ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414341

MATCHING AS COLOR IMAGES: THERMAL IMAGE LOCAL FEATURE DETECTION AND DESCRIPTION

Bhavesh Deshpande, Sourabh Hanamsheth, Yawen Lu, Guoyu Lu

Intelligent Vision and Sensing Lab, Rochester Institute of Technology

ABSTRACT

Feature detection and extraction is considered to be one of the most important aspects when it comes to any computer vision application, especially the autonomous driving field that is highly dependent on it. Thermal imaging is less explored in the field of autonomous driving mainly due to the high cost of the cameras and inferior techniques available for detection. Due to advances in technology the former does not hold true anymore and there lies tremendous scope for improvement in the latter. Autonomous driving relies heavily on multiple and sometimes redundant sensors, for which thermal sensors are a preferred addition. Thermal sensors being completely dependent on the infrared radiation emitted are able to frame and recognize objects even in the complete absence of light. However detecting features persistently through subsequent frames is difficult due to the lack of textures in thermal images. Motivated by this challenge, we propose a triplet based Siamese CNN for feature detection and extraction for any given thermal image. Our architecture is able to detect larger number of good feature points on thermal images than other best performed feature detection algorithms with superb matching performance based on our extracted descriptors.

Index Terms— Thermal Imaging, Triplet Siamese CNN, Unsupervised Network, Feature Detection and Description

1. INTRODUCTION

Extracting features on visible images is well established. Features such as SIFT [1], SURF [2], and ORB [3] provide us with good and distinct local image descriptions. There are also several other well-defined edge and corner detection algorithms such as Canny edge detector, harris corner detector [4], etc. These detectors form a core part for SFM and SLAM [5] [6] methods. However these same techniques fail to yield good quality feature detection results when used on thermal images. For the case of autonomous driving or navigation in general, visible images should be sufficient for most of the time, but in conditions of low to none visibility caused by the environment such as rain, fog, snow, night etc., even a good quality high resolution camera is not sufficient. On the other hand thermal imaging is robust to all these aforementioned conditions and provides us with high quality information of the scene. Nevertheless there are certain drawbacks when it comes to thermal imaging - The high texture quality observed in visible images is lost in case of thermal images, which also forms the basis for standard feature detection and extraction algorithms. Thermographic cameras detect infrared radiation emitted from the body of an object making it possible to have good visibility even in the absence of illumination. Warm blooded animals or object thus become easy to detect making thermal imaging useful in military and surveillance applications and has a great potential in the field of autonomous driving as well.

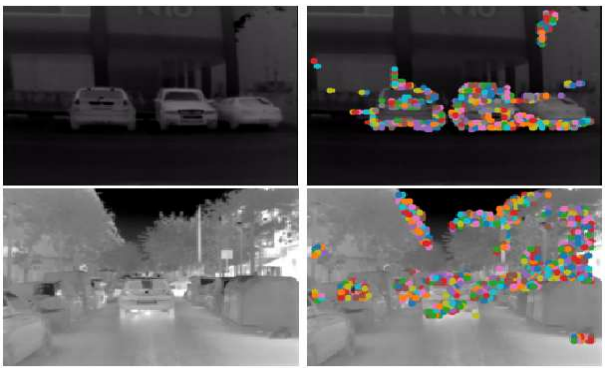

Fig. 1. Feature keypoint detection results. Left: Input image frame under consideration; Right: our network output.

In the deep learning era, obtaining unique, better and more accurate features from images has improved immensely. But almost all the approaches based on images from the visible spectrum which are rich in feature textures failing to completely tackle the aforementioned problems. BRIEF [7], SIFT [1], SURF [2], ORB [3], FAST [8] and WADE [9] algorithms are able to detect some features in thermal images but the quality of detection is rather unsatisfactory. Image-to-image translation using Adversarial networks for thermal-visible domain transfer can be used as a workaround but the generated images are based on assumptions and also may add additional artifacts which are not desirable.

