Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

Article

Cross-Spectral Local Descriptors via Quadruplet Network

Cristhian A. Aguilera [1,2,*], Angel D. Sappa [1,3], Cristhian Aguilera [4] and Ricardo Toledo [1,2]

1 Computer Vision Center, Edifici O, Campus UAB, Bellaterra 08193, Barcelona, Spain; asappa@cvc.uab.es (A.D.S.); ricard@cvc.uab.es (R.T.)
2 Computer Science Department, Universitat Autònoma de Barcelona, Campus UAB, Bellaterra 08193, Barcelona, Spain
3 Facultad de Ingeniería en Electricidad y Computación, CIDIS, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo, Km 30.5 vía Perimetral, Guayaquil 09-01-5863, Ecuador
4 DIEE, University of Bío-Bío, Concepción 4051381, Concepción, Chile; cristhia@ubiobio.cl
* Correspondence: caguilera@cvc.uab.es; Tel.: +56-9-30687318

Abstract: This paper presents a novel CNN-based architecture, referred to as Q-Net, to learn local feature descriptors that are useful for matching image patches from two different spectral bands. Given correctly matched and non-matching cross-spectral image pairs, a quadruplet network is trained to map input image patches to a common Euclidean space, regardless of the input spectral band. Our approach is inspired by the recent success of triplet networks in the visible spectrum, but adapted for cross-spectral scenarios, where for each matching pair there are always two possible non-matching patches; one for each spectrum. Experimental evaluations on a public cross-spectral VIS-NIR dataset shows that the proposed approach improves the state-of-the-art. Moreover, the proposed technique can also be used in mono-spectral settings, obtaining a similar performance to triplet network descriptors, but requiring less training data.

Keywords: cross-spectral; descriptor; infrared; CNN

1. Introduction

Over the last few years the number of consumer computer vision applications has increased dramatically. Today, computer vision solutions can be found in video game consoles [1], smartphone applications [2], and even in sports—just to name a few. Furthermore, safety-critical computer vision applications, such as autonomous driving systems, are not longer science fiction.

Ideally, we require the performance of those applications, particularly those that are safety-critical to remain constant under any external environment factor, such as changes in illumination or weather conditions. However, this is not always possible or very difficult to obtain solely using visible imagery, due to the inherent limitations of the images from that spectral band. For that reason, the use of images from different or multiple spectral bands is becoming more appealing. For example, the Microsoft Kinect 2 uses a near-infrared-camera to improve the detection performance in low light or no light conditions.

In this work we are particularly interested in cross-spectral applications, i.e., computer-vision applications that make use of two different spectral bands to provided a richer representation of a scene, that otherwise could not be obtained from just one spectral band. We present a novel CNN-based architecture, referred to as Q-Net, to learn local feature descriptors that are useful for matching image patches from two different spectral bands. Figure 1 shows an illustration of our proposal. The training network consists of four copies of the same CNN, i.e., weights are shared, which accepts as input two different cross-spectral image matching pairs[1]. In the forward pass, the

[1] A matching pair consists of two image patches that show the same scene, regardless of the spectral band of the patches. On the contrary, a non-matching pair consists of two image patches that show two different scene of the world.

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

network computes several L_2 distances between the outputs of each CNN, obtaining the matching pair with the biggest L_2 distance and the non-matching pair with the smallest L_2 distance that will be later on used during the backpropagation step. This can be seen as always using the hardest cases of matching and non-matching pairs at each iteration. At testing, our network can be used as drop-in replacements for hand-made feature descriptors such as SIFT in various cross-spectral tasks such as stereo-vision, object detection and image registration. It is important to notice that during testing it is just necessary one of the four CNNs that were used during training. This CNN will act as a feature descriptor.

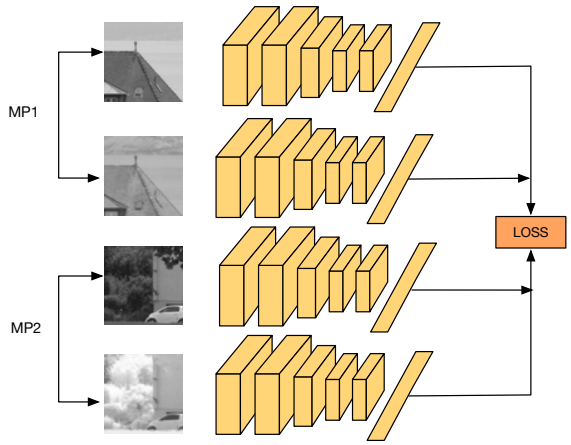


Figure 1. The proposed network architecture. It consists of four copies of the same CNN that accepts as input two different cross-spectral correctly matched image pairs (MP1 and MP2). The network computes the loss based on multiples L_2 distance comparisons between the output of each CNN, looking for the matching pair with biggest L_2 distance and the non-matching pair with the smallest L_2 distance. Both cases are then used for backpropagation of the network. This can be seen as positive and negative mining.

