Improving Edge Detection in RGB Images by Adding NIR Channel

Xavier Soria[1]
[1]*Computer Vision Center*
Universitat Autònoma de Barcelona
Barcelona, Spain
xsoria@cvc.uab.es

Angel Sappa[1,2]
[2]*Escuela Superior Politécnica del Litoral, ESPOL*
Facultad de Ingeniería en Electricidad y Computación, CIDIS,
Guayaquil, Ecuador
sappa@ieee.org

***Abstract*—The edge detection is yet a critical problem in many computer vision and image processing tasks. The manuscript presents an Holistically-Nested Edge Detection based approach to study the inclusion of Near-Infrared in the Visible spectrum images. To do so, a Single Sensor based dataset has been acquired in the range of 400nm to 1100nm wavelength spectral band. Prominent results have been obtained even when the ground truth (annotated edge-map) is based in the visible wavelength spectrum.**

***Index Terms*—Edge detection, Contour detection, VGG, CNN, RGB-NIR, Near infrared images.**

I. Introduction

Image edge detection has been a long-standing problem in computer vision, even nowadays when powerful deep learning algorithms have been proposed edge detection still motivates researchers working on this topic. The simplified feature extraction from a given image, either based on the intensity changes of luminosity, color or texture in other words image edge detection is present in most of computer vision based applications. Hence, the usage of a proper edge detection algorithm could represent a reduction on the processing time and/or an increase in the quality of results in image processing tasks such as object recognition (e.g., [1], [2]), registration [3], segmentation [4], medical imaging [5], just to mention a few.

Closely related to the edge detection problem we can find the contour and boundary detection approaches. These approaches, although similar, are intended to unveil the shape of objects contained in the scene through the extraction of closed contours. Different summaries have been proposed in the literature during last two decades [6]–[8]; proposed approaches can be classified as: pixel-based approaches, edge-based approaches, region-based approaches and machine learning based approaches. Most of current approaches are based on machine learning, more specifically Deep Learning (DL) based approaches, where Convolution Neural Networks (CNN) are the most popular DL architecture.

CNN based approaches are becoming the best option in almost every computer vision problem. Since its high impact in 2012, in the large-scale visual recognition challenge [9] where an error rate of 15.3% has been achieved, while the second best entry reached a 26.2% error rate, they are being used to solve challenging visual perception problems in domains such as self-driving cars, mobile-robotics, surveillance, remote sensing and so on [10].

Most of the approaches presented above have been intended for detecting edges in images from the visible spectrum (e.g., RGB images). The visible spectrum (VIS) ranges from 400nm to 700nm. Just after the visible spectrum we have the Near Infrared (NIR) spectral band, which cover from 700nm till 1100nm. The NIR spectral band could be added to visible spectrum (VIS) images to improve image processing issues such as restoration [11], enhancement [12], de-hazing [13] among others, to overcome the state-of-the-art results. Additionally, since the last few years, the NIR sensor technology has been more and more accessible due to the market competition, the mass access to the tech and the development of the new cheap devices like [14].

Numerous advancement have been proposed with the combined usage of VIS and NIR wavelength spectral bands (through this work images from the visible spectral band will be indistinctly referred to as RGB images or VIS images). This paper tackles information from these two spectral bands in order to detect edges present in the given image. These images have been acquired by a Single Sensor Camera (SSC), which capture in one-shot the VIS and NIR spectral bands (from 400nm till 1100nm). The proposed approach is based on the holistically-nested edge detection (HED) work presented in [15], which is modified to tackle as an input a Multi-Spectral image (MSI); it will be referred to as MSI-HED. By adding information further the 700nm wavelength more features are extracted and better edges detected. In Fig. 1, just as an illustration, it can be appreciated that more edges are detected when the multi-spectral image is considered. Note how small details are recovered (e.g., mirror) when the MSI is considered but they are missed if the VIS image is used.

The rest of the paper is organized as follow. In Section II a description of edge and contour detection related approaches is given. Then, the proposed approach is presented in Section

This work has been partially supported by the Spanish Government under Projects TIN2014-56919-C3-2-R and TIN2017-89723-P; by the "CERCA Programme / Generalitat de Catalunya"; and by the ESPOL project PRAIM. The authors gratefully acknowledge the support of the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559). Xavier Soria has been supported by Ecuador government institution SENESCYT under a scholarship contract 2015-AR3R7694.

