Image patch similarity through a meta-learning metric based approach

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

[1]ESPOL Polytechnic University, Escuela Superior Politécnica 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*—Comparing images regions are one of the core methods used on computer vision for tasks like image classification, scene understanding, object detection and recognition. Hence, this paper proposes a novel approach to determine similarity of image regions (patches), in order to obtain the best representation of image patches. This problem has been studied by many researchers presenting different approaches, however, the ability to find the better criteria to measure the similarity on image regions are still a challenge. The present work tackles this problem using a few-shot metric based meta-learning framework able to compare image regions and determining a similarity measure to decide if there is similarity between the compared patches. Our model is training end-to-end from scratch. Experimental results have shown that the proposed approach effectively estimates the similarity of the patches and, comparing it with the state of the art approaches, shows better results.**

***Index Terms*—meta-learning, metric based, Siamese Networks, Convolutional networks**

I. Introduction

One of the computer vision techniques that has always been in constant research is the determination of the similarity of the image regions, because it is the fundamental process of many vision tasks such as object recognition, stereo vision, image registration, image denoising, exemplar-based image inpainting,

The ease with which humans can differentiate whether two images are similar or not, is one of the challenges that still remain in the field of computer vision, many of the traditional techniques are based on encoding images into representation vectors, for which it is necessary to take small regions of the images to be compared and distance metrics, such as the euclidean, are used to determine the correlation between the regions of the images. Another applications that can be derived from a good management of the regions of the images is the edition of the same ones to modify the position of objects, to make changes in the texture or any other adjustment that is required to make in an image. Also, in the analysis of medical images, techniques based on regions of images (patches) are also observed in order to compare the similarity of the images with the related databases already existing to determine whether or not they are similar [1].

Another approach is learn a feature representation directly from image data, to obtain a general similarity function for comparing image patches. To formulate such a function, various CNN-based model has been designed and trained to support a wide variety of changes in image appearance [2]. Many factors could affect the comparing process of images, such as occlusion, illumination, quality of sensors, etc., for this reason multiples approaches could be developed, from hand-craft methods to a deep learning, in order to obtain this kind of information valid for high level vision problems like object recognition, classification, super-resolution, surveillance system, etc.

Several feature descriptors have been proposed last decades to resolved many computer vision problems based on principal characteristics of the images, descriptors like (e.g., SIFT [3], SURF [4], KAZE [5], among the best known). Many researchers have been working with image patches for spatial analysis, for road detection and scene understanding, which can be used for image labeling [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]. Other research have been proposed a image-adaptive wavelet transform, to form a multi-scale sparsifying global transform for the image [9].

The previous approaches have been developed using images or patches from visible spectrum and with the lower prices of cameras sensitive to several spectra, especially visible and near infrared, being even available on smart devices, capturing images of several spectra simultaneously makes it possible to propose new architectures based on convolutional neural network to learn visual similarities with success working on images in the cross-spectral domain, this information of the near infrared spectrum can help the processing of images with low levels of illumination, or to improve the quality of the images of the visible spectrum.(e.g., filtering [10], enhancement [11]). The modeling of architectures that make use of cross-spectral information is a difficult and challenging task since it implies using such information efficiently and

effectively to solve a given problem that already exists in the domain of the visible spectrum. For example, different works have recently been proposed to describe and match characteristic points in images in the cross spectral domain based on classical approaches (e.g., [12], [13],[14], [15], to mention a few). However, the results and performance obtained with these techniques still require much improvement to reach the yields achieved with the techniques using images of the visible spectrum.

Some approaches have been developed in the cross-spectral field that seek to propose architectures that can solve problems of comparisons of image patches, which show better results than traditional techniques or those implemented with CNN networks but using only images of the visible spectrum (e.g., [16], [17]. However, lately, meta-learning techniques have been proposed that allow you to generalize a model from few data, which is very useful. Also, to use other sources of data that are not labeled but plenty available, multimodal learning, transfer learning, continual learning or domain adaptation.

