Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images

Álvaro S. Hervella[a,b,*], José Rouco[a,b], Jorge Novo[a,b], Manuel G. Penedo[a,b], Marcos Ortega[a,b]

[a] *CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain*
[b] *Department of Computer Science, Universidade da Coruña, A Coruña, Spain*

Abstract

Background and objectives The analysis of the retinal vasculature plays an important role in the diagnosis of many ocular and systemic diseases. In this context, the accurate detection of the vessel crossings and bifurcations is an important requirement for the automated extraction of relevant biomarkers. Nevertheless, the localization and identification of these vascular landmarks remains challenging.

Method We propose to formulate the detection of vessel crossings and bifurcations in eye fundus images as a multi-instance heatmap regression. In particular, a fully convolutional neural network is trained in the prediction of target heatmaps that are automatically derived from the annotated target pixel coordinates. Then, the network is able to simultaneously estimate the crossings and bifurcations likelihood maps directly from the raw eye fundus images.

Results The proposed method is validated on two public datasets of reference that include detailed annotations for vessel crossings and bifurcations in the corresponding eye fundus images. The conducted experiments evidence that the propose method offers a satisfactory performance for the simultaneous detection of crossings and bifurcations. Furthermore, the proposed method outperforms previous works by a significant margin.

Conclusions The proposed multi-instance heatmap regression allows to successfully exploit the potential of modern deep learning algorithms for the simultaneous detection of retinal vessel crossings and bifurcations. Consequently, this results in a significant improvement over previous methods, which will further facilitate the automated analysis of the retinal vasculature in many pathological conditions.

Keywords: deep learning, eye fundus, blood vessels, crossings, bifurcations

[*] Corresponding author
Email address: a.suarezh@udc.es (Álvaro S. Hervella)

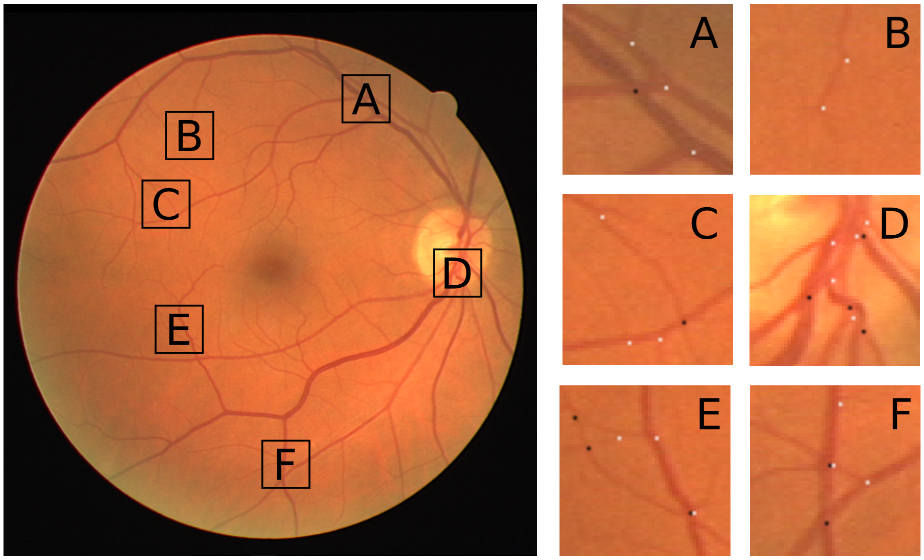


Figure 1: Example of eye fundus image including cropped regions that depict vessel crossings and bifurcations in detail. The black dots represent crossings whereas the white dots represent bifurcations.

1. Introduction

The retinal vascular tree is a complex structure formed by arteries and veins that intersect and bifurcate frequently over all the eye fundus. The analysis of this structure plays an important role in the diagnosis and follow-up of numerous diseases. In particular, the retina is the only organ of the human body where the vascular system can be studied in vivo and without invasive procedures [1]. This makes the analysis of the retinal vasculature relevant for the clinical assessment of both ocular and systemic diseases, such as age-related macular degeneration, diabetes, hypertension, or atherosclerosis, among others [2].

An exhaustive analysis of the retinal vasculature requires the recognition of the vessel crossings and bifurcations, representing the landmarks where blood vessels intersect or bifurcate, respectively. As reference, Figure 1 depicts representative examples of these characteristic points in the eye fundus. The localization and identification of these landmarks has important clinical applications. For instance, the analysis of the bifurcations provides measurements like the bifurcation angles which have been studied as biomarkers for hypertension and other cardiovascular diseases [1]. The identification of the crossings, instead, allows to study the presence of arteriovenous nicking, which happens when an artery compresses a vein. This pathological condition is associated to the development of retinal retinal vein occlusion and it is also an indicative of hypertension, among other relevant diseases [3].

Besides the direct analysis of the vessel crossings and bifurcations, these characteristic points are commonly used as reference in many heterogeneous procedures related to the automated analysis of the retinal vasculature [4, 5]. Moreover, vessel-tracking techniques that are commonly used for the measurement of vessel widths and tortuosity estimation may be affected by an inadequate identification of the constituent crossings and bifurcations [6, 1]. Additionally, these characteristic points can be used as landmarks for the registration of eye fundus images using point matching algorithms [7]. The complexity of the retinal vascular tree, which is unique for each eye, also allows the use of these landmarks as a reliable biometric pattern [8].

The importance of the vessel crossings and bifurcations means that the improvements in their identification present a potential carryover to numerous applications. In that sense, related significative problems such as vasculature segmentation [9] or microaneurysm detection [10] have benefited from the use of Deep Neural Networks (DNNs). The deep learning-based approaches do not require the ad-hoc design of complex algorithms and typically provide an improved performance in comparison with traditional methods [11]. However, the novel use of DNNs may not always be straightforward.

In the case of tasks such as segmentation or classification, a DNN can be directly trained by optimizing a similarity metric between the network outputs and the target binary labels. However, the ground truth labels for the detection of crossings and bifurcations consist of two independent sets of pixel coordinates. In that case, the selection of the most adequate training objective is not straightforward. Additionally, both the number of elements in each of these sets and their approximate spatial distribution are unknown, given that the patterns described by the retinal vasculature are unique for each eye. Thus, the challenge of this task is to adequately formulate the problem to take full advantage of the capacity of a DNN.

In this work, we propose to formulate the detection of retinal vessel crossings and bifurcations as a multi-instance heatmap regression. In particular, we use the sets of annotated pixel coordinates to generate multi-instance heatmaps, which naturally model the likelihood of a pixel being a landmark location. Then, a DNN can be easily trained to predict these multi-instance heatmaps using common regression metrics as loss function. In this setting, the simultaneous detection of crossings and bifurcations is directly enabled by training the network to predict multiple heatmaps, one for each type of target landmark. The precise location of the crossings and bifurcations is obtained by extracting the local maxima in the predicted multi-instance heatmaps. Finally, the proposed approach allows to use fully convolutional networks and, therefore, to make predictions using full images of arbitrary sizes. In order to validate our proposal, several representative experiments are performed using two public datasets of reference that include ground truth manual annotations for both vessel crossings and bifurcations.