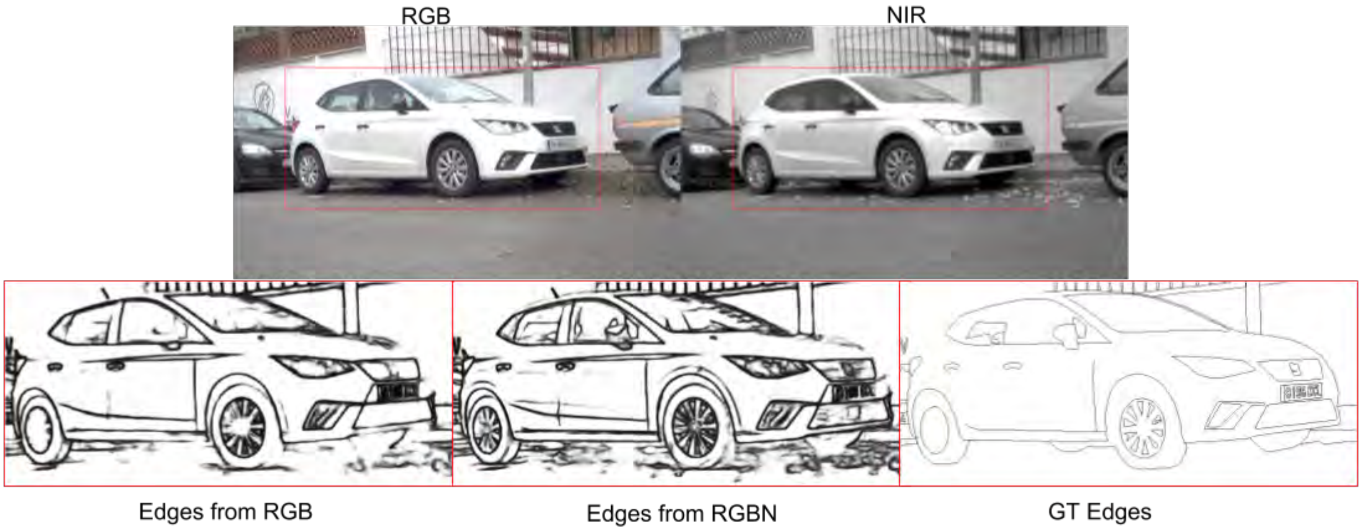


Fig. 1. ($1st$ row) Sample of images used for network training. ($2nd$ row) Edge maps from networks trained with Visible and Multi-spectral images, ground truth image is also depicted.

III; firstly, a short general description of the model is given and then the proposed architecture and the loss function are introduced. Experimental results are described in Section IV; finally, conclusions and future work are presented in Section V.

II. Related Works

Although the edge and contour detections have a long history in the computer vision community (e.g., [6], [8]), in this section just Deep Learning based edge and contour detection algorithms will be reviewed. Almost all the approaches presented bellow are based on Convolutional Neural Networks. There is just a method reviewed in this section, presented in [16] and extended in [17], which is not based on DL; it is based on the usage of another machine learning based technique (i.e., random decision forests). The main idea here is to map all the structured labels at a given node into a discrete set of labels; to this purpose, the authors propose an intermediate mapping (IM), which plays a key role in the training stage randomly generating and applying the training labels (Y) at each node. This process helps to speed up the computation and injects additional randomness into the learning process.

Inspired in AlexNet [9] the authors of [18] present a DL model for boundary detection. It was termed DeepEdge and uses four different scale patches ($64 \times 64, 128 \times 128, 196 \times 196$ and a full-size image), all these patches are resized before training the modified AlexNet architecture. The DeepEdge network uses just the first five convolutional layers of AlexNet; from these five convolutional layers they extract features by using max, average and center point pooling to assess the presence of contours in small areas. In turn, they are connected in a bifurcated sub network (two separately trained nets). The first net is trained by using binary labels (classifying contour), the second one is optimized as a regressor to predict the fraction of human labelers agreeing to the contour presence in a given point. Another deep learning contour detection based method has been proposed in [19] (DeepContour); this proposal uses the BSDS500 dataset [20], which consist of a set of 2,000,000 image patches of 45×45. In order to reach a value of 0.76 in F-measure, a 4 convolutional (conv) layers plus 2 fully connected are set. Similar to DeepEdge, to extract representative features, the first three conv layers use normalization and max pooling, the last conv just use max pooling. The cost of training is minimized by softmax loss.

The holistically-nested edge detection (HED) method [15], which is an extension of its 2015 conference paper, goes much deep than the previously presented DL based models. The essential characteristic of this architecture is the usage of a VGG network [21], deeply supervision and the end-to-end training way (image to edge map prediction). The VGG16 (16 layers) is an architecture with an excellent performance due to its fine parameter tuning; it has a lower computing cost, and a better or equal characteristics extraction than AlexNet [9]. The core contribution of this HED architecture is the holistically nested network, which is intended to produce a prediction from multiple scales extracted at the end of each blocks termed side outputs (five side-outputs). Such side-outputs are associated with a classifier and for the training purpose a class-balanced cross-entropy loss function is defined. Going deeper and inspired in the previous edge-contour detection models, [22] design a DL model by modifying VGG16 (13 conv layes and 3 fully connected layers). The modification consists of keeping the 13 conv layers and cutting the last stage of VGG (pooling and fully connected layers). Such 13 layers are split up into 5 stages, at each stage, an output layer is added termed as side-output for the deep supervision. Then, a cross-entropy loss is applied to all five stage results and the last concatenated one. For improving the results, an image pyramid is used during testing, that is, a different scale input image is tested separately and then all resulting edge maps are up-sampled with bilinear interpolation to finally average all maps in a final prediction.

