Cross-Spectral Image Patch Similarity using Convolutional Neural Network

Patricia L. Suárez[1], Angel D. Sappa[1,2], Boris X. Vintimilla[1]

[1]Escuela Superior Politcnica del Litoral, ESPOL,
Facultad de Ingeniería en Electricidad y Computación, CIDIS,
Campus Gustavo Galindo, 09-01-5863, Guayaquil, Ecuador

[2]Computer Vision Center, Edifici O, Campus UAB,
08193, Bellaterra, Barcelona, Spain

{plsuarez, asappa, boris.vintimilla}@espol.edu.ec

Abstract—**The ability to compare image regions (patches) has been the basis of many approaches to core computer vision problems, including object, texture and scene categorization. Hence, developing representations for image patches have been of interest in several works. The current work focuses on learning similarity between cross-spectral image patches with a 2 channel convolutional neural network (CNN) model. The proposed approach is an adaptation of a previous work, trying to obtain similar results than the state of the art but with a low-cost hardware. Hence, obtained results are compared with both classical approaches, showing improvements, and a state of the art CNN based approach.**

I. Introduction

Computer vision tackles problems related with object detection and recognition, texture classification, action recognition, segmentation, tracking, data retrieval, image alignment, etc. There are several techniques for performing these tasks, and usually based on representing an image using some global or local image properties, and comparing them using some similarity measure. Learning visual similarities has been presented recently with success working on images in the mono-spectral domain [1]. Images are often represented by compact region descriptors with interest points. The main idea is to extract all possible patches no matter overlapping, these patches are usually very small comparing with the original size of the image, with them we proceed with their processing to exploit interrelation between them [2].

During last decades different approaches have been proposed to generate feature descriptors (e.g., SIFT [3], SURF [4], KAZE [5], just to mention a few), those had a great impact on computer vision area. Many researchers have been working with image patches for processing spatial like prior for road detection and urban understanding, which can be used for image labeling [6]; other approaches have been proposed based on image-adaptive wavelet transform, which are tailored to sparsely represent a given image, to form a multiscale sparsifying global transform for the image in question [6]. There are some others methods based on image patch processing like a fast patch dictionary for image recovery and sparsity-based image denoising via dictionary learning and structural clustering [7], non-local means methods for image denoising [8] and image processing using smooth ordering of its patches [9].

All the approaches mentioned above have been initially proposed for working with patches obtained from similar images; generally images of the visible spectrum, in other words monospectral approaches. These days, the coexistence of cameras working at different spectral bands has increased considerably, mainly based on recent advances in imaging devices as well as the reduction on the prices of such a technology. This cross-spectral information helps to solve classical problems in poor lighting conditions or enhance visible spectrum images with information from other spectral bands (e.g., filtering [10], enhancement [11]). The current work is focussed on the usage of images from the visible spectrum (RGB images) together with images from the near infrared spectra (NIR images).

The usage of cross-spectral information, although interesting and appealing, implies new challenging and difficult problems that need to be tackled and efficiently solved. For instance, different works have been recently proposed for describing and matching feature points in cross-spectral domains based on classical approaches (e.g., [12], [13], [14], [15], to mention a few). Unfortunately, due to the natural difference between images acquired from different spectra, the obtained performance is far away from the one obtained in monospectral scenarios.

In order to overcome the aforementioned poor performance some recent approaches, based on the usage of Convolutional Neural Networks (CNNs), have been proposed with interesting results. Some times such good results are obtained using expensive dedicated GPUs. In the current work we propose to use the CNN architecture presented in [16], but modifying the number of layers and reducing both the size of patches and convolution kernels, in order to use it in a low-cost hardware (about ten times cheaper than the one used in [16]). This network consists of a unified architecture that jointly learns a 2 channel deep neural network for cross-spectral patch representation (see Fig. 1). As mentioned above, the