1.1. Related work

In the literature, several works have approached the detection of vessel crossings and bifurcations in eye fundus images. The most commonly followed strategy is to split the problem into two different tasks: the general detection of vessel junctions, and the later classification of the detected junctions as crossings or bifurcations [12, 13]. Additionally, there are several works that only tackle the first task, without facing the complex and difficult distinction between both types of landmarks [14, 15].

Regarding the first task, a recurrent approach for the detection of vessel junctions is to start by segmenting the blood vessels. Then, a thinning algorithm is used to obtain the skeleton of the vascular tree, being the vessel junctions extracted after a topological analysis of this skeleton [16, 17]. In this regard, Fahti et al. [18] propose to perform a joint analysis of both the skeleton and the segmented vessels. In these skeleton-based approaches, the most challenging part corresponds to the identification of the vessel crossings, given that, in the obtained skeletons, many crossings are represented as two close bifurcations [16]. In that sense, the classification between crossings and bifurcations is typically performed using geometrical features such as the connectivity [16], the vessel angles [19], and the vessel widths [19]. Alternatively, the vessel landmarks can be directly extracted from the segmented vascular tree by using the adequate combination of shifted Gabor filter responses [15]. Nevertheless, this approach does not allow to distinguish between crossings and bifurcations.

A common drawback of the methods applied over the segmented vessels is that their performance critically depends on the accuracy of the previous vessel tree segmentation. In that sense, several works directly assume that an accurate vascular segmentation is available and evaluate the proposed landmark detection algorithms over manually labeled blood vessels [18, 15]. However, in practice, these ground truth segmentations are not commonly available, being the manual labeling unfeasible in clinical practice routine. An alternative that do not requires an explicit segmentation of the vasculature is to use a vessel tracking algorithm guided by the intensity patterns of the retinal vessels [20]. Additionally, junction likelihood maps can be produced from the eye fundus images by using wavelets to compute orientation scores [12]. Abbasi et al. [12] combine this approach with an skeleton-based method to detect the vessel landmarks. These landmarks are later classified as crossings or bifurcations using the previously obtained dominant orientations.

The generation of junction likelihood maps has also been attempted by using DNNs [14]. However, the successful training of this task with common deep learning approaches is challenging. As reference, Uslu et al. [14] trained a multi-task network that predicts a rough estimation of the junction patterns. However, the extraction of the vessel landmarks from the network output still requires significant post-processing, similar to that applied in skeleton-based methods.

A different approach to solve the landmark detection with DNNs consists in training a patch-wise classifier [13]. Then, the predictions of overlapping patches are aggregated to obtain the final landmark estimations. Pratt et al. [13] combine this approach with a subsequent network to predict whether the patches that are identified as containing landmarks correspond to crossings or bifurcations. In this case, the vessel landmarks are both detected and classified. However, the method does not take advantage of the DNNs capacity to simultaneously perform both tasks, neither of their ability to integrate more representative information from larger contexts in comparison to the reduced analysis in small local patches.

In contrast with previous approaches, our proposal allows to succesfully generated both the crossings and bifurcations likelihood maps from the raw eye fundus images in a single simultaneous step. Besides the computational benefits, the use of a single network applied over large contexts significantly increase the feedback for learning the recognition of the vessel landmarks, which benefits the final performance. This is achieved by training a DNN in the prediction of multi-instance heatmaps that are automatically derived from the annotated pixel coordinates.

The use of a heatmap regression as surrogate task for the localization of landmarks has been previously explored in other domains. In particular, human pose estimation [21] and facial landmark detection [22] have been successfully approached by predicting landmark-derived heatmaps. Nevertheless, these tasks are typically performed over previously detected bounding boxes, which allows to only target the estimation of a known number of landmarks at a fixed scale. In contrast, the size of the blood vessels varies throughout the eye fundus whereas the number of vessel landmarks significantly varies among images. Thus, in our proposal, the networks learn to detect the required patterns at multiple scales and to generate output heatmaps containing multiple instances of the same target landmark type.

2. Materials and methods

2.1. Multi-instance heatmap regression

The detection of vessel crossings and bifurcations in eye fundus images requires the prediction of each landmark location as well as the distinction between the two possible types of landmarks: crossings and bifurcations. The straightforward alternative to tackle the detection of landmarks with fully convolutional networks would imply the prediction of a binary map where only the pixels corresponding to the ground truth location of each landmark are labeled as positive class. However, those target binary maps are heavily unbalanced given that the number of landmark coordinates is much lower than the total number of pixels in the images. As a consequence, the labels provide limited feedback for training a DNN and over-penalize wrong but close predictions to the ground truth landmarks. An improved alternative is to transform the binary ground truth maps into heatmaps where the maximum values correspond to the labeled locations and progressively lower values are assigned to the surrounding pixels. This improved heuristic strategy increases the information from the labels that is available to the network, improving the feedback for learning the detection task. Additionally, the heatmap approach takes into account the potential noise in the labels, transforming the hard binary labels into soft labels that better model the likelihood of a pixel being a target landmark location. For instance, in the considered task, the patterns that represent each crossing or bifurcation

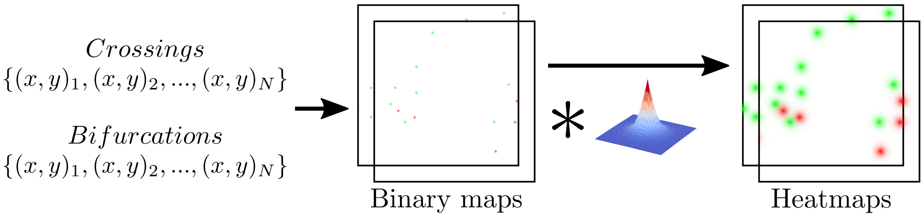

Crossings
$\{(x,y)_1,(x,y)_2,...,(x,y)_N\}$
Bifurcations
$\{(x,y)_1,(x,y)_2,...,(x,y)_N\}$
Binary maps
∗
Heatmaps

Figure 2: Generation of the target heatmaps from the annotated pixel coordinates.

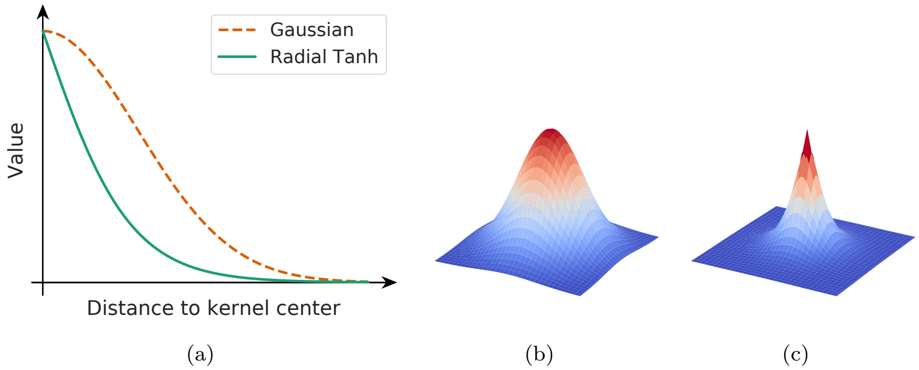


Figure 3: (a) Comparison of the different kernel profiles. ((b),(c)) Kernels represented as a three-dimensional surface. (b) Gaussian. (c) Radial Tanh.

comprise several pixels and, therefore, the precise labeling of its center is error-prone, specially for thick vessels that cover a wide region (e.g., Figure 1(A)). In addition, many of the thin vessels present low contrast, which also makes difficult the labeling (e.g., Figure 1(B)). Hence, the use of soft labels may be beneficial in these frequent scenarios.