Recently, [23] propose a CNN model and attention-gated conditional random field, termed attention-guided multi-scale hierarchical deepNet (AMH-Net). It is intended for contour detection, which can learn reacher multi-scale features than CNN models. To be precise, in [15], as well as in [22], the multi-scale feature maps are fused with the concatenation of a 1×1 convolution. In this case, such feature maps are learned to combine the latent features by attention-gated CRF model—a gate allows context-specific independence to be made explicit in the graphical model [24], in this context it permit or block the flow of information between different edge-map scales at every pixel.

Finally, [25] propose a model able to learn more details than previous approaches. Additionally, this architecture does not require the post-processing stage, like non-maximum-suppression, which is generally used in HED and RCF ([22]). The proposed model is based on the usage of Adversarial Neural Networks (GANs), considering UNET [26] for the generator and VGG16 for the discriminator, and inspired in the image to image translation work [27]. The proposed conditional GAN is similar to the work presented in [27], where the difference is that feeding noise is not used as input z.

III. Proposed Approach

This section presents the approach proposed to extract edges from the given multispectral images. The proposed MSI edge detector is based on [15]. Figure 2 presents an illustration of the modified VGG16 architecture, which is described next. Given a RGB or RGB-NIR input image, denoted as X_i^{SSR} (SSR: Spectral Sensitivity Range), and its respective ground truth Y_i^{VIS} (Y is a binary map obtained from RGB images), the estimated edge map ($\hat{Y}_i$) is obtained as follow:

$$\hat{Y}_i = \text{MSI-HED}(X_i^{SSR}, Y_i^{VIS}) \tag{1}$$

where index i is used to specify the edge estimation for a single image. More details on the proposed architecture as well as in the loss function are given below.

A. Network Architecture

Figure 2 shows five subsets (stages) with the respective convolutional hidden layers (e.g., conv1_1). This architecture has been proposed in [21] and then modified by [15] for the edge detection purpose (see description in Sec. II). Based on that modified architecture, in the current work an adaptation to the input layer is proposed in order to receive Multi-Spectral images (RGB-NIR); then the X^{SSR} size is $[image\_width, image\_height, image\_channels]$, where image_channels is 4. The size of the filters in each layers is 3×3. The size of each side-output is the same as the ground truth (e.g., *side-output-5*= $[image\_width, image\_height, 1]$). In order to preserve such a size, in each side-output a convolutional and a transpose convolution layers are incorporated. The transpose convolution or deconvolution is added at the end of such side-output to, according to [15], lead the best performance restoring the size by the up-sampling process. This process is like a bilinear interpolation but in a single layer. Those side-outputs are then used to feed a loss function, which is presented below.

B. Loss Function

As presented in Fig. 2, the MSI-HED architecture has 5 side-outputs (sdo) and 1 fuse-output (fso). Each of such outputs has a predicted edge-map $\hat{y}_i^m$. Hence, $\hat{Y}_i = y_i^{[1,...,M]} = [y_i^{(sdo\_1)}, y_i^{(sdo\_2)}, ..., y_i^{(sdo\_5)}, y_i^{(fso)}]$, $M = 6$. Hence, the loss function for MSI-HED is the same as the one presented in [15], and $\mathcal{L}_{sdo}$ and $\mathcal{L}_{fso}$ are minimized with the objective function, stochastic gradient descent, as follow:

$$(W, w, h)^* = \arg\min(\mathcal{L}_{sdo}(W, w) + \mathcal{L}_{fso}(W, w, h)) \tag{2}$$

where to get $\mathcal{L}_{sdo}(W, w)$, each single sdo has to compute the image-level loss ($\wr_{sdo}$) as follow:

$$\begin{aligned} \wr_{sdo}^m(W, w^m) = &-\beta \sum_{j \in Y_+} \log \sigma(y_j = 1 | X; W, w^m) \\ &- (1 - \beta) \sum_{j \in Y_-} \log \sigma(y_j = 0 | X; W, w^m), \end{aligned} \tag{3}$$

then,

$$\mathcal{L}_{sdo}(W, w) = \sum_{m=1}^{M-1} \wr_{sdo}^m(W, w^m). \tag{4}$$

Therefore, in Eq. 3, W is the collection of all network parameters and w is the sdo corresponding parameter, see Fig. 2 ellipsoidal-shape in green. β = $|Y_-|/|Y|$ and $(1-\beta)$=$|Y_+|/|Y|$ ($|Y_-|$ and $|Y_+|$ denote the edge and non-edge ground truth label sets).

To summarize $\sigma(y_j = 1 | X; W, w^i)$ is a sigmoid function ($\sigma(.)$) on the activation values (a_j^m) at each pixel j for each side-output. On the other hand, in Eq. 2, $\mathcal{L}_{fso}(W, w, h) = Dist(Y_i, y_i^{(fso)})$. As previously described, $y_i^{(fso)}$ is the predicted edge-map trained by a fusion weights. $Dist(.,.)$, on the other hand, is the distance between the fused predictions and the ground truth label map, which is set to be cross-entropy loss.