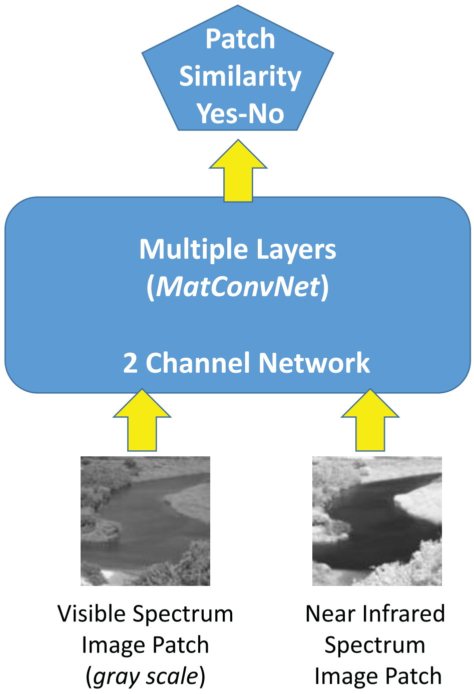

Fig. 1. 2 channel network model implemented on the current work to obtain automatic cross-spectral matchings.

main contribution of current work is to reach a performance similar to [16] but with a low-cost hardware. The rest of the paper is organized as follows. Section II describes the most recent work on CNN image patch similarity learning. Section III presents the adapted CNN architecture detailing the design and training with cross-spectral datasets. Section IV depicts the experimental results and finally, conclusion are presented in section V.

II. Related Work

Image patch's relevant representations and the corresponding similarity measures can vary significantly. Images are often represented using dense pixel-based properties or by compact region descriptors (features) often used with interest point detectors. Dense properties include raw pixel intensity or color values from image patches. There are other techniques like common compact region descriptors that include distribution based descriptors (e.g., SIFT, SURF), differential descriptors (e.g., local derivatives), shape-based descriptors using extracted edges (e.g., shape context) and others. For a comprehensive comparison of local descriptors for image matching see [17].

Although these representations and their corresponding similarity measures may vary significantly, they all share the same basic assumption, that there exists a common underlying visual property (e.g., pixels colors, intensities, edges, gradients or other filter responses), which is shared by the two image patches, and can therefore be extracted and compared across images/sequences see [18]. The comparison between the representations, using the aforementioned similarity measures, can be embedded into learning methods, which are able to find the non-linear relationship between the representations. These learning based approaches generally rely on some easy-to-compute distance metric (e.g., Hinge distance) that some times correlates with the semantic similarity. Different learning approaches have been proposed in the literature. Recently, Convolutional Neural Network based learning techniques are among the best option producing appealing results (e.g., see [19].

Convolutional Neural Networks (CNNs), are becoming the dominant tool to tackle most of computer vision problems. CNNs are a specific type of neural network widely used in deep learning algorithms. Their convolutional kernel based philosophy makes them easy to apply in the computer vision domain for classical problems. One of them is the extraction of interesting parts of an image, obtaining feature vectors needed for task like object detection, classification, segmentation, etc. Those techniques do not ignore the structure and compositional nature of images, so they can learn to extract features directly from raw images, eliminating the need of manual feature extraction.

Inspired on the network structure presented for stereo matching in [20], the authors of [16] proposes a novel approach for learning cross-spectral similarity measures. This approach avoids defining a hand-made descriptor being the CNN responsible for jointly learning the representation and the measurement. This approach will be referred in current work as 2 channel network (2ChNet). Patch matching has also been addressed in [21]; in this case, the authors propose a generalization of the siamese networks in order to speed up the matching process. The architecture of the network consists of two parts, firstly a network is used for describing the patches, then another network is proposed for the matching (metric network). Following the siamese architecture, [22] proposes to train a siamese network that compares the similarity between image patches just using L2 distance. This simple matching speeds up the whole process since it is possible to use fast approximate nearest neighbours algorithms to find the correspondences and thus improve the overall matching runtime. A comparative study between 2 channel and siamese architecture has been performed in [16]. It is shown that the 2ChNet has a considerably better performance in the cross-spectral domain.