The generation of the ground truth heatmaps is summarized in the diagram of Figure 2. In particular, the annotated pixel coordinates are used to create the binary maps with the target locations labeled as the positive class. Then, the ground truth heatmaps are generated convolving the original binary maps with an isotropic kernel of convex and monotonic decreasing kernel profile. Given that there is no prior evidence of the most adequate specific kernel profile for the considered task, we explore the use of two different alternatives: a Gaussian kernel and a Radial Hyperbolic Tangent (Radial Tanh) kernel. The Gaussian kernel has been previously explored for the localization of landmarks in other application domains [21] whereas the Radial Tanh kernel is an alternative depicting a sharper profile, which may facilitate the detection task. Figure 3 depicts a visual comparison between both kernel types. The Gaussian (K_G) and Radial Tanh (K_{RT}) kernel are defined as:

$$K_G(x,y;\sigma) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

$$K_{RT}(x, y; \alpha) = 1 + tanh\left(-\frac{\pi\sqrt{x^2 + y^2}}{\alpha}\right) \tag{2}$$

where (x, y) are the pixel coordinates with respect to the kernel center, σ is the standard deviation for the Gaussian kernel, and α is the saturation distance for the Radial Tanh kernel. Both the standard deviation of the Gaussian kernel and the saturation distance of the Radial Tanh kernel allow to control the region of influence for each landmark. In order to facilitate the comparison between both alternatives, we define an equivalent saturation distance for the Gaussian kernel. In particular, we empirically set this parameter to a value of 2.5 standard deviations, i.e., $\sigma = 0.4\alpha$.

Regarding the distinction between crossings and bifurcations, it is approached by the prediction of two independent heatmaps, one for each type of landmark. In this case, the neural network has to generate a two-channel output. Nevertheless, this setting strongly penalizes the misidentification of a crossing as bifurcation, or vice versa. For instance, using common regression metrics, the error when predicting a crossing in the bifurcation channel would be higher than the error when not predicting any landmark at all. Although this seems to be adequate for the final trained network, it complicates the learning process in the early stages of the training. Thus, the neural network is trained to predict a third channel that includes both landmarks, which further encourages the detection of vessel landmarks regardless of their type.

The simultaneous regression of the three multi-instance heatmaps is trained using the mean squared error (MSE) between the predicted and the target heatmaps as loss. Thus, the training loss is defined as:

$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathbf{y}; \alpha) = ||\mathbf{f}(\mathbf{x}) - \mathbf{y} * \mathbf{K}(\alpha)||_2^2 \tag{3}$$

where $\mathbf{x}$ is an eye fundus image, $\mathbf{y}$ the corresponding target binary map, $\mathbf{f}$ the transformation given by a DNN that generates the predicted heatmaps, and $\mathbf{K} \in \{\mathbf{K}_G, \mathbf{K}_{RT}\}$ the convolutional kernel used to generate the target heatmaps.

The pixel coordinates of the target landmarks are recovered from the heatmaps by directly detecting the local maxima. In particular, we use a maximum filter and an intensity threshold to only retrieved the most salient local maxima. The threshold is required for the predicted heatmaps given the likely slight background noise that is produced by the network, preserving only the significative landmark detections. Additionally, this threshold allows to calibrate the proposed method to different operating points according to the requirements of each specific application. The half-size of the maximum filter must be, at most, lower than the minimum expected distance between landmarks of the same type. The minimum distance between different types of landmarks, i.e., between crossings and bifurcations in this case, does not affect because they are predicted in different output channels of the network.

2.2. Network architecture and training

In order to validate the proposed multi-instance heatmap regression for the identification of crossings and bifurcations, we use an standard network architecture and training procedure. In that sense, the experiments in this work are conducted using an U-Net network architecture [23]. This network represents a reliable baseline, being commonly used in many medical image analysis procedures. Particularly, U-Net has demonstrated to produce satisfactory results for related tasks performed on eye fundus images [24, 25]. Hence, it is expected to be also adequate for the detection of crossings and bifurcations on the same domain. In brief, U-Net is characterized by an encoder-decoder structure, including skip connections between the inner layers of the encoder and the decoder. These skip connections concatenate feature maps taken from the encoder with those of the same spatial resolution in the decoder. The main building blocks of the network consists of convolutional layers with 3×3 kernels and ReLU activation functions, following the idea of the VGG networks. We use a network of the same size than the original one proposed in [23].

The network parameters are initialized with a zero-centered normal distribution following the method proposed by He et al. [26]. For the optimization, we use the Adam algorithm [27] with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The network is trained with full resolution images and batch size of one image. The initial learning rate is set to $\alpha = 1e - 4$, being reduced by a factor of 10 when the validation loss does not improve for 2500 batches. The training is early stopped after reaching a final learning rate of $\alpha = 1e - 7$. The validation set is composed of the 25% of the available training data. To avoid overfitting during training, we use spatial data augmentation consisting of random affine transformations applied to the input eye fundus images and the ground truth pixel coordinates of the target landmarks. Additionally, we also use color data augmentation consisting of random transformations of the image components in HSV color space, similar to the satisfactory application in the same domain of [28].

2.3. Datasets

The experiments in this work are performed using the publicly available DRIVE and IOSTAR datasets. In particular, the ground truth annotations for the identification of crossings and bifurcations in both datasets are provided by [12][1]. The DRIVE dataset [29] comprises 40 color fundus images that are divided by default into balanced training and test sets of 20 images each. The images present a field of view of 45° and a resolution of 565×584 pixels. In contrast, the IOSTAR dataset [12] is a collection of 24 scanning laser ophthalmoscope (SLO) images with a field of view of 45°. The images present varying resolutions but keep the same scale as the DRIVE dataset. SLO is a variant of eye fundus imaging that provides increased contrast with respect to traditional

[1] www.retinacheck.org/datasets

color fundus. In particular, the images of IOSTAR have been captured using green and infrared lasers.

The locations of the crossings and bifurcations have been annotated and reviewed by three different experts for both datasets [12]. In particular, the DRIVE dataset presents an average of 100 bifurcations and 30 crossings per image, whereas the IOSTAR dataset presents an average of 55 bifurcations and 23 crossings per image.

Following the common practices in previous works [12, 14], the DRIVE training set is used for training the networks, whereas the DRIVE test set and the IOSTAR dataset are hold out for evaluation purposes.