For the testing purpose, the MSI-HED method gives the following predicted edge-maps, $\hat{Y}_i = [y_i^{(sdo\_1)}, y_i^{(sd\_2)}, ..., y_i^{(sdo\_5)}, y_i^{(fso)}, y_i^{(av)}]$. The $y_i^{(av)}$ is the average of the 5 sdo and the fso predictions.

IV. Experiments

This section presents the implementation setup and a detailed description of the dataset used for the training and testing stages. Additionally, quantitative and qualitative evaluations are presented.

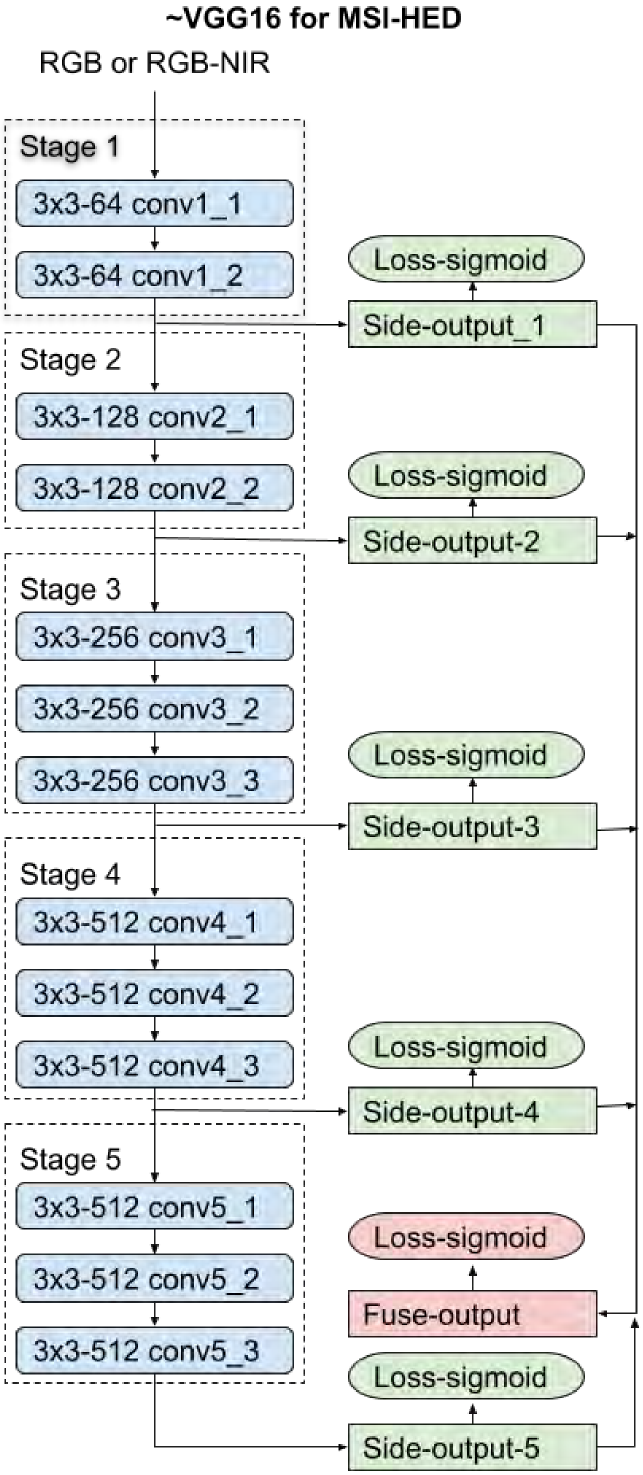


Fig. 2. VGG16 architecture modified by [15] and used in the current work. The Side-output boxes, in green, have convolutional and deconvolutional layers for the up-sampling purpose, see details in Sec. III.