Our work is based on the recent success of the triplet network presented in [3], named PN-Net, but adapted to work with cross-spectral image pairs, where for each matching pair, there are two possible non-matching patches; one for each spectrum. Results show that our technique is useful for learning cross-spectral feature descriptors that can be used as drop-in replacements of SIFT-like features descriptors. Moreover, results also show that our network can be useful for learning local feature descriptors in the visible domain, with similar performance to PN-Net but requiring less training data.

In this article, we make the following contributions:

- We propose and evaluate three ways of using triplets for learning cross-spectral descriptors. Triplet networks were originally designed to work on visible imagery, so the performance on cross-spectral images is unknown.
- We propose a new training CNN-based architecture that outperforms the state-of-the-art in a public VIS-NIR cross-spectral image pair dataset. Additionally, our experiments show that our network is also useful for learning local feature descriptors in the visible domain.
- Fully trained networks and source code is publicly available at http://github.com/ngunsu/qnet.

The rest of the paper is organized as follows. In section 2 we give a short description of near-infrared images and the VIS-NIR cross-spectral dataset used to train and evaluate our work; differences between visible and infrared images are highlighted. Additionally, section 2 also present an overview of hand-made cross-spectral descriptors and CNN-based visible spectrum descriptors. The PN-Net triplet network is described in section 3, introducing the motivations behind the

(a) (b) (c)

Figure 2. VIS-NIR cross-spectral image pairs; top images are from the visible spectrum and bottom images from the near-infrared spectrum

proposed technique, which are presented in section 4. Finally, in section 5 we show the validity of our proposal through several experiments, ending with the conclusions in section 6.

2. Background and related work

This section briefly introduces the near-infrared spectral band and highlights similarities and differences between images from this spectrum with respect to the visible spectrum. Additionally, the VIS-NIR cross-spectral image dataset used as a case study through the different sections of this manuscript is also presented. Finally, the most important methods proposed in the literature to describe images from two different spectral bands are reviewed, together with current CNN-based descriptor approaches used to describe images in the visible spectrum.

2.1. Near-infrared band

The near-infrared band (NIR: 0.7-1.4 μm) is one of the five sub-bands of the infrared spectrum (0.7-1000 μm). It is the closest infrared sub-band to the visible spectrum and images from both spectral bands share several visual similarities; in Fig. 2 three pairs of VIS-NIR images are presented. It can be appreciated that images from both spectra are visually similar but with some notable differences. For example, red visible regions disappear in NIR images (see Fig. 2(a)), the direction of the gradients can change (as in Fig. 2(b)), and NIR images are more robust to different illumination settings (as in Fig. 2(c)).

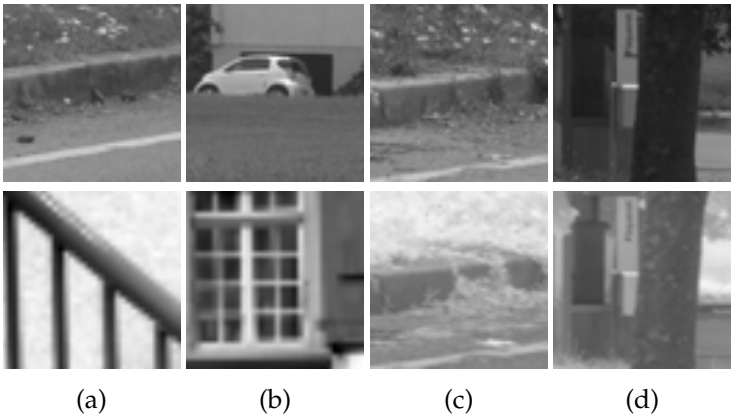
Recent advances in technology have made infrared imaging devices affordable for classical computer vision problems, from face recognition ([4]) to medical imaging ([5]). In some of these cases, infrared images are not used alone but in conjunction with images from other spectral bands (e.g., visible spectra). In these cases, infrared images need to be used in a framework that allows them to be handled in an efficient way in terms of the heterogeneity of the information ([6]), which is the main challenge to be solved and the motivation for current work.

2.2. Dataset

The dataset used in [7] has been considered in the current work to train and validate the proposed network. This dataset has been obtained from [8], and consists of more than 1 million VIS-NIR cross-spectral image pairs divided into nine different categories. Table 1 shows the distributions of patches for the different categories. The size of each image patch is 64x64 and visible images are presented in grayscale format. Finally, the number of matching and non-matching pairs is the same, so half of the samples correspond to correctly matched cross-spectral pairs and the

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

other half to non-matching pairs. Figure 3 shows four samples of cross-spectral image pairs from the dataset.



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

Figure 3. Image patches from the VIS-NIR training set. First row corresponds to grayscale images from the visible spectrum; second row to NIR images. (a) and (b): non-matching pairs; (c) and (d): correctly matched pairs.

Table 1. Shows the number of cross-spectral image pairs per category on the VIS-NIR patch dataset used to train and evaluate our work.