In this paper, we introduce a feature detection and extraction architecture based of Triplet neural network. For detection network, it takes in three image patches namely Anchor, Positive and Negative. These patches are passed through a series of convolutional and fully connected layers. The resulting vector is kept to be 128 Dimensional helping the network retain valuable feature information from the patches. Fea-

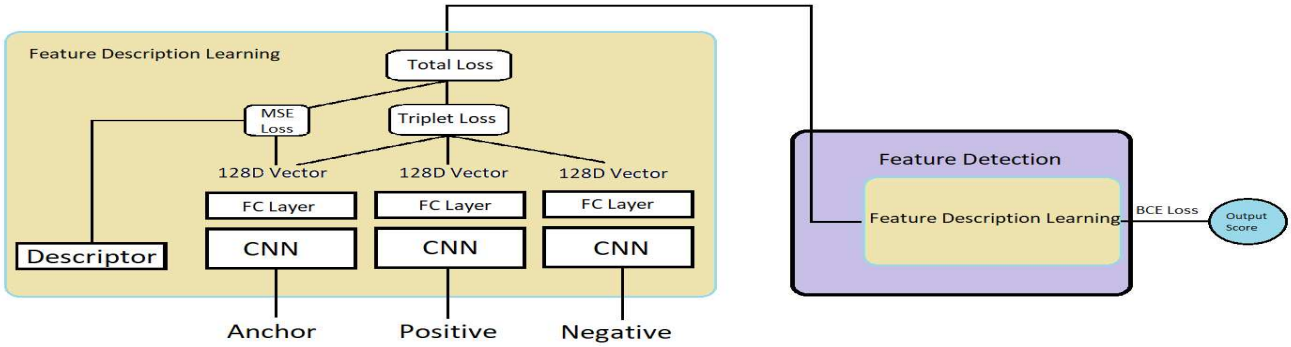

Fig. 2. Thermal image feature detection and extraction architecture. Input given are three image categories: Anchor, Positive and Negative images. RGB descriptor values are also provided to regress the anchor feature descriptor.

ture detection and description could be considered as a single step where midpoints for the image patch with a high feature embedding response is saved for feature matching purpose. However in order to acquire the high feature response, we introduce an intermediate step by adding fully connected layers to previously trained model weight values. The network is trained on 32x32 patches corresponding to features obtained from KAIST [10] and CSS [11] datasets. The network is able to learn high-quality feature descriptor for given patches and then classify good distinct features which can be identified very accurately through subsequent frames.

To summarize, the main contributions of our work are as follows: 1. We propose a novel Triplet based network to train robust feature detection on thermal scenes. 2. We propose a patch-based feature extraction network to learn 128-dimensional descriptor vectors to overcome the texture and context limitation of the thermal scenes. 3. We integrate both of the proposed detection and extraction networks into a full pipeline to enable stable and reliable feature matching on thermal images. 4. We achieve a superior performance in visual and quantitative comparison compared with other widely-used classical and most recent deep learning algorithms.

2. RELATED WORK

Classic feature detection and extraction methods. Keypoint detection and feature point matching using Canny [12] for edge, Harris [4] for corner detection and Histogram of Oriented Gradients (HOG) [13] had been typically done in the past and were used in applications such as recognition and image matching. Later SIFT was introduced, and due to its high robustness gave results with higher accuracy irrespective of the image scale, orientation, or rotation. SIFT localized and learnt good features using Difference of Gaussian (DoG) at multiple scales which made it more popular among the existing feature detection algorithms with the drawback being its higher computing time. In the later years, faster implementations similar to SIFT were introduced namely Features from Accelerated Segment Test (FAST), Binary Robust Independent Elementary Feature (BRIEF), Oriented FAST and Rotated BRIEF (ORB), Speeded up Robust Feature (SURF) which had their share of advantages and disadvantages compared to the SIFT extractor. ORB feature detector used multi-scale pyramid which were nothing but representations of the same image at different resolutions. Each level in the pyramid is a downsampled version of the image in the previous level.

