8] L. Xu and S. Luo, "A novel method for blood vessel detection from retinal images," *Biomedical engineering online*, vol. 9, no. 1, p. 14, 2010.

[9] X. You *et al.*, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognition*, vol. 44, no. 10, 2011.

[10] D. Marín *et al.*, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 1, pp. 146–158, 2011.

[11] R. Vega *et al.*, "Retinal vessel extraction using lattice neural networks with dendritic processing," *Computers in biology and medicine*, vol. 58, pp. 20–30, 2015.

[12] M. M. Fraz *et al.*, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 9, pp. 2538–2548, 2012.

[13] M. M. Fraz *et al.*, "Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification," *International journal of computer assisted radiology and surgery*, vol. 9, no. 5, pp. 795–811, 2014.

[14] C. A. Lupascu *et al.*, "Fabc: retinal vessel segmentation using adaboost," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 5, pp. 1267–1274, 2010.

[15] G. Azzopardi *et al.*, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical image analysis*, vol. 19, no. 1, pp. 46–57, 2015.

[16] Y. Yin *et al.*, "Automatic segmentation and measurement of vasculature in retinal fundus images using probabilistic formulation," *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[17] M. M. Fraz *et al.*, "Retinal vessel extraction using first-order derivative of gaussian and morphological processing," in *Advances in Visual Computing*. Springer, 2011, pp. 410–420.

[18] M. M. Fraz *et al.*, "Application of morphological bit planes in retinal blood vessel extraction," *Journal of digital imaging*, vol. 26, no. 2, pp. 274–286, 2013.

[19] S. Roychowdhury *et al.*, "Iterative vessel segmentation of fundus images," *Biomedical Engineering, IEEE Transactions on*, vol. 62, no. 7, pp. 1738–1749, 2015.

[20] B. Al-Diri *et al.*, "An active contour model for segmenting and measuring retinal vessels," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 9, pp. 1488–1497, 2009.

[21] Y. Zhao *et al.*, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retina images," *Medical Imaging, IEEE Transactions on*, 2015.

[22] F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *Image Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 1010–1019, 2001.

[23] G. B. Kande *et al.*, "Unsupervised fuzzy based vessel segmentation in pathological digital fundus images," *Journal of medical systems*, vol. 34, no. 5, pp. 849–858, 2010.

[24] J. Odstrčilík *et al.*, "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database," *IET Image Processing*, vol. 7, no. 4, pp. 373–383, 2013.

[25] T. Chakraborti *et al.*, "A self-adaptive matched filter for retinal blood vessel detection," *Machine Vision and Applications*, vol. 26, no. 1, pp. 55–68, 2014.

[26] A. Fathi and A. R. Naghsh-Nilchi, "Automatic wavelet-based retinal blood vessels segmentation and vessel diameter estimation," *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 71–80, 2013.

[27] A. Soltanipour *et al.*, "Vessel centerlines extraction from fundus fluorescein angiogram based on hessian analysis of directional curvelet subbands," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1070–1074.

[28] X. He *et al.*, "Multiscale conditional random fields for image labeling," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–695.

[29] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vision*, vol. 68, no. 2, pp. 179–201, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1007/s11263-006-7007-9

[30] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, 3rd ed. Springer, 2009.

[31] J. I. Orlando and M. Blaschko, "Learning fully-connected CRFs for blood vessel segmentation in retinal images," in *MICCAI 2014, LNCS*, P. Golland, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer, 2014, vol. 8149, pp. 634–641.

[32] S. Kadoury *et al.*, "Higher-order crf tumor segmentation with discriminant manifold potentials," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 719–726.

[33] S. Kadoury *et al.*, "Metastatic liver tumour segmentation from discriminant grassmannian manifolds," *Physics in Medicine and Biology*, vol. 60, no. 16, p. 6459, 2015.

[34] F. Fiorani and U. Schurr, "Future scenarios for plant phenotyping," *Annual review of plant biology*, vol. 64, pp. 267–291, 2013.

[35] M. Helmstaedter, "Cellular-resolution connectomics: challenges of dense neural circuit reconstruction," *Nature methods*, vol. 10, no. 6, pp. 501–507, 2013.

[36] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2012, pp. 109–117.

[37] I. Tsochantaridis *et al.*, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005, pp. 1453–1484.

[38] T. Joachims *et al.*, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[39] J. D. Lafferty *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[40] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[41] N. Komodakis *et al.*, "MRF energy minimization and beyond via dual decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 531–552, 2011.

[42] U. T. Nguyen *et al.*, "An effective retinal blood vessel segmentation method using multi-scale line detection," *Pattern Recognition*, vol. 46, no. 3, pp. 703–715, 2013.

[43] A. M. Mendonca and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *Medical Imaging, IEEE Transactions on*, vol. 25, no. 9, 2006.

[44] J. Staal *et al.*, "Ridge based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 4, pp. 501–509, 2004.

[45] A. Perez-Rovira *et al.*, "Improving vessel segmentation in ultra-wide field-of-view retinal fluorescein angiograms," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 2614–2617.

[46] A. Hoover *et al.*, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *Medical Imaging, IEEE Transactions on*, vol. 19, no. 3, pp. 203–210, 2000.

[47] J. Odstrčilík *et al.*, "Improvement of vessel segmentation by matched filtering in colour retinal images," in *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*. Springer, 2009, pp. 327–330.

[48] H. He, E. Garcia *et al.*, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[49] M. Al-Rawi *et al.*, "An improved matched filter for blood vessel detection of digital retinal images," *Computers in Biology and Medicine*, vol. 37, no. 2, pp. 262–267, 2007.

[50] T. Hastie *et al.*, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[51] B. Schölkopf, "Support vector learning," Ph.D. dissertation, Oldenbourg Verlag, Munich, 1997.

[52] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[53] S. Roychowdhury *et al.*, "Blood vessel segmentation of fundus images by major vessel extraction and sub-image classification," *Biomedical and Health Informatics, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2014.