Category	# Cross-spectral pairs
country	277504
field	240896
forest	376832
indoor	60672
mountain	151296
oldbuilding	101376
street	164608
urban	147712
water	143104

2.3. Cross-spectral descriptors

The description of two images from two different spectral bands in the same way is a challenging task that cannot always be solved with classical feature descriptors such as SIFT ([9]) or SURF ([10]), due to the non-linear intensity variations that may exist between images from two different spectral bands. Early efforts focused on modifying gradient-based descriptors to work between $[0, \pi]$ instead of $[0, 2\pi]$ to reduce the effect of changes in the gradient direction between images from two different spectral bands. For example, [11] and [12] applied this strategy to SIFT and HOG respectively, the first to match VIS-NIR local features and the second to compute the stereo disparity between a VIS and a thermal infrared camera. Although this strategy is simple, it improves the performance of those algorithms in cross-spectral scenarios.

Other works are based on the observations of [13]. In this study, concerning the joint statistics of visible and thermal images, the authors found a strong correlation between object boundaries of images from both spectra, i.e., texture information is lost and edge information remains similar between images from the different spectral bands. [14] describe cross-spectral image patches using a local version of the global EHD descriptors, focusing more on the information provided by the edges

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

rather than in image texture. In a similar way, [15] and [16] compute cross-spectral features using the EHD algorithm over the image patch response to different Log-Gabor filters.

In a more recent work, [7] tested different CNN-based networks to measure the similarity between images from the VIS-NIR and the VIS-LWIR spectra. In their experiments, they showed that CNN-based networks can outperform the state-of-the-art in terms of matching performance. However, with regard to speed, their networks are much slower than classical solutions, problem that we address in the current work, training a cross-spectral descriptor that can be used as a replacement of SIFT and many other L_2-based descriptors.

2.4. CNN-based feature descriptor approaches

During last decades carefully *hand-made* feature descriptors, such as SIFT or SURF, have been popular in the computer vision community. However, in the last few years, such approaches have been started to be outperformed by CNN-based solutions in different feature descriptors benchmarks (e.g., [3,17,18]). [17] propose a max-margin criterion over a siamese network to compute the similarity between two different image patches. [18] follows a similar approach, but instead of using a metric network, it directly minimizes the L_2 distance between the descriptor of two images in the loss function, making their trained descriptor a drop-in replacement to SIFT-like descriptors. More importantly, it notices that after a few training epochs, most of the non-matching samples used to train the network were not giving new information, making it necessary to use mining strategies to improve the performance of the networks. In the same way, [3] propose a triplet network to mine negative samples with each input triplet, improving the performance of siamese networks.

The current works is strongly based on the triplet network proposed by [3], but adapted to be used in cross-spectral image pairs. We use quadruplet instead of triplet and we do not only mine non-matching samples but also correctly matched samples. In section 3 and 4 both architectures are detailed.

Notice that in this review, we only include CNN-based feature descriptors that can be used as a replacement for common *hand-made* descriptors, since they are useful for real-time applications. Other solutions have been proposed skipping the process of description and matching, doing everything with a CNN and trading matching performance for speed. For example, the 2ch network from [17], MatchNet from [19] and the siamese network from [20].

3. PN-Net (Triplet network)

In this section we describe the architecture of PN-Net ([3]); a state-of-the-art triplet network for learning local features descriptors. Although PN-Net was not intended to be used as a network for learning cross-spectral descriptors, we propose three ways to use PN-Net in such settings. In summary, this section objective is twofold:

- As previously stated, our network is similar to the triplet network but specifically designed to learn cross-spectral local feature descriptors. A brief description of this network will help to set the basis of our proposal in Section 4.
- We explain the motivation behind our proposal through several experiments. After training PN-Net to learn cross-spectral feature descriptors, we discovered that the network performance improved when we randomly alternated between non-matching patches from both spectra.

3.1. PN-Net Architecture

Figure 4 shows the training architecture of PN-Net. The network has three inputs, where each input corresponds to a different image patch. Formally, the input is a tuple $T = \{w, x, y\}$, where w and x are two matching image patches and y is a non-matching image patch to w and x. Each one of these patches will feed one of the three CNN *towers* that the network has; CNN 1, CNN 2 and CNN 3. The three CNN *towers* of the network share the same parameters during the entire training stage.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2017 doi:10.20944/preprints201703.0061.v1

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

Finally, the output of each *tower* will be a descriptor D of configurable size, that describes each input patch.

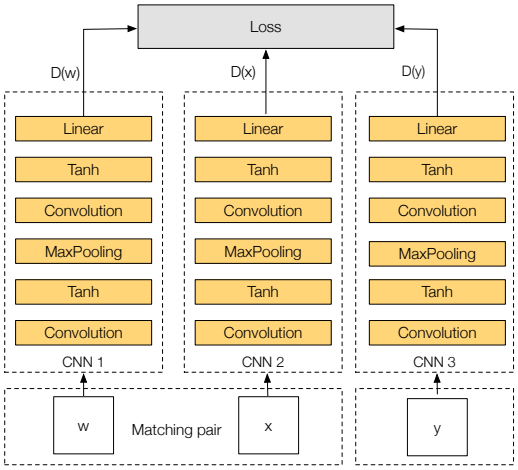


Figure 4. PN-Net training triplet architecture

3.2. PN-Net Loss

The loss function is described as follows:

$$P_m = \frac{e^{\Delta^+}}{e^{min(\Delta_1^-, \Delta_2^-)} + e^{\Delta^+}} \tag{1}$$

$$P_{nm} = \frac{e^{min(\Delta_1^-, \Delta_2^-)}}{e^{min(\Delta_1^-, \Delta_2^-)} + e^{\Delta^+}} \tag{2}$$

$$Loss(T_i) = P_m^2 + (P_{nm} - 1)^2 \tag{3}$$

where, Δ^+ corresponds to the L_2 distance between the descriptors of the matching pair w and x, $||D(w), D(x)||_2$; Δ_1^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and y, $||D(w), D(y)||_2$; and Δ_2^- corresponds to the L_2 distance between the descriptors of the second non-matching pair x and y, $||D(x), D(y)||_2$.