For meta-learning, it is important that a specific transformation of a subset of features was useful for transfer of knowledge, to obtain a distribution of patterns in the feature space that share some characteristics that may be described by a data model and easy to adapt to a new dataset of similar type to increase the representation of the features to generate a new one

As mentioned above, the main contribution of current work is to reach to a better performance compared to [17] and [16]. The rest of the paper is organized as follows. Section II describes the most recent work on image patch similarity learning. Section III presents the architecture proposed using meta-learning approach using cross-spectral datasets. Section IV depicts the experimental results and finally, conclusion are presented in section V.

II. Related Work

Several approaches for image patching similarity have been proposed in last years, some techniques are proposed based on mathematical theory, or using CCN networks. In [18] propose a novel region-based active contour model via local patch similarity measure for image segmentation. Using the spatial constraints on local region-based models to construct a patch similarity measure which balances the noise suppression and the image details reservation. Another approach [19] propose a novel deep similarity learning method that trains a binary classifier to obtain the metric of the correspondence of two image patches. The classification output is transformed to a continuous probability value, then used as the similarity score, for the comparison two commonly used metrics are presented: normalised mutual information and local cross correlation.

Also, in [20] present an approach to learn data representations using an autoencoder for defect detection. However, the texture (non-defect) area cannot be well reconstructed, which makes the pixel-wise detection inaccurate, for this reason explore similarities between different patches in the whole test image, a novel autoencoder-based fabric defect detection method is proposed. In order to maintain the texture area in the reconstructed patch, the original encoded latent variable is modified, and the cross-patch similarity is introduced for determining the modification function. In [21] present a technique to perform registration of images of different nature using SAR and optical images, using a neural network in order to build feature point descriptors and then use RANSAC algorithm to align found matches. Another approach [22] propose a deep local descriptor learning framework for cross-modality face recognition, to learn discriminant and compact local information directly from raw facial patches. Also includes a novel cross-modality enumeration loss to eliminate the modality gap on local patch level. In [23] propose an approach to use cross-spectral images to achieve a better performance with the adaptive Harris corner detector, which means improving the detection of characteristic points using cross-spectral images (NIR, G, B) and applying pruning techniques, the combination of channels for this fusion is the one that generates the largest variance based on the intensity of the merged pixels, therefore, it is that which maximizes the entropy in the resulting Cross-spectral images, with this technique the obtained results are better than those achieved with images of the visible spectra. Song et al. in [24] present an adversarial discriminative feature learning framework to close the sensing gap via adversarial learning on both raw-pixel space and compact feature space. The approach integrates cross-spectral face hallucination using generative adversarial networks and discriminative feature learning into an end-to-end adversarial network. In the feature space, an adversarial loss and a high-order variance discrepancy loss are employed to measure the global and local discrepancy between two heterogeneous distributions respectively to enhance domain-invariant feature learning and modality independent noise removing. In this work [25], the authors propose a new approach to align two images related by an unknown 2D homography where the local descriptor is learned from scratch from the images and the homography is estimated simultaneously. This technique uses a siamese convolutional neural network optimize by a single loss function. This method has been designed to align images of different modalities such as RGB and near-infra-red (NIR) without using any prior labeled data. In this paper [26], the authors present a deep coupled learning approach to solve the problem of matching polarimetric thermal face photos against a gallery of visible spectrum faces. With the polarization state information of thermal faces is possible to obtain the missing textural and geometrics details in the thermal face imagery which exist in visible spectrum. A coupled deep neural network model has been designed which leverages relatively large visible and thermal datasets to overcome the problem of overfitting, also finds global discriminative features in a nonlinear embedding space to relate the polarimetric thermal faces to their corresponding visible faces.