[54] M. A. Amin and H. Yan, "High speed detection of retinal blood vessels in fundus image using phase congruency," *Soft Computing*, vol. 15, no. 6, pp. 1217–1230, 2011.

[55] A. Budai *et al.*, "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, 2013.

Supplementary Material: A Discriminatively Trained Fully Connected Conditional Random Field Model for Blood Vessel Segmentation in Fundus Images

José Ignacio Orlando*, Elena Prokofyeva, and Matthew B. Blaschko

I. Introduction

In this supplementary material we include a more detailed comparison of our method with respect to other state of the art strategies when evaluating on DRIVE, STARE, CHASEDB1 and HRF. This information was not included in the main article due to space limitations. Though the tables do not pretend to be exhaustive, we try to cover most of the more recently published works in the field. All the comparison tables include the values of sensitivity-recall (Se-Re), specificity (Sp), precision (Pr), accuracy (Acc), area under the ROC curve (AUC), G-mean, F1-score and Matthews correlation coefficient (MCC). Works that estimated parameters such as thresholds using test data were identified with an asterisk, since the results of such approaches could potentially underestimate the actual test error and overestimate each metric in the tables. We also included a comparison between the computational time of the LNB-CRF and the FC-CRF model.

II. Evaluation on DRIVE

Table I presents a comparison with respect to other state of the art strategies evaluated on DRIVE. The Se achieved by our method is maximal, and it is possible to see other works reporting lower Sp values. At the same time, our method achieves the highest G-mean, even compared with respect to the second human observer. This represents that the FC-CRF model achieved a good balance between Se and Sp. Our F1-score and MCC values are also the highest, and the AUC and Acc values are in line with those achieved by other strategies. Additionally, it worth mentioning that our method obtained similar quality measures as those obtained by the second human observer.

III. Evaluation on STARE

As mentioned in the main article, the STARE data set comprises 20 images (10 of them containing pathologies) that are not divided into training and test. Since no standard strategy was followed by the reviewed literature at the time of the evaluation, we have reproduced some other criteria previously applied to split the images into training and test sets in order to provide additional comparisons to the existing literature.

The split followed by Amin *et al.* [15], which we refer to as STARE-A, uses the first half of STARE for training and the second half for testing. Similarly, Azzopardi *et al.* [1] trained their method on the first half of STARE and evaluated on the entire set. This last approach can lead to biased results, since images used for training are also included on the test set. We decided to run experiments on STARE-A, and we also included the results obtained by Azzopardi *et al.* on the second half of the original data set.

It must be mentioned that most of the images with serious pathologies appear in the training set, but not the test set of STARE-A. To ensure a proper evaluation of their method, Salem *et al.* [34] divided the original STARE set into training and test randomly, but ensuring that the same number of pathological and healthy images are distributed in each set. However, they did not report the names of the images used on each subset, so it is not possible to reproduce exactly the same partition. We decided to follow a similar approach, introducing an additional split, named STARE-B, obtained using the following procedure. First, images were divided into four groups: normal images (10), images with bright lesions (5), images with red lesions (3) and images with small lesions (2). Afterwards, half of each subset was assigned randomly to the training or the test sets. The configurations of STARE-A and STARE-B are listed in Table II for reproducibility. Since no other article has reported results using this split, we computed the results obtained by Marin *et al.* [6] on these images from the quantities presented in their paper.

Table III presents a detailed comparison of the results for each of the configurations on STARE. In each plit, the FC-CRF contributes to a statistically significant improvement in the results when evaluating in terms of F1-score (STARE-A: $p \approx 4\times10^{-5}$; STARE-B: $p \approx 2\times10^{-3}$), G-mean (STARE-A: $p \approx 9\times10^{-4}$; STARE-B: $p \approx 9\times10^{-4}$) and MCC (STARE-A: $p \approx 4 \times 10^{-5}$; STARE-B: $p \approx 2 \times 10^{-3}$).

In STARE-A, the pairwise potentials improve the Se value ($p \approx 9 \times 10^{-4}$) and slightly reduce the average Sp (STARE-A: $p \approx 9 \times 10^{-4}$). This can be explained by the FC-CRF introducing some false positives, as observed in the reduction of the average Pr ($p \approx 9\times10^{-4}$). The fraction of improvement in the Se value, however, is higher than the reduction on Sp

*José Ignacio Orlando is with National Council for Scientific and Technological Research (CONICET) and Pladema Institute, Tandil, Buenos Aires, 7000 Argentina. E-mail: jiorlando@conicet.gov.ar.

Elena Prokofyeva is with Inserm, U1018, University of Versailles Saint-Quentin, Villejuif, France, and Northern State Medical University, Troitsky av. 51, Arkhangelsk, Russia. E-mail: elena.prokofyeva@inserm.fr.

Matthew Blaschko is with ESAT-PSI-Visics, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. E-mail: Matthew.Blaschko@kuleuven.be. Work was done in part at Inria Scalay & CentraleSupélec, Grande Voie des Vignes, 92295 Châtenay-Malabry, France.

Manuscript received Month Day, Year; revised Month Day, Year.

TABLE I
COMPARISON OF AVERAGE Se, Sp, AUC, Acc, Pr, Re, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS AND THE 2ND HUMAN OBSERVER, WHEN EVALUATING ON THE DRIVE TEST SET.