In essence, the objective of the loss function is to penalize small L_2 distances between non-matching pairs, and large L_2 distances between matching pairs. Ideally, we want P_m to be equal to zero and P_{nm} to be equal to one, i.e, $\Delta^+ << min(\Delta_1^-, \Delta_2^-)$. Computing the minimum L_2 distances between the non-matching pairs is a type of mining strategy, were the network always performs backpropagation using the hardest non-matching sample of each triplet T, i.e, the non-matching sample with the smallest L_2 distance. The mining strategy is used to avoid the problems described in [18] and mentioned in section 2. Finally, the MSE is used to penalize values of P_m different of zero and values of P_{nm} different of one.

3.3. Cross-spectral PN-Net

One key difference between mono-spectral and cross-spectral image pairs is that for each cross-spectral matching pair we have two non-matching possible image patches; one for each spectrum. So the question is, which image patch we use as y?. We propose three simple and naive solutions: i) y is an RGB non-matching image, ii) y is an NIR non-matching image and iii) y is randomly chosen between RGB and NIR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2017 doi:10.20944/preprints201703.0061.v1

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

Table 2. Average FPR95 for each category

Train seq.	PN-Net Gray	PN-Net NIR	PN-Net Random
Country	11.79	11.63	**10.65**
Field	17.84	16.56	**16.10**
Forest	36.00	32.47	**32.19**
Indoor	48.21	47.26	**46.48**
Mountain	29.35	26.29	**25.67**
Oldbuilding	29.22	**27.25**	27.69
Street	18.23	**16.71**	16.73
Urban	**32.78**	36.61	33.35
Water	18.16	17.76	**15.84**
Average	26.84	25.84	**25.08**

We test each one of the aforementioned solutions in the dataset used in [7] and presented in section 2 . We train each network nine times, once per category and tested on the other eight categories. In Table 2, we present our results in terms of the false positive rate at 95% Recall (FPR95). In essence, we evaluate how well the descriptors can distinguish between correct and incorrect matches. Results show that randomly using y between an NIR and RGB patch was better than the other two solutions. This give us a hint, that using images from both spectra to mine the network is better than using just from one, which we expected assuming that a balanced number of non-matching images from both spectra will help to produce better results.

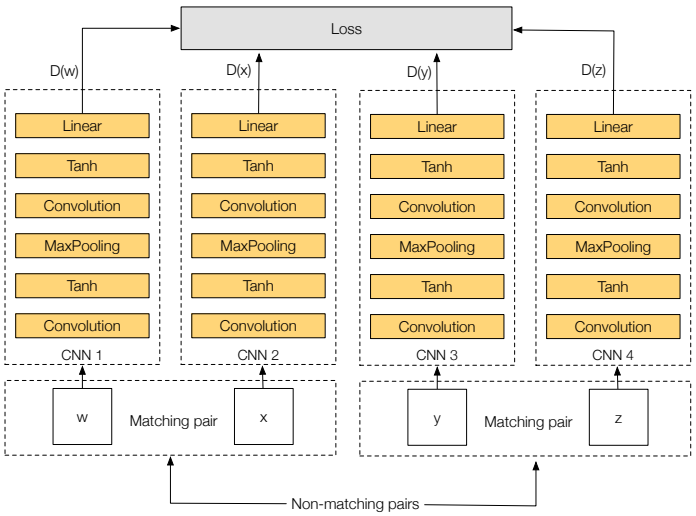


Figure 5. Q-Net training quadruplet architecture

4. Q-Net

The motivations behind our quadruplet network are straightforward. As stated before, for each cross-spectral matching pair we have at least two non-matching patches from another spatial location, each one from one of the spectra to be trained. Similar to triplets, we propose Q-Net, a quadruplet network for learning cross-spectral feature descriptors.

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

4.1. Q-Net Architecture

The architecture of Q-Net is similar to PN-Net, but using four copies of the same network instead of three (see Figure 5). The input is a tuple Q, with four different input patches $Q = \{w, x, y, z\}$, that is formed by two different cross-spectral matching pairs: (w, x), and (y, z), allowing the network to mine not just non-matching cross-spectral image pairs at each iteration, but also cross-spectral correctly matched pairs.

