DEEP MULTI-SPECTRAL REGISTRATION USING INVARIANT DESCRIPTOR LEARNING

Nati Ofir, Shai Silberstein, Hila Levi, Dani Rozenbaum, Yosi Keller, Sharon Duvdevani Bar

Elbit Systems, Bar-Ilan University, Computer Vision and Algorithms Ltd, Weizmann Institute.

ABSTRACT

In this work, we propose a deep-learning approach for aligning cross-spectral images. Our approach utilizes a learned descriptor invariant to different spectra. Multi-modal images of the same scene capture different characteristics and therefore their registration is challenging. To that end, we developed a feature-based approach for registering visible (VIS) to Near-Infra-Red (NIR) images. Our scheme detects corners by Harris and matches them by a patch-metric learned on top of a network trained using the CIFAR-10 dataset. As our experiments demonstrate, we achieve accurate alignment of cross-spectral images with sub-pixel accuracy. Comparing to contemporary state-of-the-art, our approach is more accurate in the task of VIS to NIR registration.

Index Terms— Deep-Learning, Multi-Spectral Imaging, Image Registration

1. INTRODUCTION

This work addresses the problem of multi-spectral registration, and is aimed specifically to the visible (VIS) $0.4-0.7\mu m$ and Near-Infra-Red (NIR) $0.7-2.5\mu m$ channels. Different spectra capture different appearances making their registration challenging, and it cannot be solved by state-of-the-art approaches for geometric alignment [24]. In Figure 1, the VIS channel captures the color of the scene while the NIR channel captures more details about the far objects, these images differs significantly. In this work we introduce a method for registering such images using deep learning.

Our approach is based on metric learning of cross-spectral image patches. First, we detect feature points using the Harris [9] corner detector. Then we match them to estimate the global transformation relating the input images. Since the patches around these points differ significantly, SIFT [14] matching might fail. Thus, we propose to match them using a deep-learning based approach where the network is trained on CIFAR-10 [12] dataset and is geared to classify $32\times32\times3$ patches in the RGB visible channel into 10 classes. By removing the last soft-max layer we extract an informative 64×1 descriptor for each RGB patch. We train the trimmed net from scratch on NIR patches, such that cross spectral patches of the same object are trained to produce a similar descriptor in the L_2 sense. Thus, we derive an asymmetric Siamese Convolutional-Neural-Network (CNN), the first for VIS descriptor and the second for NIR. These two networks minimize a metric between cross spectral patches, which is the Euclidean distance of the two descriptors. We show experimentally that this metric allows accurate classification of multi-spectral patches as same or different, and is used for the feature-based registration.

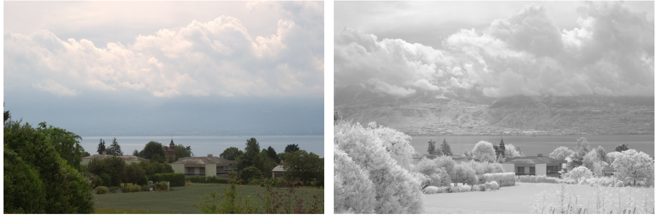

Fig. 1. Example of a pair of multi-spectral images. Left: RGB image of the VIS channel $0.4-0.7\mu m$, contains the information of the color. Right: gray scale NIR image $0.7-2.5\mu m$ captures fine details of the far mountains.

The paper is organized as follows. In Section 2 we detail previous work on the topic of multi-modal registration. In Section 3 we introduce the proposed learning scheme of a deep-descriptor invariant to different wavelengths based on a network trained using the CIFAR-10 dataset. In Section 4, we utilize the descriptor to perform multi-spectral registration. In Section 5 we compare the accuracy of the proposed registration scheme to other state-of-the-art approaches.

2. PREVIOUS WORK

Image registration is a fundamental task in computer vision studied in many works. Image registration [3, 24] is the basis for many applications such as image fusion and 3D object recovery. Early methods rely on basic approaches such as solving translation by correlation [18], while others [15] utilized images gradients and Fast-Fourier-Transform (FFT) domain representation [19]. Key-points [9, 14] and invariant descriptors were also used to estimate geometric alignment [21].

Multi-modal images registration is recently addressed by several works [6, 20]. In [16] the registration was based on maximizing the mutual information. [4, 10, 11, 13] offer to utilize contours and gradients for registration. Cross spectral registration was implemented by correlating Sobel [8], or

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

Canny [5]. A class of works relates to visible to Near-Infrared (NIR) registration [4, 11]. Aguilera et al. [1] apply FAST features [22] and unique descriptors for non linear intensity variations. [2] was the first to measure cross-spectral similarity by Convolutional-Neural-Network (CNN). Their approach classifies pairs of multi-spectral patches as same or different, but do not induce a metric. Therefore, if two patch-pairs are found to be similar their similarity can not be compared to find the best match of a feature-point. Our approach applies CNN's to quantify the similarity of cross spectral patches. This similarity paves the way for the multi-spectral registration scheme as described in Sections 3 and 4.