Method	Se	Sp	AUC	Acc	Pr	Re	F1-score	G-mean	MCC
FC-CRF	**0.7897**	0.9684	0.9506	0.9454	0.7854	**0.7897**	**0.7857**	**0.8741**	**0.7556**
Unary potentials	0.7079	0.9802	0.9474	0.9453	**0.8394**	0.7079	0.7661	0.8324	0.7401
2nd human observer	0.7760	0.9730	-	0.9473	0.8066	0.7760	0.7881	0.8680	0.7601
Supervised									
Azzopardi *et al.* [1]*	0.7655	0.9704	0.9614	0.9442	-	0.7655	-	0.8605	0.7475
Cheng *et al.* [**?**]	0.7252	0.9798	0.9648	0.9474	-	0.7252	-	0.8429	-
Dai *et al.* [2]	0.7359	0.9720	-	0.9418	-	-	-	0.8458	-
Fraz *et al.* [3]*	0.7406	0.9807	0.9747	0.9480	-	0.7406	-	0.8522	-
Niemeijer *et al.* [4]	0.6793	0.9725	0.9294	0.9416	-	0.6793	-	0.8128	-
Lupascu *et al.* [5]	0.6728	**0.9874**	0.9561	**0.9597**	-	-	-	0.8151	-
Marin *et al.* [6]*	0.7067	0.9801	0.9588	0.9452	0.8433	0.7067	0.7690	0.8322	-
Orlando and Blaschko [7]	0.7850	0.9670	-	0.9437	0.7770	0.7850	0.7810	0.8713	0.7482
Roychowdhury *et al.* [8]*	0.7249	0.9830	0.9620	0.9519	-	0.7249	-	0.8441	-
Staal *et al.* [9]	-	-	0.9520	0.9441	-	-	-	-	-
Soares *et al.* [10]	0.7283	0.9788	**0.9614**	0.9466	-	0.7283	-	0.8443	
Xu and Luo [11]	0.7760	-	-	0.9328	-	-	-	-	
You *et al.* [12]	0.7410	0.9751	-	0.9434	-	0.7410	-	0.8500	
Vega *et al.* [13]	0.7444	0.9600	-	0.9414	-	-	0.6884	0.8454	0.6617
Unsupervised									
Al-Diri *et al.* [14]	0.7282	0.9551	-	-	-	0.7282	-	0.8340	-
Amin *et al.* [15]	0.6608	0.9799	0.9360	0.9191	-	0.6608	-	0.8047	-
Bankhead *et al.* [16]	0.7027	0.9717	-	0.9371	-	-	-	0.8263	-
Budai *et al.* [17]	0.6440	0.9870	-	0.9572	-	-	-	0.7973	-
Chakraborti *et al.* [18]	0.7205	0.9579	0.9419	0.9370	-	0.7205	-	0.8308	-
Espona *et al.* [19]	0.6634	0.9682	-	0.9316	-	0.6634	-	0.8014	-
Espona *et al.* [19]	0.7436	0.9615	-	0.9352	-	0.7436	-	0.8456	-
Fathi and Naghsh-Nilchi [20]	0.7768	0.9759	0.9516	0.9581	0.7559	0.7768	0.7669	0.8707	-
Fraz *et al.* [21]	0.7152	0.9768	-	0.9430	0.8205	0.7152	0.7642	0.8358	0.7333
Fraz *et al.* [22]	0.7302	0.9742	-	0.9422	0.8112	0.7302	0.7686	0.8434	0.7359
Kande *et al.* [23]	-	-	0.9518	0.8911	-	-	-	-	-
Lam *et al.* [24]	-	-	0.9614	0.9472	-	-	-	-	-
Martínez *et al.* [25]	0.7246	0.9655	-	0.9344	-	0.7246	-	0.8364	-
Miri *et al.* [26]	0.7352	0.9795	-	0.9458	-	0.7352	-	0.8486	-
Palomera *et al.* [27]	0.6440	0.9670	-	0.9250	-	0.6440	-	0.7891	-
Odstrcilik *et al.* [28]	0.7060	0.9693	0.9519	0.9340	-	0.7060	-	0.8272	-
Roychowdhury *et al.* [29]	0.7390	0.9780	0.9670	0.9490	-	0.7390	-	0.8501	-
Saffarzadeh *et al.* [**?**]	-	-	0.9303	0.9387	-	-	-	-	-
Vlachos and Dermatas [30]	0.7468	0.9551	-	0.9285	-	0.7468	-	0.8446	-
Wang *et al.* [31]	0.7520	0.9800	-	-	-	0.7520	-	0.8585	-
Zhang *et al.* [32]	0.7120	0.9724	-	0.7120	-	0.7120	-	0.8321	-
Zhao *et al.* [33]	0.7420	0.9820	-	0.9540	-	0.7420	-	0.8536	-

and Pr, which is evidenced by an increase in the G-mean value, and both F1-score and MCC. Some examples of the segmentations obtained on STARE-A are shown in Fig. 1. It is possible to observe that the FC-CRF model incorporates a number of thinner vessels and significantly improves the connectivity of the vascular structure. When comparing with respect to Amin *et al.* [15] and Azzopardi *et al.* [1] it is possible to see that our method outperforms the other two strategies on a majority of the quality measures.

A different behavior can be observed on STARE-B, where Se is decreased ($p \approx 9 \times 10^{-4}$) but the Sp is increased ($p \approx 9 \times 10^{-4}$). G-mean, however, is improved when using the FC-CRF. A higher Pr value ($p \approx 9 \times 10^{-4}$) is obtained when pairwise potentials are used, meaning that the number of false positives was reduced. This situation can be observed in the example depicted in Fig. 2, where the FC-CRF has reduced the number of false positive pixels detected on the red lesions, and also improved the detection within the optic disc area. When compared with respect to the other method, it is possible to see that in general our FC-CRF model reports higher quality measures, even though the binary segmentations of Marin *et al.* [6] were obtained using a threshold that was estimated using the test data.