4.2. Q-Net Loss

Q-Net loss function extends the mining strategy from PN-Net presented in section 3.2. Specifically, we add two more distance comparisons to P_{nm}, making the loss suitable for cross-spectral scenarios, and we extend the mining strategy from the non-matching pairs to the correctly matched pairs. At each training step, the network uses the matching pair with larger L_2 distance and the non-matching pair with the smallest L_2 distance. The loss function is described as follows:

$$P_m = \frac{e^{max(\Delta_1^+, \Delta_2^+)}}{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)} + e^{max(\Delta_1^+, \Delta_2^+)}} \tag{4}$$

$$P_{nm} = \frac{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)}}{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)} + e^{max(\Delta_1^+, \Delta_2^+)}} \tag{5}$$

$$Loss(Q_i) = P_m^2 + (P_{nm} - 1)^2 \tag{6}$$

where, Δ_1^+ corresponds to the L_2 distance between the descriptors of the matching pair w and x, $||D(w), D(x)||_2$; Δ_2^+ corresponds to the L_2 distance between the descriptors of the matching pair y and z, $||D(y), D(z)||_2$; Δ_1^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and y, $||D(w), D(y)||_2$; Δ_2^- corresponds to the L_2 distance between the descriptors of the non-matching pair x and y, $||D(x), D(y)||_2$; Δ_3^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and z, $||D(w), D(z)||_2$; and Δ_4^- corresponds to the L_2 distance between the descriptors of the non-matching pair x and z, $||D(x), D(z)||_2$.

The proposed loss function takes into account all the possible non-matching combinations. For example, if we want to train a network to learn similarities between the VIS and the NIR spectral bands, P_{nm} will compare two VIS-NIR non-matching pairs, one VIS-VIS non-matching pair and one NIR-NIR non-matching pair; instead of using a random function as we did with PN-Net. Moreover, since we are trying to learn a common representation between the NIR and the VIS, comparing VIS-VIS and NIR-NIR cases helps the network to have more training examples. Since it is necessary to have two cross-spectral matching pairs to compute P_{nm}, it was natural to extend the mining strategy to P_m, obtaining at each step the cross-spectral matching pair with the larger L_2 distance.

Our method allows to learn cross-spectral distances, mining positives and negatives samples at the same time. This approach can also be used in monospectral scenarios, providing a more efficient mining strategy than previous works. Results that support our claim are presented in the next section. More importantly, our method can be extended to other cross-spectral or cross-modality scenarios. Even more, it can be extended to other applications such as heterogeneous face recognition, where it is necessary to learn distance metrics between faces from different spectral bands.

5. Experimental evaluation

In this section we evaluate the performance of the proposed Q-Net on 1) the VIS-NIR dataset introduced in section 2.2, and 2) the VIS standard benchmark for feature descriptors from [21]. The performance, in both cases, is measured using the FPR95 as in Section 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2017 doi:10.20944/preprints201703.0061.v1

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

Table 3. Q-Net layer descriptions

Layer	Description	Kernel	Output Dim
1	Convolution	7x7	32x26x26
2	Tanh	-	32x26x26
3	MaxPooling	2x2	32x13x13
4	Convolution	6x6	64x8x8
5	Tanh	-	64x8x8
6	Linear	-	256

5.1. VIS-NIR scene dataset

In this section we evaluate the performance of our network on the VIS-NIR dataset presented in Section 2. As in [7], we train on the country sequence and test in the remaining eight categories.

Training: Q-Net and PN-Net networks were trained using Stochastic Gradient Descent (SGD) with a learning rate of 1.1, weight decay of 0.0001, batch size of 128, momentum of 0.9 and learning rate decay of 0.000001. Trained data was shuffled at the beginning of each epoch and each input patch was normalized to its mean intensity. The trained data was split into two, where 95% of the data was used as training data and 5% as validation. Training was performed with and without data augmentation (DA), were the augmented data was obtained by flipping the images vertically and horizontally, and rotating the images by 90, 180 and 270 degrees. Each network was trained ten times to account for randomization effects in the initialization. Lastly, we used a grid search strategy to find the best parameters.

Model details: Model details are described in Table 3. The layers and parameters are the same from [3], which after several experimental results showed to be suitable for describing cross-spectral patches. Notice that for feature description shallow models are suitable, since lower layers are more general than the upper ones.

Software and hardware: All the code was implemented using the Torch framework ([22]). The GPU consisted of an NVIDIA Titan X and the network was trained in between five and ten hours when we used data augmentation.

Results are shown in Table 4. Firstly, we evaluated EHD ([14]) and LGHD ([16]), two *hand-made* descriptors that were used as a baseline in terms of matching performance. The performance of LGHD is under 10% and can be considered as state-of-art results—before the current work. Secondly, we test a siamese L_2 network based on the work of [17] that performs better than EHD, but worst than the state-of-art. Thirdly, PN-Net and its variant were tested, not being able to surpass the performance of LGHD without using data augmentation. On the other case, Q-Net showed to be better than the state-of-art even without data augmentation, showing the importance of mining on the non-matching and matching samples in cross-spectral scenarios. Additionally, we tested our model increasing the training data using the previously detailed data augmentation technique, improving the state-of-the-art by a 2.91%. For a more detailed comparison of the different feature descriptors evaluated in the current work, we provide the corresponding ROC curves in Fig. 7.

In addition, we tested the performance when different descriptor sizes were used. Fig. 6 shows the results of our experiment. From the figure we can see that there is a gain in increasing the descriptor size until 256. Descriptor sizes bigger than 256 did not perform better.