A. *Implementation*

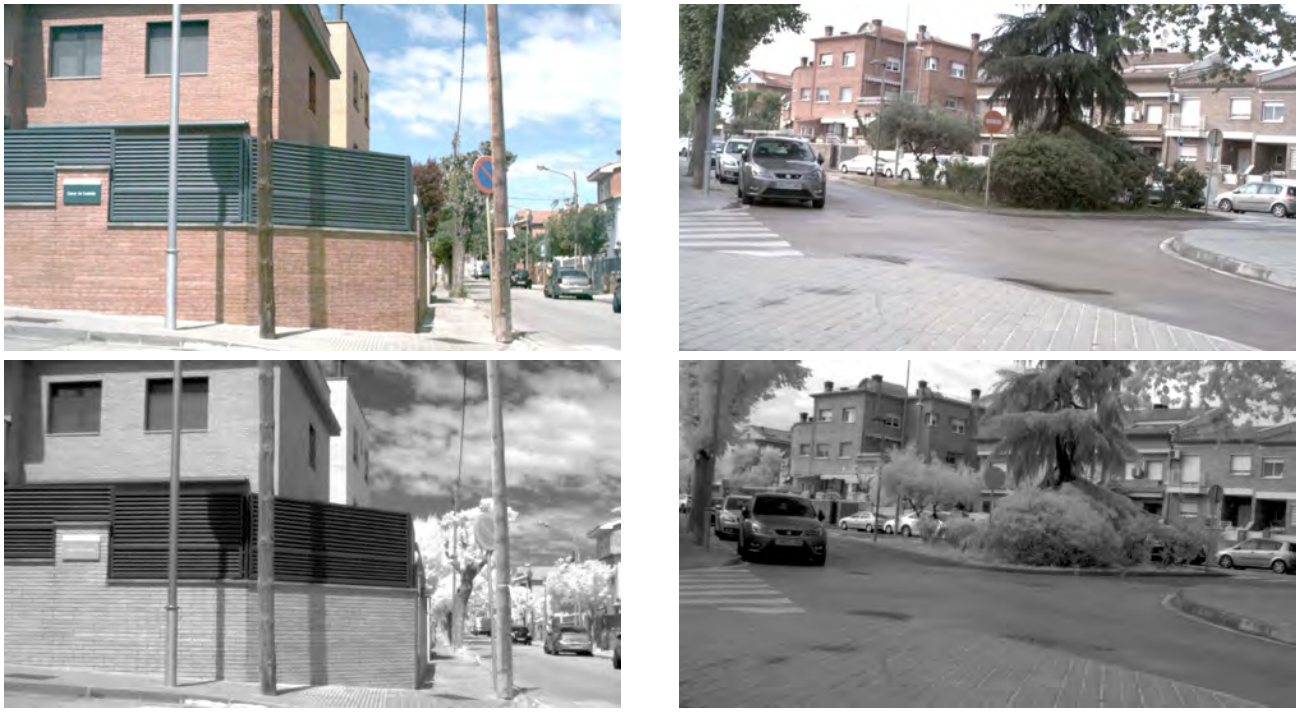
The implementation of the framework is performed by using the open-source machine learning library TensorFlow [28]. Since MSI-HED implementation does not use VGG16 pre-trained data, the training iterations are 90,000 with a 0.003 learning rate, 10 mini-batch size and 0.2 fusion weights initialization. With a difference on the number of training iteration, all the settings have the same parameters as in [15], which has the best performance in F-Score on the BSDS500 validation set. They experimented using different hyper-parameters, non-linear functions and concludes that the parameters exposed above are the best. Therefore, this work used such values. The training process took about 5 days for both, the HED and MSI-HED models, in a TITAN X GPU with the input image size $[480, 480, 4]$ (4 reefers to RGBN).

In order to enlarge the size of the data set a **data augmentation** process is performed. Firstly the given images are split up into two parts; secondly, each of the obtained sub-images are rotated in 15 different angles (cropping the maximum square of a rotated image); and finally they are horizontally flipped, which augmented the dataset by a factor of 64 . On the other hand, as the supervised learning in DL needs the ground truth (GT) for evaluating the results, a **GT generation** process is performed on the visible spectrum images (RGB). This GT is generated by using a crowdsourcing Internet-based application $labelbox.com$ to draw lines or polygons in every edges in the scenario of such image detected by human labeler. In order to avoid wrong edges, every single image is double checked by a supervisor, which review and correct errors or lack of edges in the scene.

B. *Dataset*

As the aim is the edge detection when an extra channel (NIR) is considered, and due to the lack of a dataset for this purpose, a dataset has been collected as a first step. The acquired dataset contains 203 pairs of RGB and RGB-NIR images, which are obtained by a couple of single sensor e-con system camera (e-CAM40_CUMI4682_MOD-4MP). One of these cameras has an infrared-cutoff-filter (IRCF), while the second one does not use IRCF. The acquired dataset, named SSMIHD (Single Sensor Multispectral Images in High Definition), will be available by contacting the authors. From the 203 pairs of images, 170 have been considered for the training and the rest are considered for the testing purpose. The camera maximum resolution, 2K, is set for the acquisition (2688×1520). Nevertheless, for the registration purpose (image acquired with the IRCF and image acquired without the IRCF), the final image size is 2560×1440 and to the better usage of NIR, the final SSMIHD images size is 1280×720 (High Definition size), half of the original size, after a splitting in different channels. As mentioned above, the ground truth has been obtained by using a crowdsourcing based tool on the visible images (RGB). Hence to validate obtained results the RGB and RGBN images are registered by [29].

The SSMIHD dataset contains urban scenarios, mainly consisting of vegetation, road scenes and home infrastructures.



(Sample #1) *(Sample #2)*

Fig. 3. Two samples of the SSMIHD images. ($1st$ row) RGB images acquired by a Single Sensor Camera with an infrared-cutoff-filter (IRCF). ($2nd$ row) Corresponding NIR images.

The acquisition is carried out in the spring season in the range of the day 10:00 till 18:00 when the sunset is at 20:00. Figure 3 presents two samples of such images with their respective NIR images.