Based on the results mentioned above, in the current work a 2ChNet is adapted to be used with a low-cost hardware. The adaptation consists of reducing the number of layers in the network. The objective is to obtain a similar performance to the one obtained by [16], which is considerably better to those approaches based on hand-made descriptors (e.g., [13], [14], [15]).

III. Network Architecture

As mentioned above the current work is focused on finding correspondences between images from visible and near infrared spectra. The network architecture selected to find

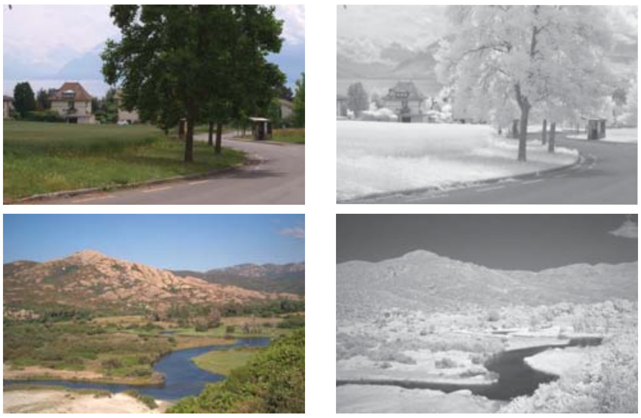

Fig. 2. Cross-spectral pairs of images obtained from [2]: (a) visible images; (b) NIR images.

correspondences between patches from these images is similar to the one presented in [16], the 2 channel network (2ChNet) model. Figure 1 shows an illustration of the model. The adaptation proposed to the 2ChNet architecture is presented in Fig. 3; this architecture contains less layers than [16] in order to train it with a low-cost hardware. As can be appreciated in this illustration, this architecture takes as input a pair of patches (one from each spectra), and then a series of convolution, ReLu and max-pooling layers are applied till the final linear layer that works as the metric network. Note that the patch from the visible spectrum (RGB image) is converted to gray scale.

The network learns the similarity by combining information from both spectra and jointly processing them through the different layers. This way of processing the information has been shown as the best solution in cross-spectral domains [16]. The training process do not rely on labels assigned to each patch, but rather on pairs of patches of different spectra with similarity or non-similarity. During the training we minimize the loss with a fully connected layer, before the loss linear layer.

The 2ChNet architecture takes as an input two patches, one from each spectra. The size of the patches is 64×64; the model consists of different layers, like convolution, ReLU, max-pooling and a final linear layer that computes the loss of each iteration of the learning process. This last layer acts as a metric, which permits to determine whether the pair of patches have or not correspondence. Figure 3 shows the adapted architecture of the model.

The network architecture described above was trained in a supervised way; each layer convolves the output of the previous one, with a filter learned at each operation. Some layers permit to change the spatial size of the output, obtaining the maximum or an average value of a previous convolution layers, or the corresponding activation function . The last layers are fully connected and multiply the output obtained with a matrix of learned parameters followed again by a non-linear activation function (ReLU). We use a margin criterion based on a *hinge loss* and squared *l2-norm* regularization term as in [1]:

$$\min_{w} \frac{\lambda}{2}||w||_2 + \sum_{i=1}^{N} max(0, 1 - y_i o_i^{net}), \qquad (1)$$

where w is the network weight, o_i^{net} is the training output for the $i-th$ training sample iteration and y is the $i-th$ training label; the value domain is -1,1 for a false and true similarity respectively, and λ denote the weight decay.

IV. Experiments Results

In order to test the proposed approach the cross-spectral data set from from [2] has been used (a couple of pairs are presented in Fig. 2). This dataset consists of 477 registered images categorized in 9 groups captured in RGB (Visible Spectrum) and NIR (Near Infrared). In order to compare with the previous approach [16] just images from the category "Country" have been used for training (150 pairs of images randomly selected). These images are the most affected in conditions of varying lighting and textures, which directly affects the variability and complexity of the detection of the feature points and therefore are the most challenging scenarios for the training process. Feature points have been obtained from SIFT, applied over the visible spectrum images. Patches of 64×64 pixels have been generated centered on those points. Then, points placed at the same position than those obtained by SIFT algorithm are placed in the NIR images and the corresponding patches with the same size extracted. With this process 150.000 patches have been generated from "Country" category (dataset with correctly matched pairs); the same amount of patches have been generated for the false pair dataset.