3. MULTI-SPECTRAL DESCRIPTOR LEARNING

We propose to align a pair of cross spectral images using a feature based approach. We introduce an approach for matching feature points from different spectral channels by their deep-descriptor.

Given a VIS channel patch P_v and a NIR patch P_n we propose to learn a metric that measures the similarity distance between them. The descriptor of P_v is computed from the trimmed network trained on CIFAR-10 [12]. Figure 1 depicts the network architecture denoted by Net_v where the descriptor of the visible channel is $D_v = Net_v(P_v)$. We aim to learn a network Net_n for the NIR channel, using the architecture as in Table 1 but with different weights. The NIR descriptor is $D_n = Net_n(P_n)$. We train the weights of Net_n such that the descriptor D_n is invariant to different wavelengths. Implying that the distance of corresponding patches, P_v and P_n

$$distance(P_v, P_n) = ||Net_v(P_v) - Net_n(P_n)||_2^2 \quad (1)$$

would be significantly less than the distance of non-corresponding patches.

We learn the weights of Net_n using a dataset [4] consisting of 900 aligned VIS and NIR images. For every image we apply the Harris corner detector [9] and extract ~ 1000 patches. Thus, we store over 100,000 corresponding pairs of cross-spectral patches, used as training examples. The input of Net_n is the NIR patch P_n, while the label is $D_v = Net_v(P_v)$. We aim to train the network to output the visible descriptor given a NIR input and maintain the invariance to spectral channels of the distance metric. Figure 2 shows the convergence of the training process and the trained network architecture. It follows that the L_2 distance is decreasing over the epochs, and that the validation graph is closed to the training graph, indicating that there is no over-fitting.

The proposed metric can be used to classify the similarity of cross-spectral patches with high accuracy.

Layer	Type	Output Dim	Kernel	Stride	Pad
1	convolution	32	5×5	1	2
2	max-pooling	32	3×3	2	0
3	ReLU	32	-	1	0
4	convolution	32	5×5	1	2
5	ReLU	32	-	1	0
6	avg-pooling	32	3×3	2	0
7	convolution	64	5×5	1	2
8	ReLU	64	-	1	0
9	avg-pooling	64	3×3	2	0
10	convolution	64	4×4	1	0

Table 1. Architecture of the trimmed network trained on CIFAR-10 dataset. This net gets an 32×32 image patch as an input (VIS of NIR) and outputs its spectral-invariant descriptor.

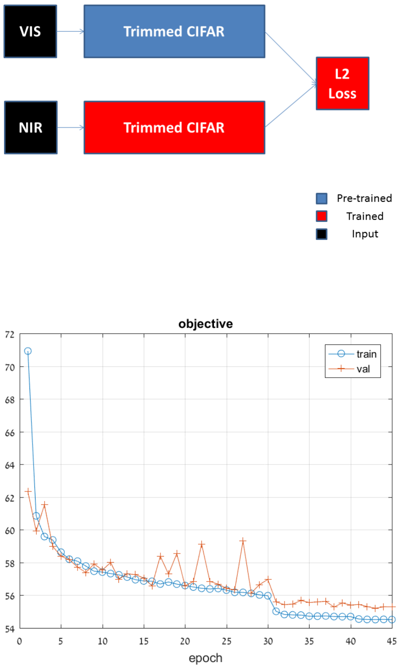

Fig. 2. Top: trained network architecture of the proposed metric learning scheme. Bottom: training process of the NIR descriptor CNN. The L_2 loss decreases over the epochs, while the validation objective is slightly above the training graph.

4. MULTI-SPECTRAL IMAGE REGISTRATION

We use the metric of cross-spectral patches described in Section 3 to form a deep feature based registration. Our approach consists of three stages: corner detection by Harris [9], corners matching using the deep descriptor and estimating the global geometric transformation using Random-Sample-Consensus (RANSAC) [7].