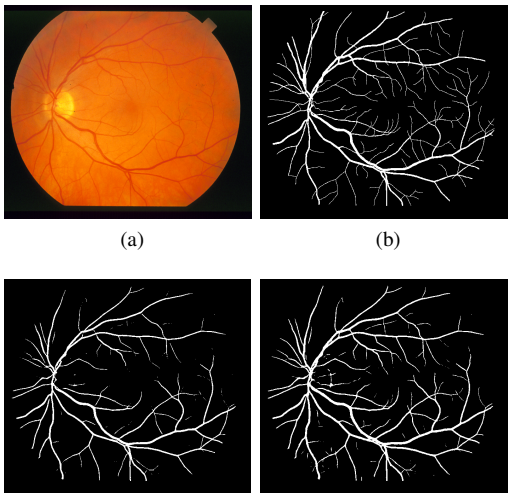
The ROC curves obtained on STARE-A and STARE-B are depicted in Fig. 3. On STARE-A the curve obtained using the FC-CRF energy surpasses the one obtained using only the unary potentials. When evaluating on STARE-B, the FC-CRF potentials are able to obtain better Se values with higher Sp than those obtained using only the unary energy. For Sp values around 0.92, the unary potentials achieve better Se values, though such Sp is not acceptable as such a threshold results in a large number of false positives.

IV. EVALUATION ON CHASEDB1

CHASEDB1 was divided into training and test by the creators of the dataset, with 8 and 20 images in each set, respectively [3]. Azzopardi *et al.* [1], however, followed a different approach, training on the first half of the set and

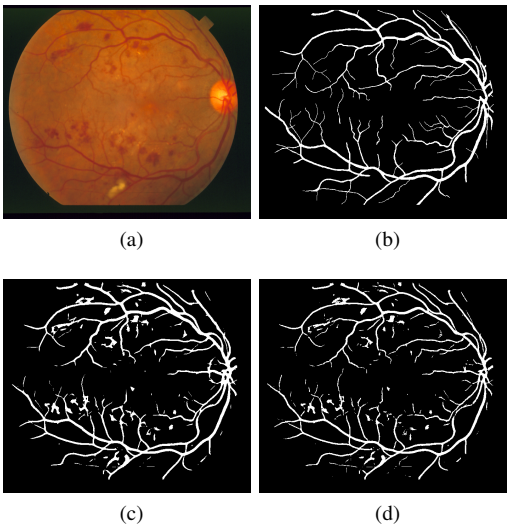
TABLE II
TRAINING AND TEST SETS CONFIGURATIONS ON STARE. IMAGES LABELLED AS *Training* BELONG TO THE TRAINING SET, AND *Test* TO THE TEST SET.

Image ID	Type	STARE-A	STARE-B
im0001	Bright lesions	Training	Training
im0002	Bright lesions	Training	Test
im0003	Bright lesions	Training	Test
im0004	Healthy	Training	Training
im0005	Red lesions	Training	Training
im0044	Bright lesions	Training	Training
im0077	Healthy	Training	Test
im0081	Healthy	Training	Test
im0082	Healthy	Training	Test
im0139	Red lesions	Training	Test
im0162	Healthy	Test	Test
im0163	Healthy	Test	Training
im0235	Small lesions	Test	Test
im0236	Small lesions	Test	Training
im0239	Healthy	Test	Test
im0240	Healthy	Test	Training
im0255	Healthy	Test	Training
im0291	Healthy	Test	Training
im0319	Bright lesions	Test	Test
im0324	Red lesions	Test	Training



(a) (b) (c) (d)

Fig. 1. Segmentation results obtained on STARE-A. (a) Image im0162. (b) Manual annotations. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF model.



(a) (b) (c) (d)

Fig. 2. Segmentation results obtained on a pathological image on STARE-B test set. (a) Image im0139. (b) Manual annotation. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF.

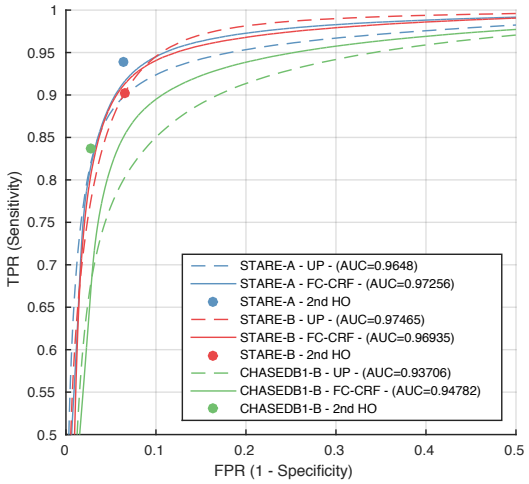


Fig. 3. ROC curves on STARE-A, STARE-B and CHASEDB1-B, using only the unary potentials (UP, dashed line) or the FC-CRF (solid line), and second human observer (HO, solid dot) performance.

evaluating on the entire set. As mentioned before, using this strategy can underestimate the actual test error. In order to enable a more direct comparison to Azzopardi *et al.*, we performed an additional experiment, named CHASEDB1-B, where we trained our method on the first half of the data set and evaluated it on the remaining images.

Table IV presents the results on both the CHASEDB1 and the CHASEDB1-B configurations, compared to other state of the art strategies. As in CHASEDB1, in CHASEDB1-B the results are increased in terms of both Se ($p \approx 6 \times 10^{-5}$) and Sp, although the improvement in this last metric is not statistically significant for a dataet of this size. This improvement in Se and Sp also increases the corresponding G-mean. The average Pr value is also improved by the FC-CRF, which is explained by the reduction in the number of false positives, as seen in Fig. 4. It is also possible to observe that the unary potentials overestimate the calibre of the narrower vessels, a setting that is improved when incorporating the pairwise potentials. When comparing to Azzopardi *et al.*, it is possible to observe that our method reports better F1-score and MCC values, with higher Pr.

The ROC curve obtained on CHASEDB1-B is presented in Fig. 3. It is possible to observe that the FC-CRF significantly outperforms the unary potentials.

V. EVALUATION ON HRF

We include in Table V results obtained on HRF, applying our method on both downsampled images (FC-CRF) and images in the original resolution (FC-CRF OR), including also a stratification by healthy images (HRF-H), diabetic retinopathy (HRF-DR) and glaucoma (HRF-G).