2.4. Evaluation

The evaluation of the proposed approach is performed by comparing the detected crossings and bifurcations against the ground truth annotations. In that regard, an independent analysis is performed for each type of landmark (crossings or bifurcations). As gold standard, a detected landmark is considered a True Positive (TP) when it is located within a specified distance d of a ground truth landmark and a False Positive (FP) otherwise. Each ground truth landmark can only be detected once, i.e., we establish a one-to-one correspondence between the set of predictions and the set of ground truth landmarks. In case of several landmarks within the range d of a prediction, the closest one is considered as its corresponding. The ground truth landmarks that remain undetected are considered False Negatives (FN). Then, TP, FP, and FN measures are used to compute Precision and Recall, which are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Additionally, we compute the F-score (F_1), which is the harmonic mean of Precision and Recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

The described analysis is performed using a distance of 5 pixels ($d = 5$) as criteria to consider the detected landmarks as valid, as defined in other works [14]. This represents a approximate real distance of $125\mu m$ and $140\mu m$ for the DRIVE and IOSTAR datasets, respectively [12].

Additionally, we also measure the localization error for the detected landmarks, which is specially relevant for applications such as registration, vascular change detection, or authentication. The localization error is computed as the average euclidean distance between the detected ground truth landmarks and their corresponding predictions. The higher bound for this localization error is given by the maximum distance required to consider a detection as valid, which in this case is 5 pixels.

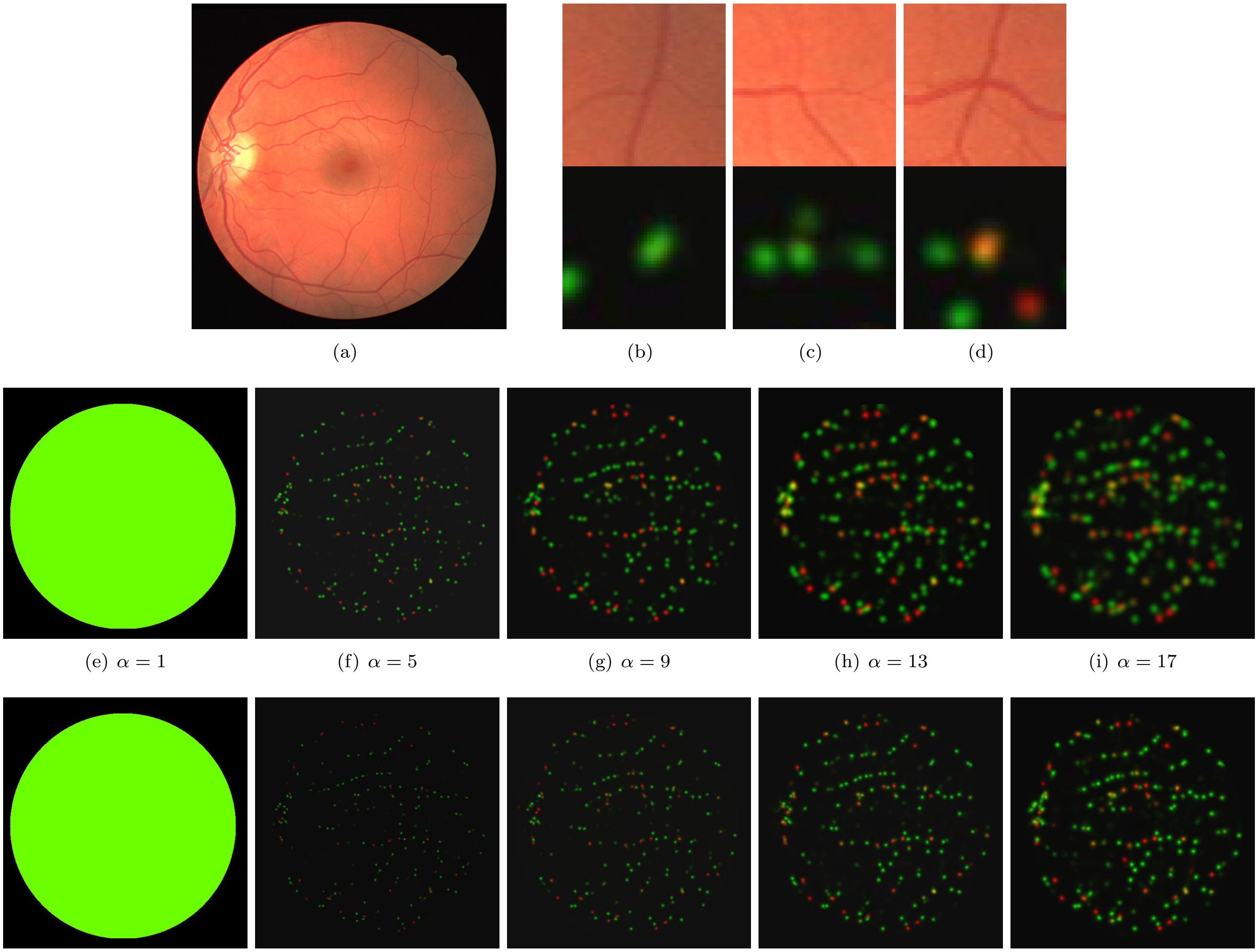
3. Results and discussion

Figure 4 depicts representative examples of predicted heatmaps for networks that were trained using Gaussian or Radial Tanh kernels. In particular, predicted heatmaps corresponding to varying kernel sizes are depicted for each kernel type. The different kernel sizes are specified by the saturation distance parameter defined in Section 2.1. In the examples, the crossings are represented in the red channel whereas the bifurcations are represented in the green channel. Each one of the blobs depicted in the images corresponds to an identified crossing or bifurcation, whose most likely location is given by the local maximum in the center of the blob region. It is observed that for some experiments the output of the network is nearly constant, which is due to the network failing to converge during training (see Figure 4 (e),(j)). This only happens for very small kernels, which make the task very similar to the prediction of the raw binary targets.

In contrast with the use of binary targets, the prediction of heatmaps offers more useful output feedback. Regarding the general appearance of the predicted heatmaps, most of the blobs present similar shape and intensity values, although some exceptions are observed. In this regard, there are elongated or low intensity blobs that differ from the model that the network learns during training. Given that the network learns to generate an specific pattern only when a crossing or bifurcation is detected, the generation of an altered output may evidence a less confident prediction. Thus, an elongated blob may indicate uncertainty in the precise location of the detected landmark (e.g., Figure 4(b)), whereas the low intensity blobs may indicate uncertainty regarding the presence of that landmark (e.g., Figure 4(c)). Additionally, the example of Figure 4(d) shows how the network successfully deals with overlapping crossings and bifurcations. In this case, the predicted crossing and bifurcation blobs partially overlap, which results in a yellowish tone in the output of this depicted example.

Regarding the comparison between Gaussian and Radial Tanh kernels, it is observed that, for the same kernel scale, the Gaussian kernel results in the generation of apparently larger and blurrier blobs. This effect is due to the more disperse distribution produced by the Gaussian kernel in comparison to the sharper one produced by the Radial Tanh variant, being the latter more concentrated around the specific identified landmark location.