Corners Detection. Denote by V the VIS channel image and by N the corresponding NIR image. We use the method described in [9] to extract the corresponding group of corners C_v and C_n. Each such corner is a local maximum in the Harris score image:

$$S = \lambda_1\lambda_2 - k(\lambda_1+\lambda_2)^2 = \det(A) - k \cdot trace^2(A), \quad (2)$$

where λ_1, λ_2 are the eigenvectors of the matrix of derivatives for each pixel:

$$A = \sum_u \sum_v w(u,v) \begin{pmatrix} I_x(u,v)^2 & I_x(u,v)I_y(u,v) \\ I_x(u,v)I_y(u,v) & I_x(u,v)^2 \end{pmatrix}. \quad (3)$$

I_x, I_y are the horizontal and vertical derivatives of the input image respectively. Since this corner detection method is based on local gradients, it is relatively invariant to multi-spectral images and therefore the group of corners C_v and C_n has a large overlap. This characteristic of the feature extraction is essential for our scheme.

Feature Matching. We aim to match the points in C_v to those in C_n to form the group of all matches M. For every point $p_v \in C_v$ we find the best match $p_n^* \in C_n$ by computing the descriptor of C_v by Net_v and the descriptors of C_n by Net_n. The complexity of the forward passes is $|C_v| + |C_n|$ making it applicable for real time registration. A match $m \in M$ is a pair (p_v, p_n^*), such that p_n^* is the nearest neighbour of p_v according to our deep metric. It means that their descriptors D_v and $D_{p_n^*}$ are the closest out of all other possible matches.

Transformation Computation. We use the set of all matches M to form the final global transformation H between the images V and N. Typically, most of the matches in M are outliers. Therefore, we compute the transformation, H by RANSAC [7]. The transformation H for every sample of matches from M is computed by least-squares. We look for the largest sub-group of M that accepts on the same transformation H, this group contains the inlier matches. H can relate to affine, rigid or translation transformations. As the number of parameters in H decreases, the registration accuracy increases. The accuracy metric is the ratio of inliers divided by $|M|$ where a score greater than 0.1 indicates a successful run of our method.

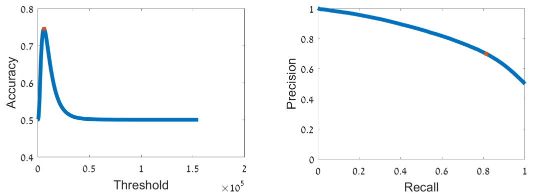

Fig. 3. Evaluation of our deep-metric as a binary classifier to same or different pairs of cross-spectral patches. Left: accuracy of the classifier as a function of threshold, the top score is 0.74. Right: precision-recall graph of the classifier, the obtained F-score is 0.75.

5. EXPERIMENTS

We trained and tested our method on cross-spectral images from the datasets of [4]. This dataset contains over 900 aligned images from the VIS and NIR channels. In Figure 6 we show images from this dataset. Our code is implemented in Matlab using the MatConvNet library [23]. The runtime of our registration is around 10 seconds per pair of images and can be further reduced by utilizing GPU and parallel computing. The training time for our network is one hour on a Titan-X GPU. We trained the network with learning rate of 0.005 and weight decay of 0.0004. To evaluate the registration accuracy we simulated transformations on the dataset of aligned cross-spectral images and estimated them using the proposed approach. For each run of a simulation we report the estimation error. We compared our approach to multiple approaches for multi-spectral registration. The first is to use edge descriptors and match them using binary correlation. Other approaches estimate only translation, using correlation of Canny [5], correlation of Sobel [8] and the maximization of mutual information. We also compare to the feature based approach of LGHD descriptor [1].

In Figure 3 we show the classification of a pair of patches to be similar or different. The positive set is the pair of patches around corners in the dataset while the negative examples are produced by random sampling. The accuracy of our binary-classifier is 0.74 when selecting the correct threshold on the L_2 distance between the descriptors. The F-measure [17], $F = \frac{2*precision*recall}{precision+recall}$, is 0.75 and it is achieved with a similar threshold for maximizing the accuracy.

Table 2 compares the different methods for estimating pure translations. It follow that the proposed scheme achieves the lowest error close to 0. In Figure 4 we depict this error for the samples of different scenes. It follows that our error is the lowest across most of the scenes.

Figure 5 shows the error of our deep-method across different scalings of the simulated transformation where we achieve an error of ~ 1 pixel in all scalings levels. In the difference of the scaling parameter we gain a negligible error < 0.002.

6. CONCLUSIONS

We introduced a novel approach to multi-spectral registration that utilizes an invariant deep descriptor of cross-spectral

Algorithm	VIS-NIR
Our method	**0.03**
Edge-Descriptor	0.08
Canny	0.07
Sobel	0.07
Mutual Information	0.11
LGHD	0.21

Table 2. Error in pixels of multi-spectral registration when estimating pure translations. Our deep method is compared to utilizing edge descriptors, correlation of Canny [5], correlation of Sobel [8], maximization of mutual-information and LGHD [1]. It follows that (in bold), the proposed deep-algorithm achieves the highest accuracy.