C. Results and Discussions

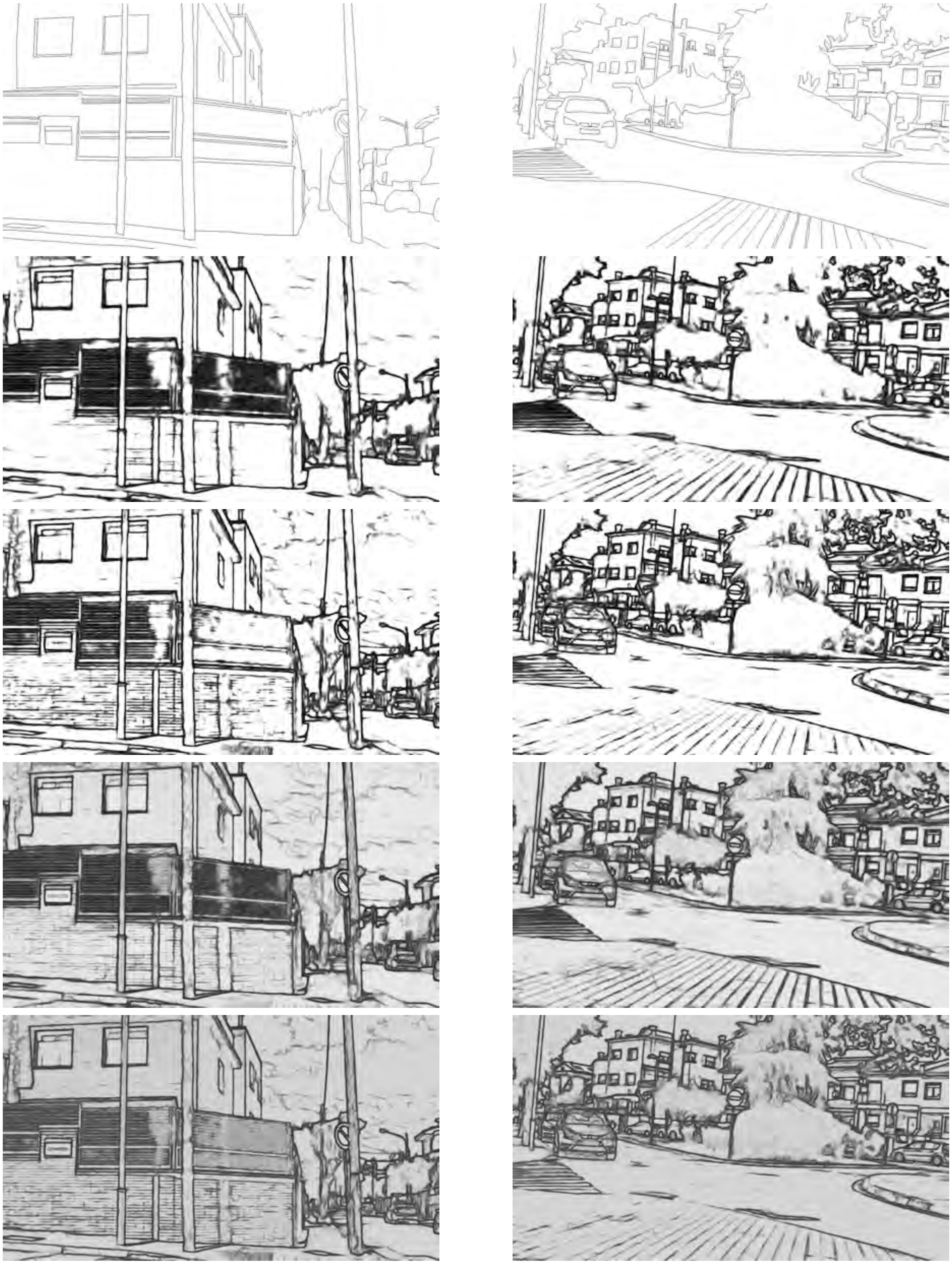
The edge detection accuracy is assessed by the measurements considered in the state-of-the-art literature [15], [17], [19]: Fixed Contour Threshold (ODS), Per-image Best Threshold (OIS), and Average Precision (AP). As in [15], before the evaluation, a non-maximal suppression process is applied to obtain thinned edges. This post-processing stage is performed by [17], a structured edge detection toolbox. For the accurate evaluation of the MSI-HED model performance, from the SSMIHD dataset, visible spectrum images are trained and tested setting up similarly to MSI (just removing the NIR images), This model is termed HED. From both testing stages, two best predictions are considered, fused $y_i^{(fso)}$ and averaged $y_i^{(av)}$.

Fig. 4 shows a couple of images with the corresponding edges extracted by the MSI-HED method for the qualitative evaluation. The first row is the labeled GT of the images in the Fig. 3. Both, RGB and RGBN based predictions, present similar amount of edges like those presented in the GT. Furthermore, both fused and averaged results present a plausible edge even in challenging scenes (note in the illustration in the second column how small details like windows in the buildings are detected). However, in the first column, the first light post (left to right) is not formed properly when the edge is predicted from the VIS images (see in the top, the post through the window). Nevertheless, both fused and averaged predictions from the MSI images, when the light post is considered, has a presence of edges in the whole of its contour. More illustrative results are presented in Fig. 5.