5.2. Multi-view stereo correspondence dataset

Although the proposed approach has been motivated to tackle the cross-spectral problem, in this section we evaluate the proposed architecture when a visible spectrum dataset is considered. This is intended to evaluate the validity of the proposed approach in classical scenarios.

For the evaluation we used the *multi-view stereo correspondence dataset* from [21], which is considered a standard benchmark for testing local feature descriptors in the visible domain (e.g.,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2017 doi:10.20944/preprints201703.0061.v1

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

Table 4. FPR95 performance on the VIS-NIR scene dataset. Each network, i.e., siamese-L2, PN-Net and Q-Net, were trained in the country sequence and tested in the other eight sequences as in [7]. Smaller results indicate better performance. In brackets the standard deviation is provided.

Descriptor/Network	Field	Forest	Indoor	Mountain	Oldbuilding	Street	Urban	Water	Mean
EHD	48.62	23.17	30.25	33.94	19.62	27.29	3.72	23.46	26.26
LGHD	18.80	3.73	8.16	11.34	8.17	6.66	7.39	13.90	9.77
Siamese-L2	38.47	12.46	7.94	22.36	15.70	16.85	11.06	29.18	15.50
PN-Net RGB	25.33 (1.08)	4.41 (0.28)	7.00 (0.32)	19.37 (1.07)	7.31 (0.32)	10.21 (0.46)	5.00 (0.27)	17.79 (0.67)	12.05 (0.40)
PN-Net NIR	24.74 (0.98)	4.45 (0.14)	6.54 (0.25)	15.75 (0.44)	7.78 (0.19)	10.82 (0.25)	4.66 (0.14)	16.49 (0.34)	11.40 (0.15)
PN-Net Random	24.56 (1.00)	3.91 (0.20)	6.56 (0.43)	15.99 (0.60)	6.84 (0.31)	9.51 (0.36)	4.407 (0.34)	15.62 (0.61)	10.92 (0.34)
Q-Net 2P-4N (ours)	20.80 (0.81)	3.12 (0.20)	**6.11** (0.27)	12.32 (0.49)	5.42 (0.13)	6.57 (0.40)	3.30 (0.11)	11.24 (0.50)	8.61 (0.14)
PN-Net Random DA	20.09 (0.65)	3.27 (0.27)	6.36 (0.14)	11.53 (0.57)	5.19 (0.20)	5.62 (0.20)	3.31 (0.28)	10.72 (0.36)	8.26 (0.24)
Q-Net 2P-4N DA (ours)	**17.01** (0.33)	**2.70** (0.17)	6.16 (0.18)	**9.61** (0.38)	**4.61** (0.18)	**3.99** (0.09)	**2.83** (0.13)	**8.44** (0.14)	**6.86** (0.09)

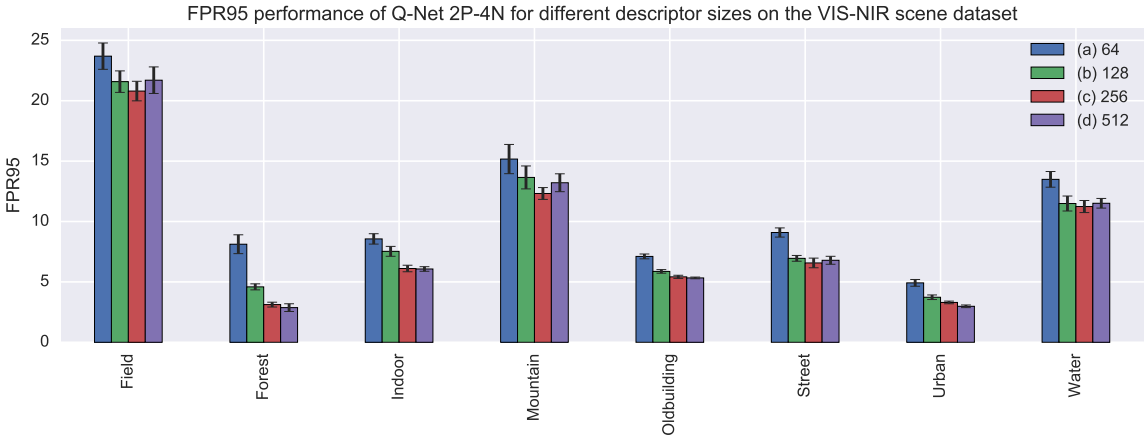


Figure 6. FPR95 performance on the VIS-NIR scene dataset for Q-Net 2P-4N using different descriptor sizes ((a) 64, (b) 128, (c) 256 and (d) 512). Shorter bars indicate better performances. On top of the bars standard deviation values are represented with segments.

[3,17–19]). The dataset contains more than 1.2 million patches of size 64x64 divided into three different sets: Liberty, Notredame and Yosemite, where each image patch was computed from *Difference of Gaussian* (DOG) maxima. We followed the standard protocol of evaluation, training our network three times, one at each sequence, and testing the FPR95 in the remaining two sequences. In our evaluation, we compared our model against two other learned L_2 descriptors, the first from [17] and the second from [3]; which can be considered state-of-the-art.