Deep learning based detection and extraction methods. With the increase in popularity of deep learning-based methods, focus was shifted towards learning based. Patch based feature descriptor learning has been also implemented using Siamese network however most of the work has been done in the RGB domain. Faiz et al. [14] depicts a Siamese network trained for detection of change in satellite imagery with the network architecture containing two VGG16 [15] networks. PN-Net [16] took a Triplet based approach to generate descriptor which could be used in traditional matching setup. In contrast to Hinge Embedding loss [17, 18], they introduced SoftPN loss where the pairs of patches represented a soft negative mining. Another Siamese network L2-Net [19] specifically trained for descriptor learning from patches in Euclidean space showed state of the art performance. They had an all convolutional structure with a stride of 2 to achieve downsampling,and a loss function having three error terms.SuperPoint [17] used an encoder-decoder based approach having a shared encoder and two different decoders for description and detection of features. Having a good performance, it was limited to RGB images and failing to produce comparable results for thermal images. All the above-mentioned methods can be used for feature descriptor extraction but only on RGB/grayscale images. Our model can be considered as the first patch-based descriptor learning scheme designed for a more challenging thermal image dataset.

3. FEATURE DETECTION AND EXTRACTION ARCHITECTURE

To enable the network to learn meaningful features on thermal images, we make use of the Triplet Network which is trained on image pair patches extracted from visible-thermal image pairs from the KAIST and CSS datasets. The image

patches are selected such that they contain feature information which are robust to scale and illumination changes. The size of the extracted patches is set to 32x32 to encapsulate enough information to be identified as a strong feature. For selection of patches we make use of SIFT keypoints obtained on visible images and extract a 32x32 patch around it. Another patch with the same coordinates is extracted from the corresponding thermal image, thus allowing us with visible - thermal patches containing features. This method of extracting patches is not entirely accurate as there are many features that are observed in visible images which are absent in thermal images. To overcome this situation we further filter the image patch pair by taking into account the variance observed on thermal patches as shown in Fig. 3, thus only selectively choosing patches for finer training. We also make use of heavy data augmentations on these patches to add robustness.

Fig. 3. Patch selection criterion on the basis of variance measured.

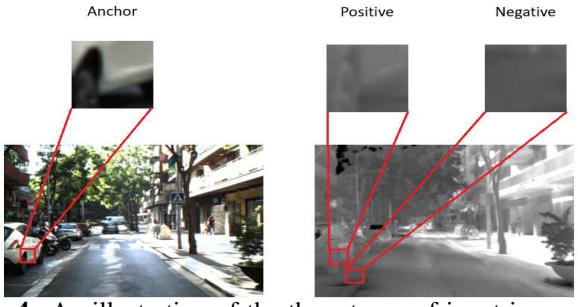

Fig. 4. An illustration of the three types of input images to the detection network: Anchor from RGB images; Positive and negative patches from thermal scenes.

For training Triplet Network, we require 3 images: Anchor, Positive Negative. The Anchor is the patch containing feature keypoint extracted from the visible image. The Positive is the thermal patch corresponding to the Anchor and the Negative is any patch but Positive. The Triplet network learns embeddings of the positive patches but also of the negative patches thus allowing us to accurately localize the keypoint.

3.1. Feature Description Network

The Triplet model requires three images (in our case patches corresponding to SIFT features) for learning the similarity between images. The Triplet network learns distributed embedding representation of data points where contextually similar data points are projected in the nearby region and dissimilar data points are projected far away from each other. We use SIFT feature to detect the keypoints and from those detected keypoints we extract 32 x 32 patches in the RGB and thermal image. **The Anchor(A) -** This patch corresponds to the detected SIFT feature in the RGB image. On obtaining the anchor patch, the keypoint for it is saved. **Positive(P) -** The saved keypoint location from the anchor patch is used. A 32x32 patch with the previously detected keypoint at the center is chosen as the positive patch. **Negative(N) -** For the negative patches we randomly generate keypoint locations and patches with the generated keypoint at the center are selected. Here an additional step is included where we implement patch selection based on the standard deviation of both RGB and thermal patch under consideration. Only those patches are selected which have a deviation value above a threshold indicating a good feature response. This additional step further helps in making the learning better and more robust. The selected patches are then fed into the Triplet network for feature description learning. The network consists of five convolutional layers followed batch normalization and ReLU activation function. Dropout is used to add regularization in both the models and finally two fully connected layers along with a Sigmoid function are used to output a 128-dimension vector.