Table I summarizes a statistic comparison in ODS, OIS and AP of fused and averaged predicted edge-maps ($\hat{y}_i^{(fso)}$ and $\hat{y}_i^{(av)}$) with VIS and MSI images. Even though the GTs are annotated in the visible spectrum images, the results in both, fused and merged, have the same value (0.8) according to OIS. On the other hand, when the ODS evaluation measurement is considered the HED overcome MSI-HED results in fused, but just in 0.01; while in averaged both approaches have the same results. On the contrary, when the AP evaluation measurement is considered, the MSI-HED proposed model overcomes the results from HED in 0.01 in both predictions.

TABLE I
QUANTITATIVE RESULTS OF FUSED AND AVERAGED PREDICTIONS, $y_i^{(fso)}$ AND $y_i^{(av)}$. HED REEFERS TO THE MODEL TRAINED AND TESTED WITH VIS IMAGES; MSI-HED REEFERS TO THE PROPOSED MODEL TRAINED WITH MSI IMAGES.

Output Edges	**HED (RGB)**			**MSI-HED (RGBN)**		
	ODS	***OIS***	***AP***	***ODS***	***OIS***	***AP***
Fused	**0.79**	**0.80**	0.80	0.78	**0.80**	**0.81**
Averaged	**0.78**	**0.80**	0.82	**0.78**	**0.80**	**0.83**



(Sample #1) *(Sample #2)*

Fig. 4. Edge-maps predicted by HED and MSI-HED to images in Fig. 3: ($1st$ row) ground truth (manual annotation using $labelbox$; ($2nd$) fused prediction ($\hat{y}_i^{(fso)}$) in VIS image; ($3rd$) fused prediction ($\hat{y}_i^{(fso)}$) in MSI image; ($4th$) averaged prediction ($\hat{y}_i^{(av)}$) in VIS image; ($5th$) averaged prediction ($\hat{y}_i^{(av)}$) in MSI image.