In order to quantitatively evaluate the final objective of the proposed methodology, we perform the analysis described in Section 2.4. The local maxima are extracted from the predicted heatmaps as indicated in Section 2.1. In this case, we use a variable threshold, which allows to plot the Precision-Recall (PR) curves represented in Figure 5. The curves are depicted for both Gaussian and Radial Tanh kernels, as well as for different kernel scales. Additionally, the maximum F-score is computed for every experiment, which provides a representative operating point for subsequent comparisons. Simultaneously, we measure the average localization error for each experiment and threshold value in the PR curves. Figure 6 depicts these results by plotting the localization error against the recall measures. To facilitate the comparison with other results, the points

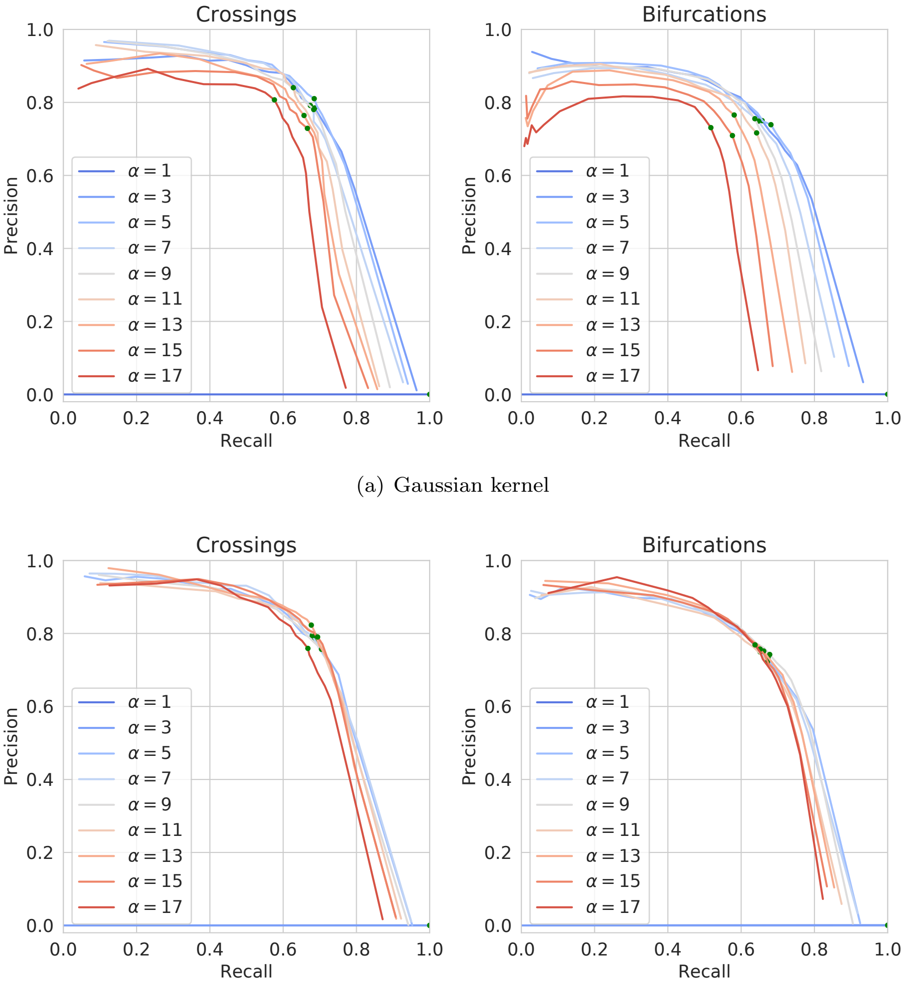


(a) (b) (c) (d)

(e) $\alpha = 1$ (f) $\alpha = 5$ (g) $\alpha = 9$ (h) $\alpha = 13$ (i) $\alpha = 17$

(j) $\alpha = 1$ (k) $\alpha = 5$ (l) $\alpha = 9$ (m) $\alpha = 13$ (n) $\alpha = 17$

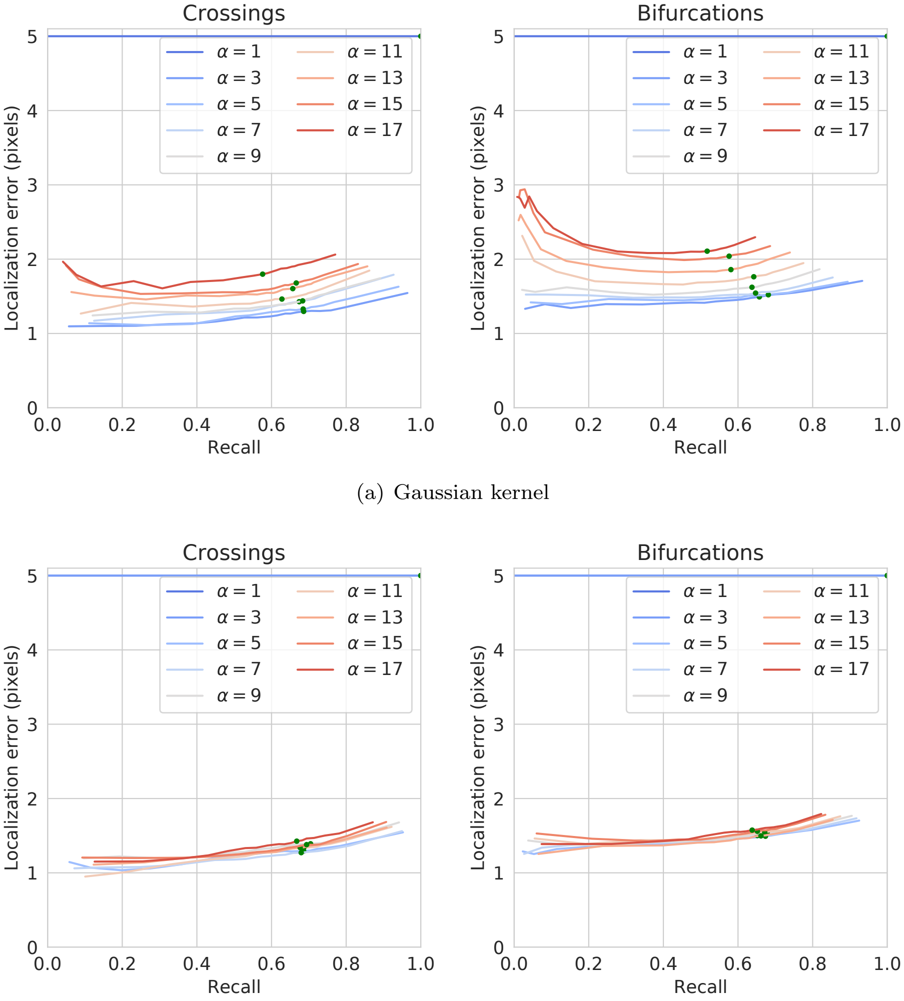
Figure 4: Examples of predicted heatmaps where crossings are represented in the red channel and bifurcations in the green channels. (a) Eye fundus image from the DRIVE test set. (b-d) Regions cropped from (a) that depict both the original image and the predicted heatmaps in detail. (e-i) Predicted heatmaps for (a) using the Gaussian kernel at progressive varying kernel scales. (j-n) Predicted heatmaps for (a) using the Radial Tanh kernel at progressive varying kernel scales.



(a) Gaussian kernel

(b) Radial Tanh kernel

Figure 5: Precision-Recall curves for the detection of crossings and bifurcations in the DRIVE test set at progressive varying kernel scales. The green dots represent the operating points of maximum F-score.