TABLE III
COMPARISON OF AVERAGE Se, Sp, AUC, Acc, Pr, Re, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS AND THE 2ND HUMAN OBSERVER, BASED ON THE STARE DATA SET.

Method	Se	Sp	AUC	Acc	Pr	Re	F1-score	G-mean	MCC
STARE-A									
FC-CRF	**0.7773**	0.9789	**0.9726**	**0.9571**	0.8039	**0.7773**	**0.7871**	**0.8723**	**0.7654**
Unary potentials	0.6516	**0.9916**	0.9648	0.9552	**0.8928**	0.6516	0.7505	0.8038	0.7393
2nd human observer	0.9385	0.9365	-	0.9378	0.6389	0.9385	0.7593	0.9375	0.7433
Amin *et al.* [15]	0.7261	0.9681	-	0.9199	-	0.7261	-	0.8384	-
Azzopardi *et al.* [1]	0.7588	0.9780	-	0.9553	0.8031	0.7588	0.7772	0.8615	0.7546
STARE-B									
FC-CRF	0.7845	0.9768	-	**0.9576**	**0.7835**	0.7845	**0.7819**	0.8754	**0.7596**
Unary potentials	**0.8483**	0.9564	0.9561	0.9456	0.6794	0.8483	0.7516	**0.9007**	0.7288
2nd human observer	0.9022	0.9341	-	0.9324	0.6249	0.9022	0.7272	0.9180	0.7133
Marin *et al.* [6]*	0.7246	0.9811	-	0.9563	0.8148	0.7246	0.7609	0.8432	-
STARE									
FC-CRF	0.7680	0.9738	0.9570	0.9519	0.7740	0.7680	**0.7644**	0.8648	**0.7417**
Unary potentials	0.7487	0.9744	0.9561	0.9505	0.7743	0.7487	0.7557	0.8541	0.7317
2nd human observer	0.8951	0.9387	-	0.9352	0.6424	0.8951	0.7401	0.9166	0.7225
Supervised									
Cheng *et al.* [**?**]	0.7813	**0.9843**	**0.9844**	**0.9633**	-	0.7813	-	0.8769	-
Dai *et al.* [2]	0.7769	0.9550	-	0.9364	-	-	-	0.8614	-
Fraz *et al.* [3]*	0.7548	0.9763	0.9768	0.9534	-	0.7548	-	0.8572	-
Marin *et al.* [6]*	0.6944	0.9819	0.9769	0.9526	**0.8227**	0.6944	0.7531	0.8257	-
Roychowdhury *et al.* [8]*	0.7719	0.9726	0.9688	0.9515	-	0.7719	-	0.8665	-
Staal *et al.* [9]	-	-	0.9614	0.9516	-	-	-	-	-
Soares *et al.* [10]	0.7200	0.9750	0.9671	0.9480	-	0.7200	-	0.8379	-
You *et al.* [12]	0.7260	0.9751	-	0.9497	-	0.7260	-	0.8414	-
Vega *et al.* [13]	0.7019	0.9671	-	0.9483	-	-	0.6082	0.8239	-
Unsupervised									
Budai *et al.* [17]	0.5800	0.9820	-	0.9386	-	0.5800	-	0.7547	-
Chakraborti *et al.* [18]	0.6786	0.9586	-	0.9379	-	0.6786	-	0.8065	-
Fathi and Naghsh-Nilchi [20]	**0.8061**	0.9717	0.9680	**0.9591**	0.7027	**0.8061**	0.7509	**0.8850**	-
Fraz *et al.* [21]	0.7409	0.9665	-	0.9437	0.7363	0.7311	0.7386	0.8462	0.7003
Fraz *et al.* [22]	0.7318	0.9660	-	0.9423	0.7294	0.7318	0.7306	0.8408	0.6908
Kande *et al.* [23]	-	-	0.9298	0.8976	-	-	-	-	-
Martínez *et al.* [25]	0.7506	0.9569	-	0.9410	-	0.7506	-	0.8475	-
Lam *et al.* [24]	-	-	0.9739	0.9567	-	-	-	-	-
Palomera *et al.* [27]	0.7790	0.9409	-	0.9260	-	0.7790	-	0.8561	-
Odstrcilik *et al.* [28]	0.7847	0.9512	0.9569	0.9341	-	0.7846	-	0.8642	-
Roychowdhury *et al.* [29]	0.7320	0.9840	0.9670	0.9560	-	0.7320	-	0.8487	-
Saffarzadeh *et al.* [**?**]	-	-	0.9431	0.9483	-	-	-	-	-
Wang *et al.* [31]	0.7800	0.9780	-	-	-	0.7800	-	0.8734	-
Zhang *et al.* [32]	0.7177	0.9753	-	0.9484	-	0.7177	-	0.8366	-
Zhao *et al.* [33]	0.7800	0.9780	-	0.9560	-	0.7800	-	0.8734	-

TABLE IV
COMPARISON OF AVERAGE Se, Sp, AUC, Acc, Pr, Re, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS AND THE 2ND HUMAN OBSERVER, WHEN EVALUATING ON THE CHASEDB1 DATA SET. WORKS MARKED WITH AN ASTERISK USED THE TEST DATA TO ESTIMATE PARAMETERS TO OBTAIN THE BINARY SEGMENTATIONS.