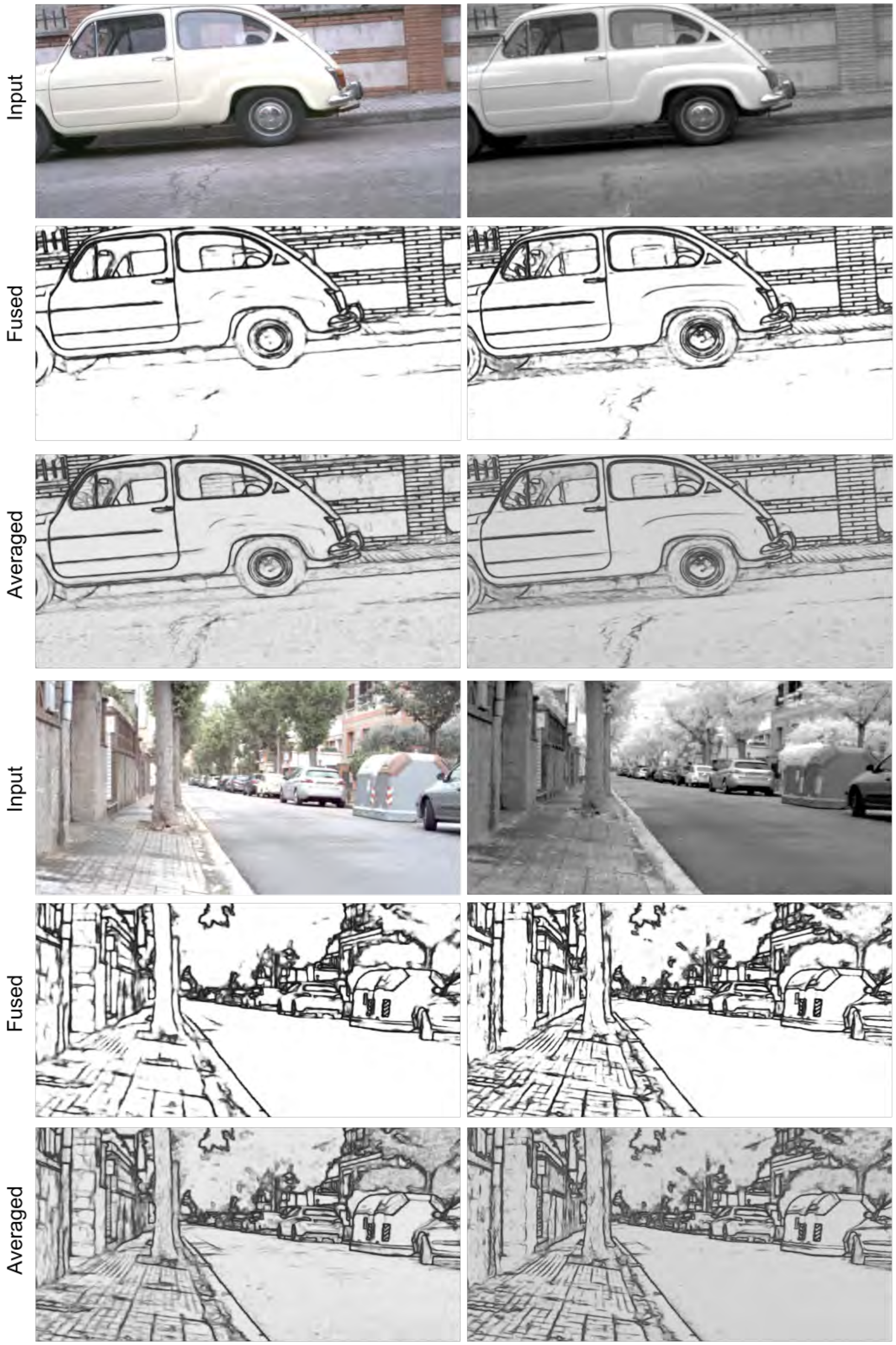


Fig. 5. Illustration of edges extracted in two pairs of images. (Input row) RGB and NIR images. (Fused row) Edge maps corresponding to fused predictions ($\hat{y}_i^{(fso)}$) by HED ($left-side$) and MSI-HED ($right-side$). (Averaged row) Edge maps corresponding to averaged predictions ($\hat{y}_i^{(av)}$) by HED ($left-side$) and MSI-HED ($right - side$).

V. Conclusions and Future Works

A Multi-Spectral Deep Learning model based on [15] has been proposed. Quantitative results show that the proposed approach reaches the same performance as the one obtained with HED model (just training the network with images from the visible spectrum). It should be mentioned that these evaluations have been performed using as ground truth a dataset annotated in the visible spectrum (users were given the images from the visible spectrum to draw objects' edges). In spite of the fact both, MSI-HED and HED, result in similar quantitative values, qualitatively MSI-HED gives better results since it was able to extract more edges from the given scenes (i.e., small details are correctly recovered). Hence, it is expected, that training the MSI-HED model with edges from multispectral images will help to obtain better quantitative results. As a future work we are going to explore the possibility of training the MSI-HED model with ground truth from multi-spectral images as well as from images resulting from the fusion of different channels. Furthermore, different models will be evaluated to obtain the best architecture for the MSI edge detection.

Acknowledgment

We are grateful for the free usage of *labelbox.com* for the dataset annotation.

References

[1] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1270–1281, 2008.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[3] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[4] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] R. Pourreza, Y. Zhuge, H. Ning, and R. Miller, "Brain tumor segmentation in mri scans using deeply-supervised neural networks," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 320–331.

[6] D. Ziou, S. Tabbone *et al.*, "Edge detection techniques-an overview," *Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 8, pp. 537–559, 1998.

[7] G. Papari and N. Petkov, "Edge and line oriented contour detection: State of the art," *Image and Vision Computing*, vol. 29, no. 2-3, pp. 79–103, 2011.

[8] X.-Y. Gong, H. Su, D. Xu, Z.-T. Zhang, F. Shen, and H.-B. Yang, "An overview of contour detection approaches," *International Journal of Automation and Computing*, Jun 2018. [Online]. Available: https://doi.org/10.1007/s11633-018-1117-z

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[11] X. Soria, A. Sappa, and R. Hammoud, "Wide-band color imagery restoration for rgb-nir single sensor images." *Sensors (Basel, Switzerland)*, vol. 18, no. 7, p. 2059, 2018.

[12] S. Matsui, T. Okabe, M. Shimano, and Y. Sato, "Image enhancement of low-light scenes with near-infrared flash images," *Information and Media Technologies*, vol. 6, no. 1, pp. 202–210, 2011.

[13] L. Schaul, C. Fredembach, and S. Süsstrunk, "Color image dehazing using the near-infrared," in *Proc. IEEE International Conference on Image Processing (ICIP)*, no. LCAV-CONF-2009-026, 2009.

[14] Z. Chen, X. Wang, and R. Liang, "Rgb-nir multispectral camera," *Optics express*, vol. 22, no. 5, pp. 4985–4994, 2014.

[15] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, no. 1-3, pp. 3–18, 2017.

[16] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1841–1848.

[17] ——, "Fast edge eetection using structured forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1558–1570, 2015.

[18] G. Bertasius, J. Shi, and L. Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4380–4389.

[19] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3982–3991.

[20] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2010.161

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5872–5881.

[23] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe, "Learning deep structured multi-scale features using attention-gated crfs for contour prediction," in *Advances in Neural Information Processing Systems*, 2017, pp. 3961–3970.

[24] J. Winn, "Causality with gates," in *Artificial Intelligence and Statistics*, 2012, pp. 1314–1322.

[25] Z. Zeng, Y. K. Yu, and K. H. Wong, "Adversarial network for edge detection," in *International Conference on Informatics, Electronics & Vision (ICIEV)*, 2018.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5967–5976.

[28] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[29] G. Evangelidis, "Iat: A matlab toolbox for image alignment," 2013.