The model is trained using Stochastic Gradient Descent with a weight decay (λ) of 0.0007, a learning rate of 0.05, a momentum of 0.9 and batches size of 80 samples. All input patches were normalized by its intensity mean, previous to normalization the values of intensities must be in the $\{0,1\}$ range. (80%) of the data set generated as mentioned above has been used for training, while 20% used for validation. We use MatconvNet toolbox for Matlab that implements Convolutional Neural Networks [23]. The 2ChNet model was trained during 6 days, on a 3.2 eight core processor with 4Gb of memory with a NVIDIA GeForce GTX970 GPU.

Once the 2ChNet has been trained with images from the "Country" category it has been evaluated with other cross-spectral images from the "Country" category together with other categories. Thus, 300 pairs from each of the following categories have been selected: "Country", "Indoor", "Olbuilding" and "Urban" respectively. The results obtained from this evaluation were compared with those obtained with a classical feature descriptor (SIFT) to highlight the improvements in performance reached with the proposed approach. The FPR95% rate, which is the ratio between the number of negative coincidences wrongly categorized as positive (false positives) and

Descriptor-Network	Country	Indoor	Oldbuilding	Urban
SIFT [3]	46.6	12.4	21.3	13.27
2ch Network (from [16])	**0.23**	4.4	**2.3**	**1.58**
2ch Network (Proposed)	0.27	**3.3**	3.4	4.6

TABLE I
EVALUATIONS (FPR95%) ON VISIBLE-NIR PATCH DATASETS [2] FROM DIFFERENT CATEGORIES (THE SMALLER THE BETTER, BOLD FACES CORRESPOND TO THE BEST RESULTS IN THAT CATEGORY).