III. Proposed Approach

One of the challenges posed by meta-learning techniques is the design of a deep training model that using only a few

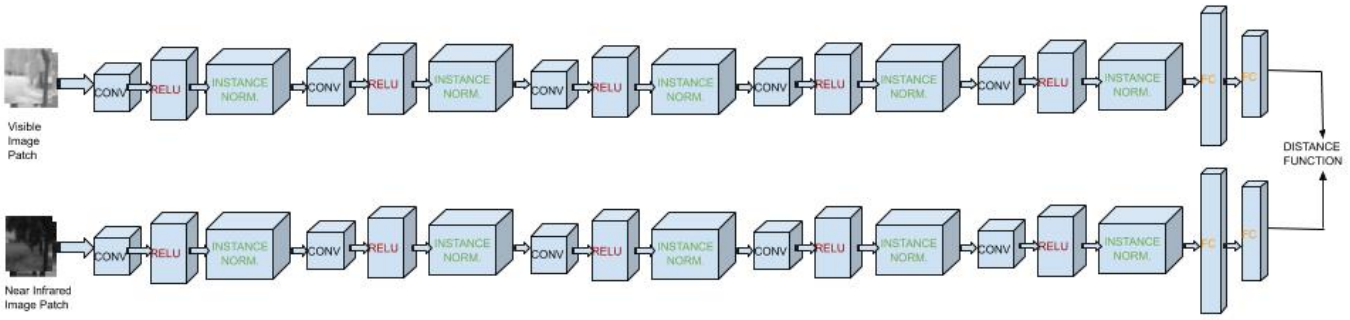

Fig. 1. Siamese General Schema implemented on the current research.

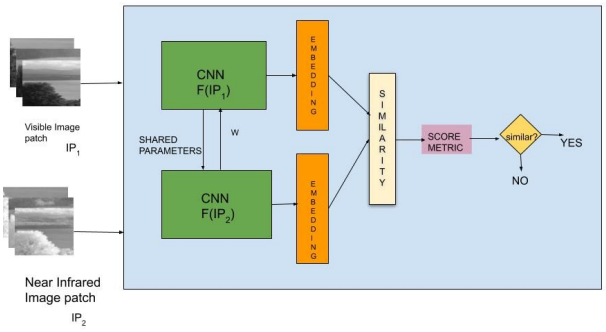

Fig. 2. Siamese General Schema implemented on the current research.

training data given the previous experience taken from very similar learning tasks. This technique is known as learn from few data shots, trying to simulate the human capacity to learn from one or a few examples and what is proposed in this work is to create an architecture that is capable of detecting the similarity of patches of cross-spectral images, in our case It has been proposed to generate a similarity metric based solely on $K - shotsinN - ways$", learning in which we are given little training data (for example, images of certain classes such as urban, oldbuildings, etc) to determine whether or not similarity exists between the K classes with N data in each.

Once the model have been trained, the similarity metric can deduce the pattern of the common characteristics that represents the images evaluated by the meta-learning architecture optimized by the model parameters already obtained. The model parameters have been designed to be shared and its optimization model are :

$$\theta^* = arg\min_{\theta} \mathbb{E}_D \sim_{p\ (\mathrm{C})} [\mathcal{L}_\theta(C)] \tag{1}$$

where θ^* tries to optimize the model to obtain a semantic embedding space based on few shot samples and labels $\mathbb{E}_D \sim_p$ through the learning process to generate the representation vectors and determine the patching similarity using the corresponding dataset D.

The proposed approach is based on a cross-spectral metric based siamese network approach, see Fig. 2 that look for a model capable of determine the similarity of the cross-spectral image patches of five different pattern (field, mountain, indoor, oldbuilding, urban) categorized separately, with 16 examples in each class, having 50% of these examples with images of the visible spectrum and the rest with images of the near infrared spectrum, with a 8 samples on each class per spectra, for the training process and for the test process there are three classes not seen by the training process with 8 examples in each, of which half are images of the visible spectrum and the rest are of the near infrared spectrum.

The meta-learning model proposed in this work has been designed to be trained over a variety of classes and at the same time to obtain a good performance on the learning of metric similarity of image patches. Being C the cross-spectral dataset of all image patches of the all classes to be considered in the training process to perform the learning similarity tasks and optimized for the best accuracy. Each task is associated with a cross-spectral dataset C, containing both patch images representation and their corresponding labels.

