Image Feature Matching Based on Deep Learning

Yinyang Liu
Beijing Advanced Innovation Center for Future Internet Technology
Beijing University of Technology
Beijing, China
e-mail: liuyinyang09@qq.com

Xiaobin Xu *
Beijing Advanced Innovation Center for Future Internet Technology
Beijing University of Technology
Beijing, China
e-mail: xuxiaobin@bjut.edu.cn

Feixiang Li
Beijing Advanced Innovation Center for Future Internet Technology
Beijing University of Technology
Beijing, China
e-mail: lifeixiang126@126.com

***Abstract*—Image feature matching is an integral task for many computer vision applications such as object tracking, image retrieval, etc. The images can be matched no matter how the image changes owing into the geometric transformation (such as rotation and translation), illumination, etc. Also due to the successful application of the deep learning in image processing, the deep learning method has an advantage in feature extraction of images. In this paper, we adopt a deep Convolutional neural network (CNN) model, which attention on image patch, in image feature points matching. CNN obtains the feature by convolution kernel which parameters are achieved by learning. So it has strong ability to express feature. Compared with other methods, experimental results indicate the proposed method has higher accuracy and completed efficiently.**

Keywords-image matching; convolutional neural network; deep learning

I. INTRODUCTION

Image matching is defined as judging the similarity by analyzing the similarity and consistency between images. Image matching has a wide range of applications in many fields, such as image recognition, 3D modeling, target recognition, image stitching, image retrieval, etc.

Image matching methods generally are divided into two broad categories: gray-based image matching and feature-based image matching. The matching based on gray image information is simple and accurate, but it is weaker for image changes such as nonlinear deformation, illumination and scale change. In feature-based image matching, the image features are extracted and the features are quantified by some mathematical means. Matching by this method has a great important relationship between the accuracy of matching and the selection of features. The higher the robustness of the feature is, the higher the correctness of matching is. The traditional feature points matching algorithm, such as the feature extracted by Scale-invariant feature transform (SIFT) algorithm [1], which has rotation invariance and translation invariance. This algorithm is suitable for image matching with different scales. It has high robustness and high computational complexity. Therefore, the speed of calculation is slow and cannot meet real-time requirement. Based on the SIFT algorithm, Speed Up Robust Features (SURF) algorithm [2] is improved the calculation speed, but the robustness of its features is reduced, and the effect of large-scale image matching is generally satisfactory.

With the successful application of deep learning [3-6] in the field of image processing, image feature matching, extracted by Convolutional Neural Network (CNN), achieves better results than traditional methods. With the development of CNN, a series of network structures, e. g. Alexnet [7], VGG [8], Resnet [9], etc., have been developed. In recent research, it has been shown that using CNN to extract image features can improve image matching accuracy. Reference [10] trained a piecewise linear regression to detect invariant feature points of outdoor pictures with dramatic changes because of illumination and weather. References [11][12] extracted feature through the image patch training based Siamese network and matching by similarity measure. However, because the image patch is too small, the size and robustness of the network are limited. Also, Reference [12] focused on the difference of the images in appearance. References [13-15] compensated for the lack of image matching performance in the big dataset by the following methods: increasing the negative example to train the network, increasing the robustness of the network, and matching the correlation between the positive and negative examples. Reference [16] focused on the improvement of the triplet loss function, and introduced a novel algorithm to train the CNN with weakly-labeled datasets. In reference [17], a hybrid network is proposed for feature point detection, direction estimation and key point description in the network. However, it is difficult to train and the speed of feature points detection is slow in large-scale data set. The reference [18] proposed a data-driven descriptor that matched in

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

Euclidean space, also modified the loss function to train the network. On the basis of [18], reference [19] modified the loss function and the descriptor outperforms in the image match with extreme condition, and also can be used to image retrieval. In this paper, a deep CNN model is used to describe the feature points of matching the image pairs. The CNN model in this paper uses triplet loss function for network training, and its input comes from the level of image patch.

The rest of this paper is organized as follows: the next section briefly describes CNN model. Image Matching based on CNN model is presented in detail in section 3. Experimental results about this algorithm and discussion are presented in section 4. In section 5, the conclusion is given.

II. CNN MODEL

In recent years, the CNN model has achieved great success in image processing. The main characteristic of CNN is that features are obtained by learning the convolution kernels in each convolutional layer. The input image is obtained image features by multi-layer convolution kernel operations, down sampling, etc. operations. The training of the model is to obtain the predictive information by transmitting the layer information of the convolutional layer. Through the back propagation, the loss between the predicted value and the real value is transmitted back, and the partial derivative of the parameter of each layer is passed through the loss function. The parameters of each layer are updated by using the gradient descent algorithm. Through continuously learning and updating of parameters, the network has a strong ability to express the image. In this paper, we adopted the structure of the deep convolutional neural network. Through the multi-layer convolution network, the extraction of image feature description is more accurate. To enhance the robustness of the network, the network used the triple loss function to train the network and the method to update the parameters is stochastic gradient descent algorithm.