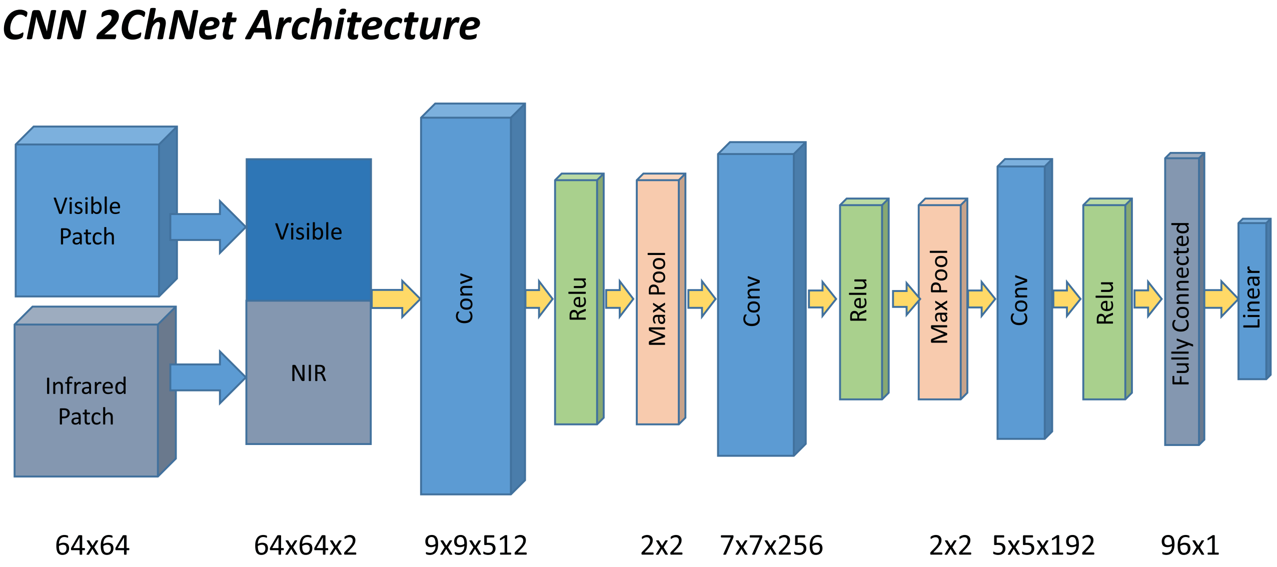

Fig. 3. Layer architecture of the 2ChNet adapted in the current work (note that both inputs are converted to gray level representations).

the total number of actual negative coincidences (regardless of classification), is used to measures the obtained results. Additionally, these values have been compared with the ones presented in [16]. It can be used to evaluate results from the same categories, Table 1 shows the obtained performances. As expected, it can be appreciated the large improvements reached with respect to SIFT. Additionally, it can be appreciated that in spite of the hardware limitations, the results are similar to the one presented in [16], actually, in one case the result is even better than the ones obtained in [16].

V. Conclusion

This paper tackles the challenging problem of cross-spectral image patch similarity, by adapting a state of the art architecture with a low-cost hardware. The results show that even with a low-cost hardware the obtained performance is quite similar to the state of the art, as well as it is shown that outperforms classical SIFT feature based descriptors. As a future work other strategies will be considered for improving results, but always keeping in mind the limitation of hardware.

Acknowledgment

This work has been partially supported by the ESPOL projects: ”Pattern recognition: case study on agriculture and aquaculture” (M1-DI-2015) and ”Integrated system for emergency management using sensor networks and reactive signaling” (G4-DI-2014); and by the Spanish Government under Project TIN2014-56919-C3-2-R.

References

[1] S. Zagoruyko and N. Komodakis, “Learning to compare image patches via convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[2] M. Brown and S. Süsstrunk, “Multi-spectral SIFT for scene category recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 177–184.

[3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014

[5] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, “Kaze features,” in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ser. ECCV’12, 2012, pp. 214–227.

[6] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, “Convolutional patch networks with spatial prior for road detection and urban scene understanding,” *arXiv preprint arXiv:1502.06344*, 2015.

[7] W. Dong, X. Li, L. Zhang, and G. Shi, “Sparsity-based image denoising via dictionary learning and structural clustering,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 457–464.

[8] S. Dang, Y. Zhang, and D. Gong, "A patch-based non-local means method for image denoising," in *International Conference on Intelligent Science and Intelligent Data Engineering*. Springer, 2012, pp. 582–589.

[9] I. Ram, M. Elad, and I. Cohen, "Image processing using smooth ordering of its patches," *IEEE transactions on image processing*, vol. 22, no. 7, pp. 2764–2774, 2013.

[10] H. Honda, R. Timofte, and L. Van Gool, "Make my day-high-fidelity color denoising with near-infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 82–90.

[11] X. Zhang, T. Sim, and X. Miao, "Enhancing photographs with near infra-red images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[12] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and Multi-spectral Registration for Natural Images," in *ECCV*, Zurich, Switzerland, Sep 2014, pp. 309–324.

[13] C. Aguilera, F. Barrera, F. Lumbreras, A. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, vol. 12, no. 9, pp. 12 661–72, Jan. 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/9/12661

[14] C. A. Aguilera, A. D. Sappa, and R. Toledo, "LGHD: A feature descriptor for matching across non-linear intensity variations," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 178–181.

[15] T. Mouats, N. Aouf, A. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *ITS*, vol. PP, no. 99, pp. 1–15, Sep 2014.

[16] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Jun 2016, p. 9.

[17] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[18] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[19] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[20] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1592–1599.

[21] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *CVPR*, 2015.

[22] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative Learning of Deep Convolutional Feature Point Descriptors," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Dec 2015.

[23] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.