Margin ranking loss with margin of 0.2 is used as the loss function which computes the distance between the input images trying to reduce distance between correct matches while increasing the other, with ADAM [20] having a learning rate of 0.001 and as the optimizer. The equation for loss is as:

$$L(A,P,N) = \max\{||f(A)-f(P)||^2-||f(A)-f(N)||^2+\alpha, 0\} \quad (1)$$

Along with the above-mentioned loss we also use a MSE constraint by training the anchor patches with their respective descriptor values as ground truth to improve the learning. We noticed that, by doing so the model learned features not only based on the pixel intensity values. The introduction of descriptor value causes the final 128-dimensional anchor feature vector to regress to the provided descriptor values. This further improves the model performance on low resolution dataset such as KAIST. The equation for loss is given as:

$$L_{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-y_{ij})^2 \quad (2)$$

And the overall loss function for the feature detection network is $L_{detection} = \lambda_1 L_{(A,P,N)} + \lambda_2 L_{MSE}$.

3.2. Keypoint Detection Network

The 128-dimensional output from the above learned Triplet architecture is connected to the keypoint detection network, which shares the same structure of the feature description network, thus allowing one-step detection and description. This step helps the network classify good and bad feature patches to further improve the feature description and detection. All the intermediate layers are frozen and two Fully connected

layers followed by ReLU activation function dedicated for detection are added. A Sigmoid activation is used at the end of the fully connected output. The input to this network is 32x32 thermal-thermal patch correspondences. The patches here include augmentation by scaling, flipping etc. Learning is performed on positive and negative patches producing an output score compared to 0 or 1. Binary cross entropy loss (BCE) along with Adam optimization is used for classification.

$$L_{det}(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (p(y_i)) + (1-y_i) \cdot \log(1-p(y_i)) \quad (3)$$

Once a patch with good feature is detected, mid points of the patch in the image coordinate are located and stored as keypoint. After the model is successfully trained and keypoints are obtained we use simple distance measurement amongst the output scores of descriptor values to find the corresponding matches in the two thermal frames. With the shared network and common layers, the feature detection and description can be much accelerated.

4. EXPERIMENTS

For training we use KAIST [10] and Cross-spectral Stereo dataset (CSS) [11]. These datasets are chosen to make our model more robust to low resolution (KAIST) as well as much high resolution (CSS) thermal patches. The image frames are selected such that they consist of unique features for efficient learning. This does cause reduction in the total number of images in the dataset but it is compensated by augmentation where the frames are flipped (horizontally and vertically), scaled and even jitter is added to the frames before training. The training dataset consists of 10,000 to 15,000 images and 32x32 sized patches are extracted from the images taking our total dataset count to more than 150,000 images. For testing we use around 1500 images with patches extracted from them in a similar manner as in training.

We compared feature description and matching from SIFT and deep network SuperPoint on different datasets. The results of which are shown in Fig. 5 and 6. It can be observed from Fig. 5 that our method is able to generate denser and more reliable matchings on the challenging thermal scenes, compared with classic SIFT algorithm and learning based SuperPoint. We also compare feature point detection for different techniques. From the feature detection result in Fig. 6, we can see that even though SIFT has a relatively good number of feature detection for the high resolution CSS dataset, it is unable to produce a comparable result for low resolution KAIST dataset. This is mainly because SIFT is designed for images with high textures and hence gives a poor detection.

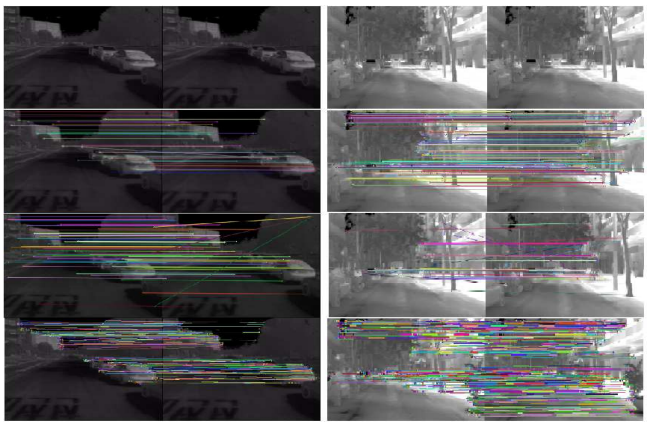

Fig. 5. Feature matching comparisons: Input images (Top), SIFT feature matching (second row), Superpoint matching (third row) and finally our model (Bottom).