(a) Gaussian kernel

(b) Radial Tanh kernel

Figure 6: Localization error (in pixels) against Recall for the detection of crossings and bifurcations in the DRIVE test set at progressive varying kernel scales. The green dots represent the operating points of maximum F-score.

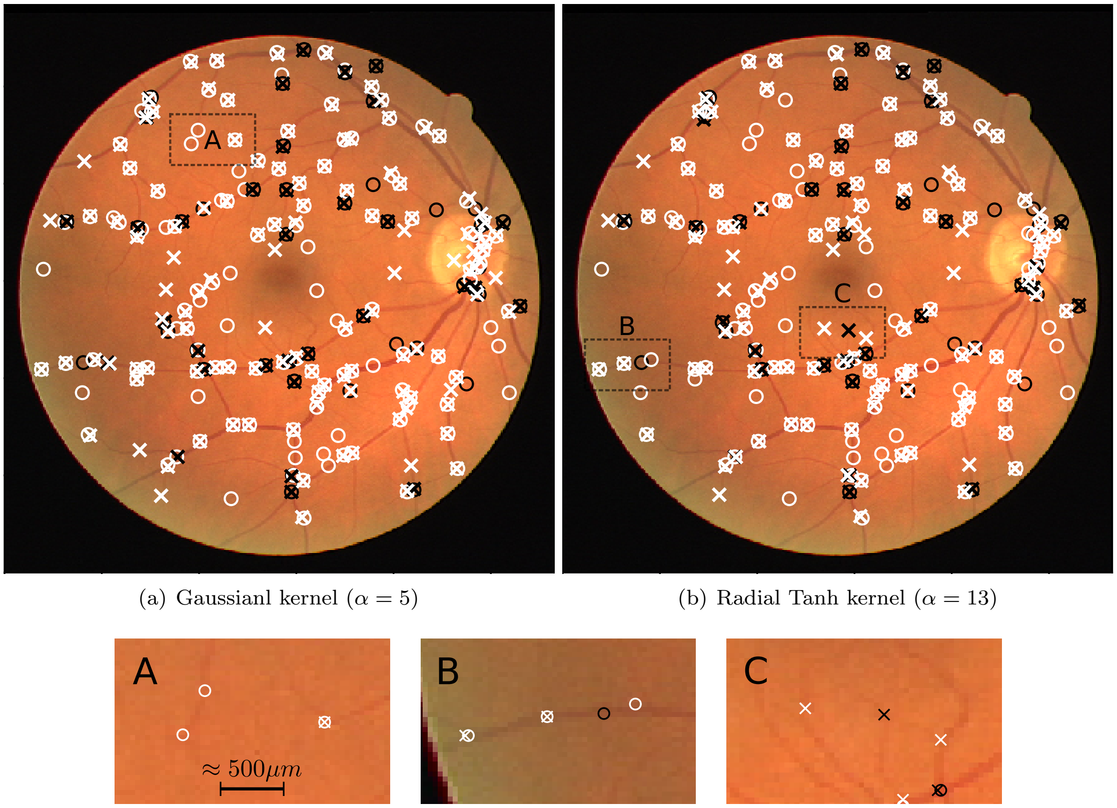
of maximum F-score are also indicated.

As previously seen in the examples of Figure 4, the experiments with the smallest kernels do not converge and result in almost zero precision for any applied threshold. This matches with the constant output depicted in Figures 4(e),(j). Also, for those experiments, the localization error is 5 pixels, which is the maximum for the performed evaluation. However, once the kernel size is increased to the minimum required for convergence, the performance improves drastically. In this regard, it should be noticed that the smallest kernels will produce little change in the original binary maps, resulting in almost binary heatmaps that provide limited feedback for training the networks. However, slightly increasing the kernels size, the region of influence for each landmark is also increased. This results in an improved heuristic for learning the detection task.

In the case of the Gaussian variant, the best performance is obtained for the smallest kernels (after removing the non-convergence case) and it is gradually reduced with the increase of the kernel size. This happens in terms of both PR analysis and localization error. In contrast, for the Radial Tanh variant, a similar performance is obtained for the different kernel sizes. The only exception is the largest kernel when evaluating the detection of crossings. Nevertheless, if the analysis is reduced to the high recall region, the smaller kernels are able to produce higher recall values. This trend is similar to that of the Gaussian kernels, albeit on a smaller scale.

Figure 7 depicts representative examples of detected crossings and bifurcations over an analyzed eye fundus image from the DRIVE test set. The detected landmarks are represented with crosses whereas the ground truth landmarks are represented with circles. At the same time, the black color denotes crossings and the white denotes bifurcations. The provided examples correspond to the operating points of highest F-score, which are marked in the plots of Figures 5 and 6. These examples show that the method detects the majority of the landmarks, while it simultaneously distinguish between crossings and bifurcations. Regarding the missing landmarks and false detections, most of them correspond to secondary tiny vessels (as reference, see Figure 7(c)). In these cases, the crossings and bifurcations are very difficult to appreciate and, therefore, their analysis is typically not considered in the clinical practice. Moreover, the small size and low contrast of these tiny vessel also makes the labeling more error-prone, which complicates both the training and evaluation. Discarding these extreme scenarios, in general, the method offers and adequate performance for both main and secondary branches of the vascular tree. Additionally, the examples show that the results obtained with the two different kernels are similar, at least when an adequate kernel scale is selected. In particular, many of the missing landmarks and false detections are the same for both variants.

In summary, the obtained results demonstrate that the multi-instance heatmap regression approach is adequate for the detection of crossings and bifurcations in eye fundus images. In the performed experiments, the use of very small kernels led to the networks failing to converge during training. However, as said before, the smallest kernels in our experiments are almost equivalent to not using any



(a) Gaussianl kernel ($\alpha = 5$)

(b) Radial Tanh kernel ($\alpha = 13$)

(c) Cropped regions in detail

Figure 7: Examples of detected crossings (in black) and bifurcations (in white) over an eye fundus image from the DRIVE test set. The circles denote ground truth annotations whereas the crosses denote detected landmarks. (a-b) Complete eye fundus images. (c) Cropped regions from (a) and (b) depicting representative examples of missing landmarks and false detections.

kernel at all and, instead, directly training the prediction of the binary target maps. This means that it is precisely the proposed approach which makes possible the detection of vessel crossings and bifurcations using fully convolutional networks.

Regarding the comparative between both types of kernels, the main difference is the higher dependency of the Gaussian variant with respect to the kernel size. In that sense, even if the same or superior performance can be achieved using the Gaussian kernel, in practice its use requires more tuning of the kernel scale. Therefore, its use for related applications depends on the availability of sufficient data, time and computational budget for the validation. In that sense, the advantage of the Radial Tanh kernel is due to the sharper profile. This kernel produces well-defined maxima even when the kernel size is significantly increased. At the same time, it still facilitates the training of the detection task. A trend that is observed for both kernels in the high recall region of the PR curves (Figure 5) is the reduction in recall with the increase of the kernel size. This may be explained by a less defined maxima when the generated blobs get larger as well as the possible overlap of very close landmarks of the same type, which makes extremely complicated to differentiate each one of them. Nevertheless, this happens to a lesser extent for the Radial Tanh kernel, given the mentioned genuine sharper profile.