Method	Se	Sp	AUC	Acc	Pr	Re	F1-score	G-mean	MCC
CHASEDB1									
FC-CRF	0.7277	0.9712	0.9524	0.9458	0.7438	0.7277	0.7332	0.8407	0.7046
Unary potentials	0.7110	0.9707	0.9284	0.9436	0.7386	0.7110	0.7209	0.8308	0.6919
2nd human observer	0.7425	0.9793	-	0.9539	0.8090	0.7425	0.7686	0.8527	0.7475
Fraz *et al.* [3]*	0.7224	0.9711	0.9712	0.9469	0.7415	0.7224	0.7318	0.8376	-
Fraz *et al.* [35]*	0.7259	0.9770	0.9760	0.9524	0.7732	0.7259	0.7488	0.8421	-
CHASEDB1-B									
FC-CRF	0.7565	0.9655	0.9478	0.9467	0.6810	0.7565	0.7149	**0.8546**	0.6878
Unary potentials	0.7224	0.9647	0.9371	0.9428	0.6686	0.7224	0.6896	0.8348	0.6617
2nd human observer	0.8362	0.9724	-	0.9602	0.7501	0.8362	0.7900	0.9017	0.7700
Azzopardi *et al.* [1]*	0.7623	0.9556	-	0.9387	0.6338	0.7623	0.6876	0.8535	0.6602

In the case of the downsampled images, it can be observed that, in general, our method performs better than the one presented in [28] when evaluated in terms of G-mean, with the exception of results on HRF-G, where the method by Odstrcilik *et al.* reports a higher value. When decomposed into Se and Sp, it is possible to see that the average Se of our method is higher in all the configurations with the exception of HRF-G. Figure 5 depicts different segmentations obtained

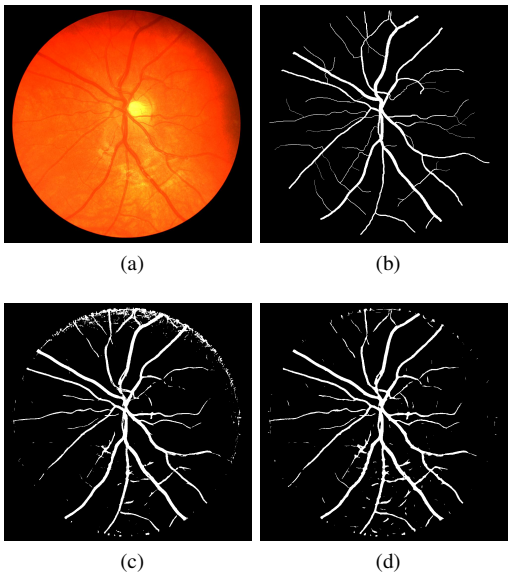


Fig. 4. Segmentation results obtained on CHASEDB1-B. (a) Image_11L. (b) Manual annotations. (c) Segmentations obtained using only the unary potentials. (d) Segmentations obtained using the FC-CRF model.

on HRF-DR, HRF-G and HRF-H. It can be observed that the FC-CRF improves the detection of narrower vessels and also the connectivity between structures that were isolated when using only the unary potentials.

If the FC-CRF is applied on images at the original resolution of the data set, the F1-score and the MCC are increased with respect to using the same model on the downsampled images. This is a consequence of a significant improvement in the Pr value, which is higher than the obtained using the Odstrcilik method. In contrast, when using the fully connected model on lower resolution images, a significantly higher G-mean value is obtained. When analyzed in terms of Se and Sp, it is possible to see that that the FC-CRF achieved a higher Se value on the downsampled images, though the Sp value is lower. When comparing the results qualitatively, it is possible to see that some of the detections obtained when using the original resolution images are isolated and not connected to the main structure. This can be a consequence of using a θ_p value that was adapted with a simple scaling strategy. Additionally, a number of other thin vascular structures that are detected on lower resolution images are still missing at higher resolutions. This can be explained by a largest variability in the calibres of the vessels, that cannot be captured by the relatively small number of feature scales considered. A degradation in the capability of the method to deal with the central reflex of the thicker arteries can be observed at higher resolution images. As is mentioned in the main article, features used in this work are state of the art and were not originally designed to deal with this property. As when using the downsampled images the artifact is reduced, the largest scales of the unary features can better deal with them. Evaluating other features in further works would be extremely valuable to address this issue.

VI. Comparison of computational costs

In Fig 6 it is possible to observe a comparison of the computational cost of the inference in the LNB-CRF and the FC-CRF models. As explained in the main article, these times were calculated by taking the average time of segmenting each image on each of the test sets. It is possible to see that both methods have similar speed. However, it must we highlighted that the LNB-CRF does not offer any improvement with respect of using only the unary potentials, so only the fully connected model yields an improvement.

References

[1] G. Azzopardi *et al.*, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical image analysis*, vol. 19, no. 1, pp. 46–57, 2015.

[2] E. Cheng *et al.*, "Discriminative vessel segmentation in retinal images by fusing context-aware hybrid features," *Machine Vision and Applications*, vol. 25, no. 7, pp. 1779–1792, 2014.

[3] P. Dai *et al.*, "A new approach to segment both main and peripheral retinal vessels based on gray-voting and gaussian mixture model," *PloS one*, vol. 10, no. 6, p. e0127748, 2015.

[4] M. M. Fraz *et al.*, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 9, pp. 2538–2548, 2012.

[5] M. Niemeijer *et al.*, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 648–656.

[6] C. A. Lupascu *et al.*, "Fabc: retinal vessel segmentation using adaboost," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 5, pp. 1267–1274, 2010.

[7] D. Marín *et al.*, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 1, pp. 146–158, 2011.

[8] J. I. Orlando and M. Blaschko, "Learning fully-connected CRFs for blood vessel segmentation in retinal images," in *MICCAI 2014, LNCS*, P. Golland, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer, 2014, vol. 8149, pp. 634–641.

[9] S. Roychowdhury *et al.*, "Blood vessel segmentation of fundus images by major vessel extraction and sub-image classification," *Biomedical and Health Informatics, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2014.

[10] J. Staal *et al.*, "Ridge based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 4, pp. 501–509, 2004.

[11] J. V. Soares *et al.*, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *Medical Imaging, IEEE Transactions on*, vol. 25, no. 9, 2006.

[12] L. Xu and S. Luo, "A novel method for blood vessel detection from retinal images," *Biomedical engineering online*, vol. 9, no. 1, p. 14, 2010.

[13] X. You *et al.*, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognition*, vol. 44, no. 10, 2011.