Table 5. Matching results in the *multi-view stereo correspondence dataset*. Evaluations were made on the 100K image pairs groundtruth recommended from the authors. Results correspond to FPR95. Smallest results indicates better performance.In brackets the standard deviation is provided.

Training Testing	Notredame Yosemite	Liberty	Notredame Liberty	Yosemite	Yosemite Notredame	Liberty	
Descriptor							mean
Siamese-L2	15.15	20.09	12.46	8.38	18.83	6.04	13.49
PN-Net size = 128, patches = 2560000	8.47 (0.20)	9.50 (0.48)	9.17 (0.17)	10.82 (0.49)	4.47 (0.18)	4.16 (0.10)	7.77 (0.17)
PN-Net size = 128, patches = 3840000	8.46 (0.46)	**8.77** (0.23)	8.86 (0.11)	10.78 (0.57)	4.37 (0.14)	3.98 (0.10)	7.53 (0.16)
Q-Net 2P-4N size = 128, patches = 2560000	**7.69** (0.52)	9.34 (0.71)	**7.64** (0.31)	**10.22** (0.60)	**4.07** (0.18)	**3.76** (0.13)	**7.12** (0.22)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2017 doi:10.20944/preprints201703.0061.v1

Peer-reviewed version available at Sensors 2017, 17, , 873; doi:10.3390/s17040873

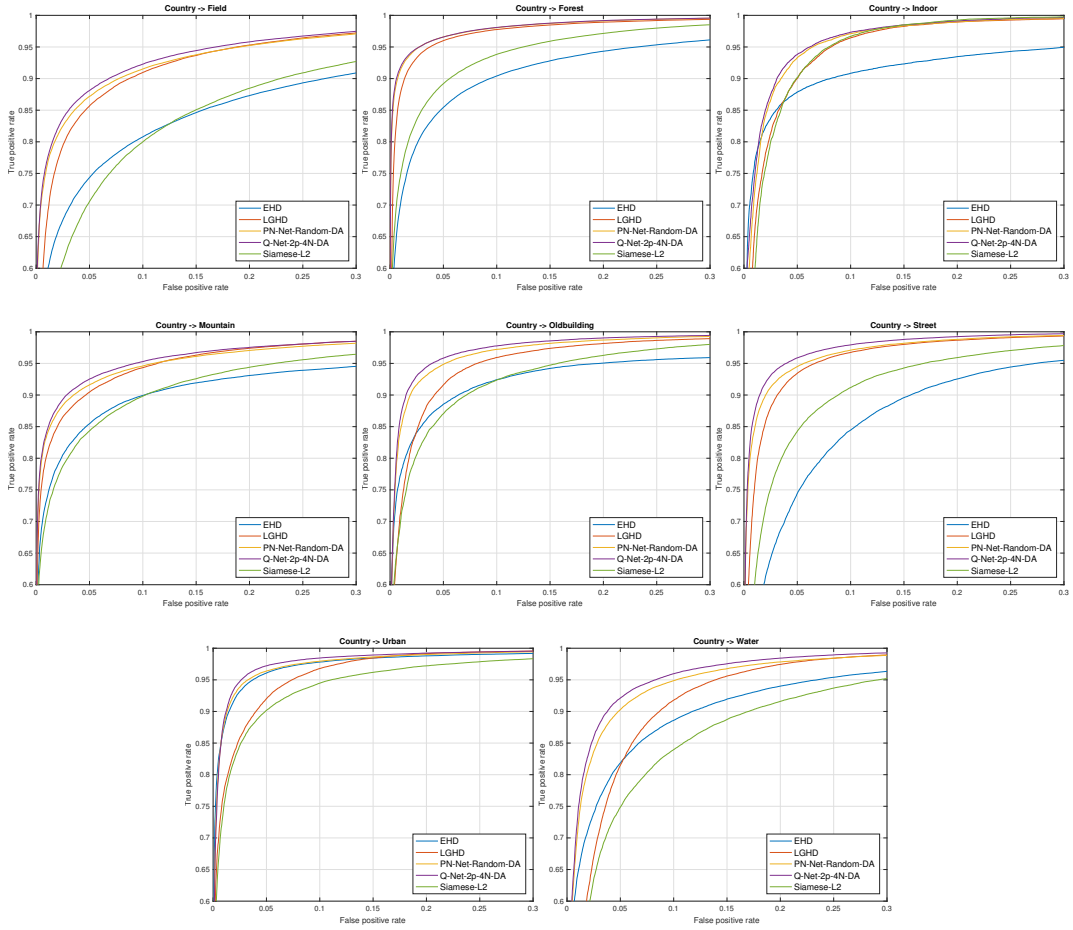


Figure 7. ROC curves for the different descriptors evaluated on the VIS-NIR dataset. For Q-Net and PN-Net, we selected the network with the best performance.**This figure is best viewed in color.**

Training. Quadruplets networks were trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.1, weight decay of 0.0001, batch size of 128, momentum of 0.9 and learning rate decay of 0.000001. Trained data was shuffled at the beginning of each epoch and each input patch was normalized using zero-mean and unit variance. We split up each training sequence into two sets, where 80% of the data was used as training data and the 20% left as validation data. We used the same software and hardware from the previous experiment. As in the previous experiment, Q-Net and PN-Net networks were trained ten times to account for randomization effects in the initialization.