Being IP_1 and IP_2 a pair of image patches from visible and near infrared spectra respectively and let L being their corresponding label; "0" for a similar image patch pair class and "1" for a non similar image patch pair, including cross-spectral image pairs existing in the training and test database. Let W be the shared weights in the siamese network architecture, see Fig. 1 which will be optimized incrementally as the proposed model is generalized. Having a generator function $G_w(ip)$ instantiated by a Siamese architecture with a weight vector W. Being the siamese net instantiated $G_(ip_1)$ and $G_w(ip_2)$ to obtain a embedding vector representation on each side of the network to measure the distance between those embeddings and determine the similarity of the patches feed it into the network. This similarity function $G_w(IP_1, IP_2)$ is defined as :

$$G_w(IP_1, IP_2) = \|G_(ip_1) - G_w(ip_2)\|^2, \quad (2)$$

A. Instance Normalization

Deep learning is a technique that allow to learn multiple levels of representation and abstraction to transform data in order to resolve an specific problem. Many researches with deep learning are focusing on developing techniques to stabilize training. Thus, some architectures are known to be unstable (during training) and very sensitive to changes over the hyper-parameter values of the model. Another field of analysis has emerged around the style of an image evaluated by the statistics of convolutional neural network filters, a renewed interest in the texture generation and image stylization problems to obtain qualitative improvement in the generated image.

Ulyanov et al. [27] shows that it is possible to apply a method named Instance Normalization over the data of the model for a better stylization and texture synthesis, that derive entropy loss which improves samples diversity. This method prevents instance-specific mean and covariance shift simplifying the learning process. The instance normalization layer is applied at test time as well as at training time. According to [27] the generator network should discard contrast information in the content image to learn a highly nonlinear contrast normalization function as a combination of such layers. Let $x \in \mathbb{R}^{NCWH}$ an input tensor containing a batch of N images, where C, W and H are the depth, width and high respectively of the tensor and let x_{tijk} denote its $tijk$-th element of t tensor, where k and j span spatial dimensions, i is the feature channel (in the case of an RGB image being used as an input, it would represent a color channel). Thus, a simple version of instance normalization is defined as:

$$y_{tijk} = \frac{x_{tijk}}{\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm}}. \quad (3)$$

A small change in the stylization architecture proposed by [27] presents a qualitative improvement in the generated embedding vector. The change is limited to swapping batch normalization with instance normalization, and to apply the latter both at training and testing times. The resulting method can be used to train high-performance architectures for real-time embedded vector generation. Our architecture uses this normalization, applied in feed-forward style transformation, to improve the quality of the embedded feature representation generated by the model.

B. Contrastive Loss

To be able to differentiate if the images that are fed to the model are similar or not, a representation of smaller size must be obtained that conserves the information of the structure and semantics of the image, for which the contrastive loss based on a maximum margin has been used previously defined. According to [28] a meaningful mapping from high to low dimensional space maps similar input vectors to nearby points on the output manifold and dissimilar vectors to distant points. This loss function whose minimization runs over pairs of samples. Let $\overrightarrow{\mathrm{X}_1}$, $\overrightarrow{\mathrm{X}_2} \in$ I a set of images, be a pair of input vectors shown to the system. Let Y be a binary label assigned to this pair. Y = 0 if $\overrightarrow{\mathrm{X}_1}$, and $\overrightarrow{\mathrm{X}_2}$ are similar, and Y = 1 if they are dissimilar. Define the parameterized distance function to be learned D_W between $\overrightarrow{\mathrm{X}_1}$, $\overrightarrow{\mathrm{X}_2}$ as the euclidean distance between the outputs of G_W. This loss is defined as:

$$\mathcal{L}(\mathcal{W}) = \sum_{m=1}^{P} (L, W(Y, \overrightarrow{\mathrm{X}_1}, \overrightarrow{\mathrm{X}_2})^i) \quad (4)$$