[14] R. Vega *et al.*, "Retinal vessel extraction using lattice neural networks with dendritic processing," *Computers in biology and medicine*, vol. 58, pp. 20–30, 2015.

[15] B. Al-Diri *et al.*, "An active contour model for segmenting and measuring retinal vessels," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 9, pp. 1488–1497, 2009.

[16] M. A. Amin and H. Yan, "High speed detection of retinal blood vessels in fundus image using phase congruency," *Soft Computing*, vol. 15, no. 6, pp. 1217–1230, 2011.

[17] P. Bankhead, C. N. Scholfield, J. G. McGeown, and T. M. Curtis, "Fast retinal vessel detection and measurement using wavelets and edge location refinement," *PloS one*, vol. 7, no. 3, p. e32435, 2012.

[18] A. Budai *et al.*, "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, 2013.

[19] T. Chakraborti *et al.*, "A self-adaptive matched filter for retinal blood vessel detection," *Machine Vision and Applications*, vol. 26, no. 1, pp. 55–68, 2014.

TABLE V
COMPARISON OF AVERAGE Se, Sp, AUC, Acc, Pr, Re, $F1$-SCORE, G-MEAN AND MCC VALUES OF OUR METHOD WITH RESPECT TO OTHER EXISTING BLOOD VESSEL SEGMENTATION ALGORITHMS WHEN EVALUATED ON THE DIFFERENT SUBSETS OF THE HRF DATA SET. FC-CRF OR CORRESPONDS TO RESULTS OBTAINED APPLYING THE FULLY CONNECTED MODEL IN THE ORIGINAL RESOLUTION IMAGES.

Method	Se	Sp	AUC	Acc	Pr	Re	F1-score	G-mean	MCC
HRF									
FC-CRF	**0.7874**	0.9584	0.9360	0.9424	0.6630	**0.7814**	0.7158	**0.8687**	0.6897
FC-CRF OR	0.7201	**0.9713**	-	0.9478	**0.7199**	0.7201	0.7168	0.8361	0.6900
Unary potentials	0.7315	**0.9680**	0.9455	0.9459	0.7012	0.7315	0.7127	0.8415	0.6851
Odstrcilik *et al.* [28]	0.7794	0.9650	**0.9679**	**0.9479**	0.6950	0.7794	**0.7324**	0.8672	**0.7065**
HRF-DR									
FC-CRF	**0.7869**	0.9575	-	0.9415	0.6566	**0.7869**	0.7117	**0.8680**	0.6855
FC-CRF OR	0.7198	**0.9709**	-	**0.9473**	**0.7150**	0.7198	0.7143	0.8384	0.6872
Unary potentials	0.7302	0.9675	-	0.9452	0.6958	0.7302	0.7091	0.8405	0.6813
Odstrcilik *et al.* [28]	0.7708	0.9652	**0.9579**	0.9470	0.6921	0.7708	**0.7263**	0.8625	**0.7002**
HRF-G									
FC-CRF	0.7844	0.9614	-	0.9448	0.6799	0.7844	0.7253	0.8684	0.6989
FC-CRF OR	0.7229	**0.9728**	-	**0.9493**	**0.7337**	0.7229	0.7256	0.8385	0.6994
Unary potentials	0.7263	0.9698	-	0.9470	0.7143	0.7263	0.7176	0.8393	0.6900
Odstrcilik *et al.* [28]	**0.7875**	0.9658	**0.9695**	**0.9493**	0.7063	**0.7875**	**0.7425**	**0.8721**	**0.7172**
HRF-H									
FC-CRF	**0.7910**	0.9564	-	0.9410	0.6524	**0.7910**	0.7103	**0.8698**	0.6848
FC-CRF OR	0.7174	**0.9703**	-	0.9468	**0.7109**	0.7174	0.7105	0.8343	0.6835
Unary potentials	0.7382	0.9667	-	0.9455	0.6935	0.7382	0.7113	0.8448	0.6840
Odstrcilik *et al.* [28]	0.7799	0.9640	**0.9764**	**0.9473**	0.6865	0.7799	**0.7254**	0.8671	**0.7023**

[20] L. Espona, M. J. Carreira, M. Penedo, and M. Ortega, "Retinal vessel tree segmentation using a deformable contour model," in *19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[21] A. Fathi and A. R. Naghsh-Nilchi, "Automatic wavelet-based retinal blood vessels segmentation and vessel diameter estimation," *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 71–80, 2013.

[22] M. M. Fraz *et al.*, "Retinal vessel extraction using first-order derivative of gaussian and morphological processing," in *Advances in Visual Computing*. Springer, 2011, pp. 410–420.

[23] ——, "Application of morphological bit planes in retinal blood vessel extraction," *Journal of digital imaging*, vol. 26, no. 2, pp. 274–286, 2013.

[24] G. B. Kande *et al.*, "Unsupervised fuzzy based vessel segmentation in pathological digital fundus images," *Journal of medical systems*, vol. 34, no. 5, pp. 849–858, 2010.

[25] B. S. Lam, Y. Gao, and A.-C. Liew, "General retinal vessel segmentation using regularization-based multiconcavity modeling," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 7, pp. 1369–1381, 2010.

[26] M. E. Martinez-Perez, A. D. Hughes, S. A. Thom, A. A. Bharath, and K. H. Parker, "Segmentation of blood vessels from red-free and fluorescein retinal images," *Medical Image Analysis*, vol. 11, no. 1, pp. 47–61, 2007.

[27] M. S. Miri and A. Mahloojifar, "Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 5, pp. 1183–1192, 2011.

[28] M. A. Palomera-Pérez, M. E. Martinez-Perez, H. Benítez-Pérez, and J. L. Ortega-Arjona, "Parallel multiscale feature extraction and region growing: application in retinal blood vessel detection," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 500–506, 2010.

[29] J. Odstrčilík *et al.*, "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database," *IET Image Processing*, vol. 7, no. 4, pp. 373–383, 2013.