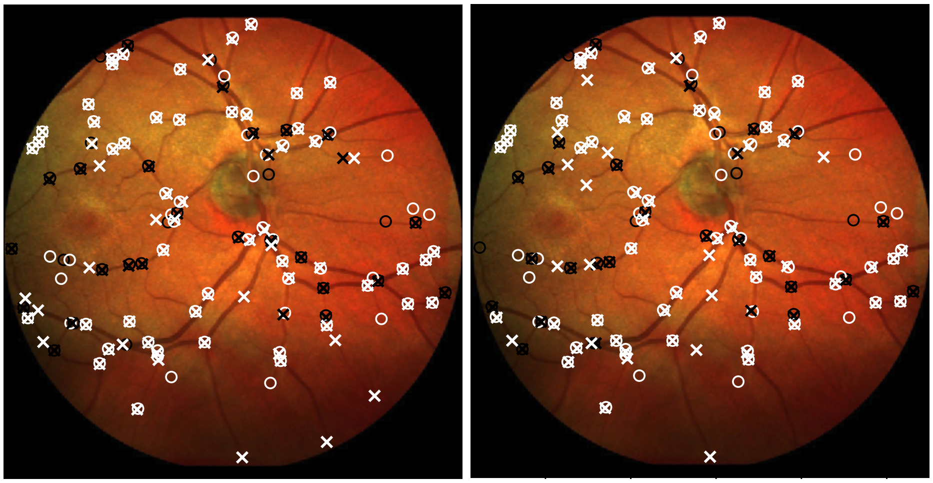
3.1. Comparison with the state-of-the-art

In this section, we compare the performance of the proposed approach against those state-of-the-art works that were evaluated on the same public datasets. To that end, we select the kernel sizes that provide the best performance by means of maximum F-score on the DRIVE training set. Then, the comparison is performed for both the DRIVE test set and the IOSTAR dataset. As reference, Figure 8 depicts examples of detected crossings and bifurcations for the IOSTAR dataset.

In contrast with the proposed approach, previous works typically address the detection of junctions followed by their classification between crossings and bifurcations. This is reflected in their evaluation, which is independently performed for these two steps (detection and classification). To provide an adequate comparison, we reevaluate the trained networks as junctions detectors by merging the predicted sets of crossings and bifurcations. Additionally, the performance as binary classifiers is evaluated over the set of correctly detected junctions. In this case, the crossings are considered as positive samples and the bifurcations as negative ones [12, 13].

Figure 9 depicts the comparison for the detection of junctions. It is observed that the proposed method significantly outperforms previous approaches in both the DRIVE and IOSTAR dataset. Furthermore, the improvement is independent of the selected operating point, given that the performance of the other approaches is always under the PR curves of the proposed method.

In the literature, there are some additional works that reported competitive performance regarding the detection of junctions. However, it should be considered that, in some cases, the evaluation datasets present significantly



(a) Gaussian kernel ($\alpha = 5$) (b) Radial Tanh kernel ($\alpha = 13$)

Figure 8: Examples of detected crossings (in black) and bifurcations (in white) over an eye fundus image from the IOSTAR dataset. The circles denote ground truth annotations whereas the crosses denote detected landmarks.

less detailed annotations [16], whereas, in others, the methods are applied over manually segmented vessels [18]. In that regard, Uslu et al. [14] evaluate their method on both eye fundus images and manually labeled vessels. However, in order to produce an even comparison among all the methods, we do not include the results corresponding to the manual segmentations. Additionally, regarding the provided comparisons, Pratt et al. [13] report individual results for the annotations of three different experts. In this case, we only include the results with the highest accuracy, which is provided by the first expert in their work.

Table 1 depicts the results and comparison for the binary classification between crossings and bifurcations. Given that the classification is evaluated over the correctly detected junctions, the results can vary depending on the operating point for the detection of junctions. Thus, we report classification results for several recall levels in the detection of junctions. These results show that the proposed method outperforms previous approaches at the same level of detection sensitivity. Additionally, our approach also keeps an adequate performance when the detection sensitivity is increased, i.e., when more landmarks are detected.

In summary, the proposed approach leads to a remarkable improvement over the previous existing methods. In that sense, although the use of DNNs had been previously explored, existent works did not achieve a significant improvement over other methodologies. This evidences that the advantage of the methodology in our experiments is due to the proposed multi-instance heatmap regression. This represents a novel alternative to successfully exploit the DNNs capacity for the detection of relevant vascular landmarks as vessel crossings and bifurcations

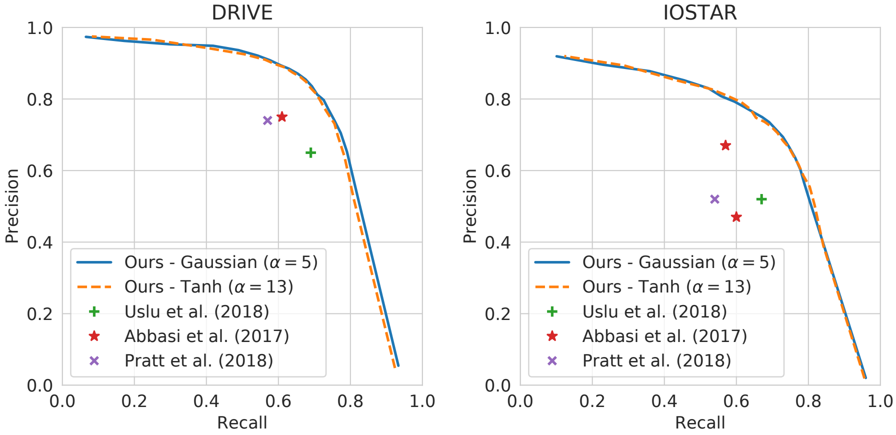


Figure 9: Precision-Recall curves for the detection of vessel junctions without considering their distinction between crossings and bifurcations. Comparison of state-of-the-art works and the proposed approach.

Table 1: Performance for the binary classification between crossings (positive samples) and bifurcations (negative samples). Comparison of state-of-the-art works and the proposed approach. Acc, Sp, and Sn denote accuracy, specificity, and sensitivity, respectively.

Method	Acc (%)	Sp (%)	Sn (%)	Support set
Evaluation on DRIVE				
Abbasi et al. (2016) [12]	83.00	91.00	59.00	Detected with 61.00% recall
Pratt et al. (2018) [13]	80.27	84.82	69.89	All[*]
Ours – Gaussian	93.56	96.91	85.88	Detected with 60.90% recall
	93.83	97.09	86.17	Detected with 71.01% recall
	95.93	97.42	92.27	Detected with 82.61% recall
Ours – Radial Tanh	94.39	97.22	87.53	Detected with 59.24% recall
	93.82	96.62	87.05	Detected with 70.94% recall
	94.93	96.60	90.76	Detected with 81.55% recall
Evaluation on IOSTAR				
Abbasi et al. (2016) [12]	83.00	93.00	67.00	Detected with 57.00% recall
Pratt et al. (2018) [13]	64.79	61.27	74.35	All*
Ours – Gaussian	94.22	95.87	90.37	Detected with 59.14% recall
	92.83	95.22	87.36	Detected with 70.76% recall
	95.59	97.70	90.57	Detected with 80.19% recall
Ours – Radial Tanh	93.93	96.17	88.60	Detected with 61.20% recall
	92.72	94.60	88.24	Detected with 71.23% recall
	95.29	96.53	92.29	Detected with 81.43% recall

[*] The classifier was evaluated on the whole set of ground truth annotations.

in eye fundus images.