	Keypoint Detected	Keypoint Matched
SIFT	KAIST - 121 CSS - 274	KAIST - 77 CSS - 218
Superpoint	KAIST - 97 CSS - 212	KAIST - 63 CSS - 163
Our Model	KAIST - 562 CSS - 778	KAIST - 405 CSS - 582

Table 1. Comparisons on detection and matching.

Besides the visual comparison, we also report the number of the detected keypoints and the matched keypoints from our trained model on the two datasets, compared with SIFT and SuperPoint in Table 1. It can be noticed that either for keypoint detection and the final matching stage, our pipeline is able to produce stable and promising features points.

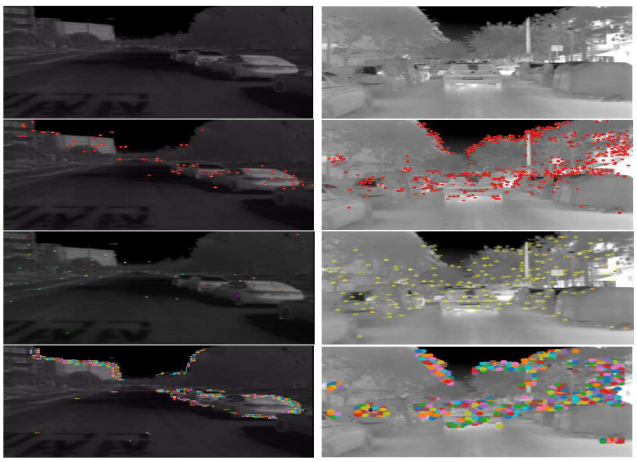

Fig. 6. Feature detection comparison. Top-bottom: Input Images; SIFT matches; Third row: Superpoint matches; Final row: Our method.

5. CONCLUSION

We propose a throughout feature detection and description network for thermal descriptor learning based on Triplet Siamese network, which designs an effective method for extracting descriptor values to be learned along with the intensity images to obtain much better feature extraction. Both the learning scheme and loss constraint demonstrate an effective solution compared to other available methods. Our method is easy to implement to be used in practical applications.

6. REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[3] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[4] Konstantinos G Derpanis, "Harris corner detector," *York University*, pp. 1–2, 2004.

[5] Josep Aulinas, Marc Carreras, Xavier Llado, Joaquim Salvi, Rafael Garcia, Ricard Prados, and Yvan R Petillot, "Feature extraction for underwater visual slam," in *OCEANS 2011 IEEE-Spain*. IEEE, 2011, pp. 1–7.

[6] Jan Hartmann, Jan Helge Klüssendorff, and Erik Maehle, "A comparison of feature descriptors for visual slam," in *2013 European Conference on Mobile Robots*. IEEE, 2013, pp. 56–61.

[7] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2548–2555.

[8] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[9] Samuele Salti, Alessandro Lanza, and Luigi Di Stefano, "Keypoints from symmetries by wave propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2898–2905.

[10] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *CVPR*, 2015, pp. 1037–1045.

[11] Julien Poujol, Cristhian A Aguilera, Etienne Danos, Boris X Vintimilla, Ricardo Toledo, and Angel D Sappa, "A visible-thermal fusion based monocular visual odometry," in *Robot 2015: Second Iberian Robotics Conference*. Springer, 2016, pp. 517–528.

[12] John Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 679–698, 1986.

[13] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 886–893.

[14] Faiz Rahman, Bhavan Vasu, Jared Van Cor, John Kerekes, and Andreas Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 958–962.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[16] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk, "Pn-net: Conjoined triple deep network for learning local image descriptors," 2016.

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[18] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.

[19] Yurun Tian, Bin Fan, and Fuchao Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 661–669.

[20] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.