[30] S. Roychowdhury *et al.*, "Iterative vessel segmentation of fundus images," *Biomedical Engineering, IEEE Transactions on*, vol. 62, no. 7, pp. 1738–1749, 2015.

[31] V. M. Saffarzadeh *et al.*, "Vessel segmentation in retinal images using multi-scale line operator and k-means clustering," *Journal of medical signals and sensors*, vol. 4, no. 2, p. 122, 2014.

[32] M. Vlachos and E. Dermatas, "Multi-scale retinal vessel segmentation using line tracking," *Computerized Medical Imaging and Graphics*, vol. 34, no. 3, pp. 213–227, 2010.

[33] L. Wang, A. Bhalerao, and R. Wilson, "Analysis of retinal vasculature using a multiresolution hermite model," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 2, pp. 137–152, 2007.

[34] B. Zhang, L. Zhang, L. Zhang, and F. Karray, "Retinal vessel extraction by matched filter with first-order derivative of Gaussian," *Computers in biology and medicine*, vol. 40, no. 4, pp. 438–445, 2010.

[35] Y. Zhao *et al.*, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retina images," *Medical Imaging, IEEE Transactions on*, 2015.

[36] S. A. Salem, N. M. Salem, and A. K. Nandi, "Segmentation of retinal blood vessels using a novel clustering algorithm (RACAL) with a partial supervision strategy," *Medical & biological engineering & computing*, vol. 45, no. 3, pp. 261–273, 2007.

[37] M. M. Fraz *et al.*, "Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification," *International journal of computer assisted radiology and surgery*, vol. 9, no. 5, pp. 795–811, 2014.

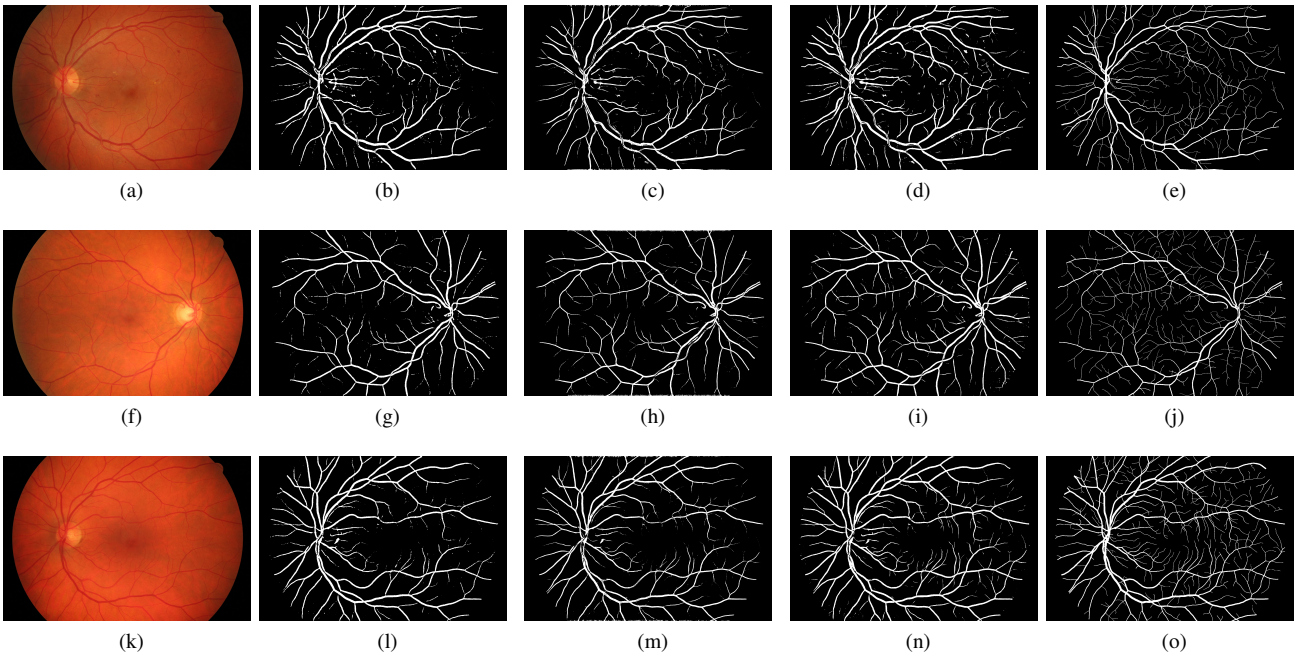


Fig. 5. Segmentation results obtained on HRF-DR, HRF-G and HRF-H. First column: images 07_dr (a), 10_g (f) and 03_h (k). Second column: Segmentations obtained using only the unary potentials in the downsampled images (b,g,l). Third column: Segmentations obtained using the FC-CRF model on the original images (c,h,m). Fourth column: Segmentation obtained using the FC-CRF model on the downsampled images (d,i,n). Last column: Manual annotations (e,j,o).

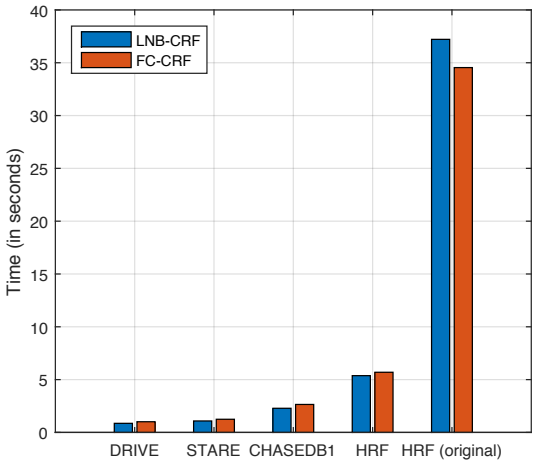


Fig. 6. Computational cost of the LNB-CRF and the FC-CRF inference in all the data sets used for evaluation.