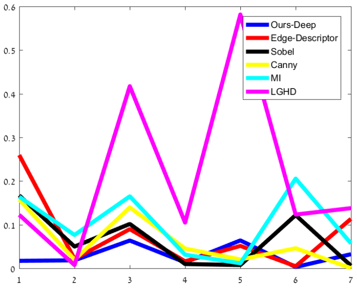

Fig. 4. Error in pixels of cross-spectral registration methods over sample of scenes. It follows that the proposed deep method achieves the lowest error in most of the examples.

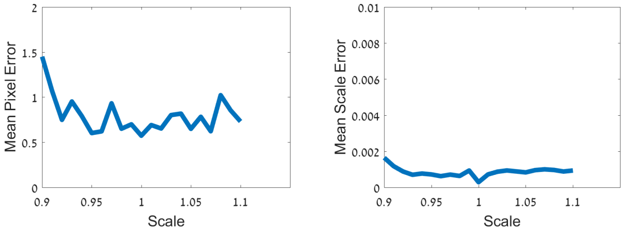

Fig. 5. Evaluation of registration error across simulated scaling transformations. Left: error of the translation parameters when solving scales from 0.9 ro 1.1. Right: error of the scaling parameter acrros the same range of scalings between the cross-spectral images. The translation error is ~ 1 pixels while the scaling error is negligible.

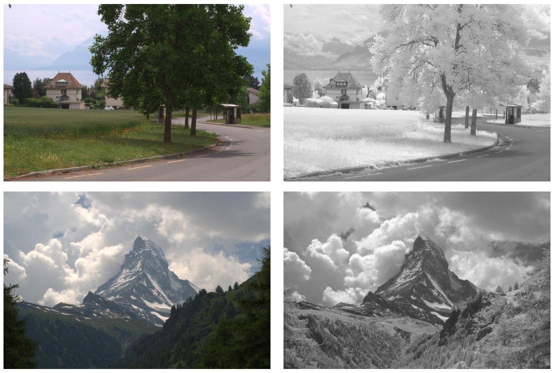

Fig. 6. Pairs of aligned cross-spectral images from the dataset we used to train and evaluate our method [4]. We used 90% of the images for training and 10% for testing

patches. For that end, we trained a CNN to extract such a descriptor for NIR patches. This CNN alongside the trimmed network pre-trained using CIFAR-10 for RGB patches, form a metric between multi-spectral patches. Our experiments demonstrate that the proposed metric-learning scheme is useful for classifying pair of patches as same or different. Moreover, it paves the way for accurate multi-spectral registration. In future we propose to derive a fully end-to-end network that will carry out all the stages of our feature based registration including corner detection and feature matching.

7. REFERENCES

[1] C. 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*, page 5. IEEE, Sep 2015.

[2] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, and R. Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[3] L. G. Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.

[4] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, June 2011.

[5] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[6] C. Chen, Y. Li, W. Liu, and J. Huang. Sirf: simultaneous satellite image registration and fusion in a unified framework. *IEEE Transactions on Image Processing*, 24(11):4213–4224, 2015.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[8] W. Gao, X. Zhang, L. Yang, and H. Liu. An improved sobel edge detection. In *Computer Science and Information Tech-*

nology (ICCSIT), 2010 3rd IEEE International Conference on, volume 5, pages 67–71. IEEE, 2010.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Manchester, UK, 1988.

[10] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Computer Vision, 1998. Sixth International Conference on*, pages 959–966. IEEE, 1998.

[11] Y. Keller and A. Averbuch. Multisensor image registration via implicit similarity. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):794–801, 2006.

[12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[13] H. Li, B. Manjunath, and S. K. Mitra. A contour-based approach to multisensor image registration. *IEEE transactions on image processing*, 4(3):320–334, 1995.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[15] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[16] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on medical imaging*, 16(2):187–198, 1997.

[17] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[18] W. K. Pratt. Correlation techniques of image registration. *IEEE transactions on Aerospace and Electronic Systems*, (3):353–358, 1974.

[19] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8):1266–1271, 1996.

[20] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. In *European Conference on Computer Vision*, pages 309–324. Springer, 2014.

[21] M. Subramanyam et al. Automatic feature based image registration using sift algorithm. In *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on*, pages 1–5. IEEE, 2012.

[22] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod. Unified real-time tracking and recognition with rotation-invariant fast features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 934–941. IEEE, 2010.

[23] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM International Conference on Multimedia*, 2015.

[24] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.