Applying this loss to our model, where the image pairs are the image patches of visible and near infrared spectra. It is defined as :

$$(L, W(Y, \overrightarrow{\mathrm{X}_1}, \overrightarrow{\mathrm{X}_2})^i) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}(max(0, m - D_W)^2 \quad (5)$$

Algorithm 1 Image Patch Similarity for *country*, *urban* and *indoor* classes; i, j are the number image patches per category $epochs$ is the number of iterations for training process; n is the number of images per batch;t is the number of images of the training dataset.

for `Number image patches in training set` (t), $\overrightarrow{\mathrm{IP}_1}$ **do**
  Determine $\overrightarrow{\mathrm{IP}_2}$ pair from the complete sample training set: $S_{\overrightarrow{IP}}$ where $\{\overrightarrow{\mathrm{IP}_1}\}_{i=1}^{t}$ are similar $\{\overrightarrow{\mathrm{IP}_2}\}_{j=1}^{t}$ and calculating their corresponding label, Y_{ij}, so that Y_{ij} = 0 when $\{\overrightarrow{\mathrm{IP}_1}\}_{i=1}^{t}$ = $\{\overrightarrow{\mathrm{IP}_2}\}_{j=1}^{t}$ and Y_{ij} = 1 otherwise.
end for
for $epochs$ `steps` **do**
  for n `batch steps` **do**
    Initialize network weights
    Instance the architecture of the Siamese network
    Contrastive loss calculation
    Neural Net Optimization
    Contrastive loss calculation, by minimization
  end for
  Fine-tuning weights of the net based on the error rate
end for

IV. Experiments Results

A. Results and comparisons

To check the proposed 8-shot 1-way" meta-learning metric based network model, the cross-spectral data set of cite brown2011multi has been used (in Fig. Ref fig: some examples of the data some examples of pairs of some are presented categories of images). This data set consists of 477 registered images classified in 9 groups captured in RGB (visible spectrum) and NIR (near infrared). To make the comparison with the previous approaches [17] and [16]

Descriptor-Network	Country	Indoor	Oldbuilding	Urban
SIFT[3]	46.6	12.4	21.3	13.27
2ch Network (from [17])	0.23	4.4	2.3	**1.58**
2ch Network (from [16])	0.27	3.3	3.4	4.6
Metric Based Network (cross-spectral proposed)	**0.22**	**3.1**	**2.2**	1.63

TABLE I
EVALUATIONS (FPR95%) ON CROSS-SPECTRAL IMAGE PATCH DATASETS [29] FROM DIFFERENT CATEGORIES (THE SMALLER THE BETTER, BOLD FACES CORRESPOND TO THE BEST RESULTS IN THAT CATEGORY).

Descriptor-Network	Country	Indoor	Oldbuilding	Urban
Metric Based Network (cross-spectral proposed)	0.22	3.1	2.2	1.63
Metric Based Network (Visible only proposed)	**0.17**	**1.6**	1.9	1.21
Metric Based Network (Near infrared only proposed)	0.19	2.6	**1.6**	**1.43**

TABLE II
EVALUATIONS (FPR95%) ON VISIBLE AND NEAR INFRARED IMAGE PATCH DATASETS EVALUATED SEPARATELY [29] FROM DIFFERENT CATEGORIES (THE SMALLER THE BETTER, BOLD FACES CORRESPOND TO THE BEST RESULTS IN THAT CATEGORY).

the images of the category "Country" have been used for training. For the experiments (8 pairs of randomly selected images from visible and near infared has been selected). These images are the most affected in lighting conditions and variable textures, which directly affects the complexity of the process of establishing their similarity through the detection of characteristic points and, therefore, are the most challenging scenarios for the training process. First, the characteristic points of patches of the visible spectrum images have been obtained using the SIFT algorithm, to search these points in their corresponding on the near infrared spectrum images. To carry out the experiments, 64 $times$ 64 pixels patches have been generated centered on the previously detected points in both the visible and near infrared spectrum images, since the images are perfectly aligned, then the corresponding patches are extracted with the previously defined size. For training, a total of 16 perfectly balanced cross-spectral image patches (matched and not matched) have been prepared for each category, It have been used Adam optimizer with a learning rate of 0.0002, with a stochastic gradient descent, minimizing the contrastive loss to converge the model.