Table 5 shows the results of our experiments. Q-Net and PN-Net performed better than the Siamese-L2 network proposed by [17], which is an expected result, since the siamese-L2 network was not optimized for L_2 comparison during training as the other two networks were. Q-Net performed better than PN-Net by a small margin but using much less training data. When comparing both techniques with the same amount of data, the difference becomes bigger. Meaning that our network needs less data to train than PN-Net, i.e., Q-Net needs less training data than PN-Net and it converges more quickly.

Regarding training time, both networks perform similarly. In our experiments, PN-Net was about 9% faster than Q-Net when both networks where trained with the same amount of patches. In essence, the improved accuracy performance of Q-Net is related to a small loss in training speed.

6. Conclusions and future work

This paper presents a novel CNN-based architecture to learn cross-spectral local feature descriptors. Experimental results with a VIS-NIR dataset showed the validity of the proposed approach, improving the state-of-the-art by almost 3%. The experimental results showed that the proposed approach is also valid for training local feature descriptors in the visible spectrum, providing a network with similar performance to the state-of-the-art, but requiring less training data.

Future work might consider using the same architecture for different cross-spectral applications such as heterogeneous face recognition.

Acknowledgments: This work has been partially supported by the Spanish Government under Project TIN2014-56919-C3-2-R and the Chilean Government under project Fondef ID14I10364. Cristhian A. Aguilera has been supported by Universitat Autònoma de Barcelona.

Author Contributions: The work presented here was carried out in collaboration between all authors. Cristhian A. Aguilera and Angel D. Sappa defined the research topic and implementation. Cristhian Aguilera provided resources to carry on the experiments and helped to edited the manuscript with Ricardo Toledo.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* **2012**, *19*, 4–10.
2. You, C.W.; Lane, N.D.; Chen, F.; Wang, R.; Chen, Z.; Bao, T.J.; Montes-de Oca, M.; Cheng, Y.; Lin, M.; Torresani, L.; others. CarSafe app: alerting drowsy and distracted drivers using dual cameras on smartphones. Proceeding of the 11th annual international conference on Mobile systems, applications, and services. ACM, 2013, pp. 13–26.
3. Balntas, V.; Johns, E.; Tang, L.; Mikolajczyk, K. PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors. *CoRR* **2016**, *abs/1601.05030*.
4. Yi, D.; Lei, Z.; Li, S.Z. Shared representation learning for heterogenous face recognition. Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. IEEE, 2015, Vol. 1, pp. 1–7.
5. Ring, E.; Ammer, K. The technique of infrared imaging in medicine. In *Infrared Imaging*; IOP Publishing, 2015.
6. Klare, B.F.; Jain, A.K. Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 1410–1422.
7. Aguilera, C.A.; Aguilera, F.J.; Sappa, A.D.; Aguilera, C.; Toledo, R. Learning cross-spectral similarity measures with deep convolutional neural networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, 2016, p. 9.
8. Brown, M.; Susstrunk, S. Multi-spectral SIFT for scene category recognition. CVPR; , 2011; pp. 177–184.
9. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **2004**, *60*, 91–110.
10. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. European conference on computer vision. Springer, 2006, pp. 404–417.
11. Firmenichy, D.; Brown, M.; Süsstrunk, S. Multispectral interest points for RGB-NIR image registration. ICIP; , 2011; pp. 181–184.
12. Pinggera, P.; Breckon, T.; Bischof, H. On Cross-Spectral Stereo Matching using Dense Gradient Features. Proc. British Machine Vision Conference, 2012, pp. 526.1–526.12.
13. Morris, N.J.W.; Avidan, S.; Matusik, W.; Pfister, H. Statistics of Infrared Images. 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–7.
14. Aguilera, C.; Barrera, F.; Lumbreras, F.; Sappa, A.; Toledo, R. Multispectral image feature points. *Sensors* **2012**, *12*, 12661–72.
15. Mouats, T.; Aouf, N.; Sappa, A.D.; Aguilera, C.; Toledo, R. Multispectral Stereo Odometry. *IEEE Transactions on Intelligent Transportation Systems* **2015**, *16*, 1210–1224.
16. Aguilera, C.A.; Sappa, A.D.; Toledo, R. LGHD: A feature descriptor for matching across non-linear intensity variations. Image Processing (ICIP), 2015 IEEE International Conference on, 2015, pp. 178–181.

17. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4353–4361.
18. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative Learning of Deep Convolutional Feature Point Descriptors. Proceedings of the International Conference on Computer Vision (ICCV), 2015.
19. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. CVPR, 2015.
20. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research* **2016**, *17*, 1–32.
21. Winder, S.; Hua, G.; Brown, M. Picking the best DAISY. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 178–185.
22. Collobert, R.; Kavukcuoglu, K.; Farabet, C. Torch7: A matlab-like environment for machine learning. BigLearn, NIPS Workshop, 2011, number EPFL-CONF-192376.