Finally, the results provided in this section show that the performance for the detection of junctions on the IOSTAR dataset is not as good as that on the DRIVE test set. In this case, it should be considered that these datasets correspond to two slightly different image modalities, namely color fundus and SLO. Moreover, following the approach of previous works [14, 13], we reserve the whole IOSTAR dataset for evaluation due to its small size. Hence, there is a certain domain shift between training and test in the case of the evaluation on IOSTAR.

4. Conclusions

The automated detection of vessel crossings and bifurcations in eye fundus images represents an important task with numerous practical applications. In that sense, despite the direct analysis for clinical purposes, the detection of these representative landmarks is commonly required as an intermediate step for several automated procedures. In this work, we propose a novel approach for the detection of crossings and bifurcations in eye fundus images. In particular, we reformulate the detection task as a multi-instance heatmap regression which can be performed using a fully convolutional neural network. This allows to simultaneously predict the crossings and bifurcations likelihood maps directly from the raw eye fundus images.

Several experiments are conducted to analyze the proposed approach, including the study of different alternatives to construct the multi-instance heatmaps for training the neural networks. In order to validate the proposal, we use two public datasets of reference with detailed annotations of vessel crossings and bifurcations. Finally, the obtained results demonstrate the advantages of our proposal over the previous existing methods. In that sense, the multi-instance heatmap regression approach allows to further take advantage of modern deep learning algorithms. This leads to a significant improvement in the detection of crossing and bifurcations in eye fundus images.

Acknowledgments

This work is supported by Instituto de Salud Carlos III, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project, and by Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the DPI2015-69948-R and RTI2018-095894-B-I00 research projects. The authors of this work also receive financial support from the ERDF and European Social Fund (ESF) of the EU, and Xunta de Galicia through Centro Singular de Investigación de Galicia, accreditation 2016-2019, ref. ED431G/01, Grupo de Referencia Competitiva, ref. ED431C 2016-047, and the predoctoral grant contract ref. ED481A-2017/328.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] N. Patton, T. M. Aslam, T. MacGillivray, I. J. Deary, B. Dhillon, R. H. Eikelboom, K. Yogesan, I. J. Constable, Retinal image analysis: Concepts, applications and potential, Progress in Retinal and Eye Research 25 (2006) 99 – 127.

[2] M. D. Abramoff, M. K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Reviews in Biomedical Engineering 3 (2010) 169–208.

[3] T. Y. Wong, R. Klein, B. E. Klein, J. M. Tielsch, L. Hubbard, F. Nieto, Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality, Survey of Ophthalmology 46 (2001) 59 – 80.

[4] S. Akbar, M. U. Akram, M. Sharif, A. Tariq, U. ullah Yasin, Arteriovenous ratio and papilledema based hybrid decision support system for detection and grading of hypertensive retinopathy, Computer Methods and Programs in Biomedicine 154 (2018) 123 – 141.

[5] E. Grisan, M. Foracchia, A. Ruggeri, A novel method for the automatic grading of retinal vessel tortuosity, IEEE Transactions on Medical Imaging 27 (2008) 310–319.

[6] S. Kalaie, A. Gooya, Vascular tree tracking and bifurcation points detection in retinal images using a hierarchical probabilistic model, Computer Methods and Programs in Biomedicine 151 (2017) 139 – 149.

[7] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), 2018.

[8] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, M. J. Carreira, Retinal verification using a feature points-based biometric pattern, EURASIP Journal on Advances in Signal Processing 2009 (2009) 235746.

[9] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, L. V. Gool, Deep Retinal Image Understanding, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016.

[10] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, A. Hunter, Microaneurysm detection using fully convolutional neural networks, Computer Methods and Programs in Biomedicine 158 (2018) 185 – 192.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60 – 88.

[12] S. Abbasi-Sureshjani, I. Smit-Ockeloen, E. Bekkers, B. Dashtbozorg, B. t. H. Romeny, Automatic detection of vascular bifurcations and crossings in retinal images using orientation scores, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 189–192.

[13] H. Pratt, B. M. Williams, J. Y. Ku, C. Vas, E. McCann, B. Al-Bander, Y. Zhao, F. Coenen, Y. Zheng, Automatic detection and distinction of retinal vessel bifurcations and crossings in colour fundus photography, Journal of Imaging 4 (2018).

[14] F. Uslu, A. A. Bharath, A Multi-task Network to Detect Junctions in Retinal Vasculature, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 92–100.

[15] G. Azzopardi, N. Petkov, Automatic detection of vascular bifurcations in segmented retinal images using trainable COSFIRE filters, Pattern Recognition Letters 34 (2013) 922–933.

[16] D. Calvo, M. Ortega, M. G. Penedo, J. Rouco, Automatic detection and characterisation of retinal vessel tree bifurcations and crossovers in eye fundus images, Computer Methods and Programs in Biomedicine 103 (2011) 28 – 38.

[17] A. M. Aibinu, M. I. Iqbal, A. A. Shafie, M. J. E. Salami, M. Nilsson, Vascular intersection detection in retina fundus images using a new hybrid approach, Computers in Biology and Medicine 40 (2010) 81–89.

[18] A. Fathi, A. R. Naghsh-Nilchi, F. A. Mohammadi, Automatic vessel network features quantification using local vessel pattern operator, Computers in Biology and Medicine 43 (2013) 587–593.

[19] H. Hamad, D. Tegolo, C. Valenti, Automatic detection and classification of retinal vascular landmarks, Image Analysis Stereology 33 (2014) 189–200.

[20] Chia-Ling Tsai, C. V. Stewart, H. L. Tanenbaum, B. Roysam, Model-based method for improving the accuracy and repeatability of estimating vascular bifurcations and crossovers from retinal fundus images, IEEE Transactions on Information Technology in Biomedicine 8 (2004) 122–130.

[21] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: International Conference on Computer Vision, 2015.

[22] D. Merget, M. Rock, G. Rigoll, Robust facial landmark detection via a fully-convolutional local-global context network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[23] O. Ronneberger, P.Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[24] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal image understanding emerges from self-supervised multimodal reconstruction, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018.

[25] A. S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised deep learning for retinal vessel segmentation using automatically generated labels from multimodal data, in: International Joint Conference on Neural Networks (IJCNN), 2019.

[26] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: International Conference on Computer Vision (ICCV), 2015.

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.

[28] P. Liskowski, K. Krawiec, Segmenting Retinal Blood Vessels with Deep Neural Networks, IEEE Transactions on Medical Imaging 35 (2016) 2369–2380.

[29] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, B. van Ginneken, Ridge based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging 23 (2004) 501–509.