Once the meta-learning metric based model has been trained with images from the "Country" category, it has been evaluated with other categories cross-spectral in addition to those of images from the "Country" category together with other categories. Thus, 8 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 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 [17] and [16]. It can be used to evaluate results from the same categories, Table I shows the obtained performances. As expected, it can be appreciated the large improvements reached with respect to SIFT. Additionally, it can be appreciated also a better results than those presented in [17] and [16], actually, only in the "Urban" category previous approaches remain a bit better than the ones obtained in our approach. Also we have evaluated our model to validate the similarity between only visible and near infrared image patches separately, just to show that our model is able to measure similarity without having to train the model again showing that it is adaptable to various measurement tasks. the results are shown in the Table II

V. CONCLUSION

This research paper addresses the challenging problem of measuring the similarity of images belonging to different categories and to different spectra, for this case, visible and near infrared, for which, a metric-based meta-appendix model has been used, which allow the model to learn categories of objects from few examples, and at a rapid pace, trying to simulate the learning that humans do and that fundamentally does not depend on a great computational power, but that the designed architecture is able to synthesize in a way Efficient and effective an embedded representation that allows learning new classes from existing information on different classes previously learned.

The results show that it is possible even with a few show samples to obtain a performance 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 architectures and normalization techniques will be considered for improving results. As a work in progress, we have been working on the design of a meta-learning model based on a matching network for the classification of volcanic stone samples.

ACKNOWLEDGMENT

This work has been partially supported by: the ESPOL project PRAIM (FIEC-09-2015); the Spanish Government under Projects TIN2014-56919-C3-2-R and TIN2017-89723-P; and the "CERCA Programme / Generalitat de Catalunya".

(*a*) (*b*)

Fig. 3. Cross-spectral pairs of images obtained from [29]: (*a*) visible images; (*b*) NIR images.

The authors gratefully acknowledge the support of the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559). The authors would also like to thank NVIDIA for GPU donations.

References

[1] G. Wu, B. C. Munsell, Y. Zhan, W. Bai, G. Sanroma, and P. Coupé, *Patch-Based Techniques in Medical Imaging: Third International Workshop, Patch-MI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Proceedings*. Springer, 2017, vol. 10530.

[2] 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.

[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, Jun. 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] 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.

[13] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1210–1224, 2014.

[14] 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.

[15] 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

[16] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Cross-spectral image patch similarity using convolutional neural network," in *Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), 2017 IEEE International Workshop of*. IEEE, 2017, pp. 1–5.

[17] 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.

[18] H. Yu, F. He, and Y. Pan, "A novel region-based active contour model via local patch similarity measure for image segmentation," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 24 097–24 119, 2018.

[19] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2018.

[20] H. Tian and F. Li, "Autoencoder-based fabric defect detection with cross-patch similarity," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[21] D. Abulkhanov, I. Konovalenko, D. Nikolaev, A. Savchik, E. Shvets, and D. Sidorchuk, "Neural network-based feature point descriptors for registration of optical and sar images," in *Tenth International Conference on Machine Vision (ICMV 2017)*, vol. 10696. International Society for Optics and Photonics, 2018, p. 106960L.

[22] C. Peng, N. Wang, J. Li, and X. Gao, "Dlface: Deep local descriptor for cross-modality face recognition," *Pattern Recognition*, vol. 90, pp. 161–171, 2019.

[23] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Adaptive harris corner detector evaluated with cross-spectral images," in *International Conference on Information Theoretic Security*. Springer, 2018, pp. 732–744.

[24] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] J. Dong, B. Boots, F. Dellaert, R. Chandra, and S. Sinha, "Learning to align images using weak geometric supervision," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 700–709.

[26] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi, "Deep cross polarimetric thermal-to-visible face recognition," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 166–173.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6924–6932.

[28] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

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