A. The Structure of CNN

We adopt a deep CNN structure [18-19] (see Fig.1) to process the image patch which represented the image feature point, and obtain the feature point description of the image. As shown in Figure 1, the structure of CNN contains seven convolution layers which consist of six convolutional layers with a 3*3 kernel and one convolutional 8*8 layer, separated by Batch Normalization and ReLU activations.

The purpose of Batch Normalization (BN) is to prevent the convolutional neural network from gradually shifting or changing the distribution of data input to the active layer during neural network training, so that in the case of back propagation, the gradient of the network disappears, resulting in slower network convergence during training, that is, the layer can accelerate convergence. At the same time, the problem that the convolution network is insensitive to the initialization weight can be reduced, and the over-fitting problem can be controlled.

$$\mathrm{F(x)} = \gamma * \frac{x\text{-}E(x)}{\sqrt{\sigma^2+\epsilon}} + \beta \qquad (1)$$

where $\mathrm{E(x)} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i\text{-}E(x))^2$, γ and β is the training parameters.

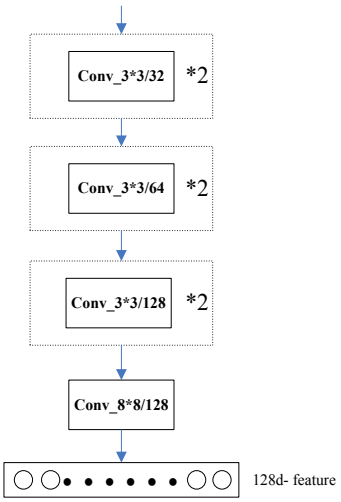

Figure 1. The structure of deep CNN model.

Rectified Linear Unit (ReLu), which is an activation function that performs nonlinear operations on the feature data processed by batch normalization. The expression of the function is shown as follows:

$$\mathrm{F(x)} = \max(0, \mathrm{x}) \qquad (2)$$

The Relu function has unilateral suppression, which makes the CNN sparsely activating, can better mine image features, and fit training data, and has strong expression ability.

The Dropout layer, which appears after the sixth layer of ReLu, aims to reduce the over-fitting problem of the network while reducing the coupling between different parameters. Since this convolutional neural network structure uses the BN layers which can also reduce the over fitting problem, the Dropout layer is used only once.

B. The Loss Function

Triplet loss function [20] is utilized to train the model. The input data include the positive pair and negative pair patches. We generate the positive pair patches form the different patch of the same 3D points. As the negative patch pair, patch is selected randomly in other image. These three patches are set as a group during the training process, the group is input into the same network at one time and get 3 sets of features x_i^a, x_i^p, x_i^n. The loss is calculated by the following formula.

$$L = \arg min \sum_{i}^{N} (||x_i^a\text{-}x_i^p||_2^2 - ||x_i^a\text{-}x_i^n||_2^2 + threshold)_+ \qquad (3)$$

where $(\mathrm{x})_+ = \max(0, \mathrm{x})$, $||\cdot||_2$ is the L_2 norm.

III. IMAGE MATCHING BASED ON CNN MODEL

The feature-based matching of the images includes two parts: feature detection and matching. The feature detection requires two steps. First, we need to get the local feature

points and then describe them. The traditional detection of feature points algorithms are SIFT, SIRF, Harris, etc. SIFT feature points are chose to be the feature points detector, because SIFT features are insensitive to changes in scale, rotation and brightness. The second step is the description of these feature points for matching calculations. Because the input of the deep CNN model is the image patch, we centering on the feature point to crop an image patch to represent the feature points that can be the input of the model. Through the deep CNN model, we get the feature description and match a pair of images by them.

A. *Feature Points Detection*

SIFT detectors is utilized to achieve feature points. For SIFT feature point detection, it is necessary to establish the image pyramid. It is obtained by the difference of Gaussian scale space, which is shown as follows:

$$\mathrm{L}(x,y,\sigma) = G(x,y,\sigma)*\mathrm{I}(x,y) \tag{4}$$

$$\mathrm{D}(x,y,\sigma) = L(x,y,k\sigma)\text{-}L(x,y,\sigma) \tag{5}$$

where $\mathrm{G}(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$. Employs the image pyramid to find extreme point as feature points. In these points, there are some low contrast points and some unstable edge response points should be eliminated.

B. *Feature Points Description*

In the traditional method, such as the feature descriptor of SIFT, its composition is complex, computationally intensive and takes a long time. Deep CNN model is used to achieve the feature points description. The structure of the deep CNN model is shown in Figure 1.

By intercepting the image patch centered on the feature point (see Fig. 2) and turning it into a grayscale image, an image patch of 32*32*1 is obtained. This image patch is taken as an input of the trained model. The patch performs multiple times convolution, normalization, finally generate a 128-dimensional feature description to represent the image points, which represents the description of the feature point.

Figure 2. Example of take the feature points as center to cut the image patches. Left is the image and right is the image patches of some feature points.

C. *Image Match by Feature Points*

The image pairs are matched by using KNN algorithm. After processing by above two steps, two kinds of information can be obtained for each picture. One is coordinate information about the feature points, and the other is the feature description corresponding to the feature points. The feature descriptions are used to establish the KD-tree. Through comparing the feature descriptions with KD-tree of the two images, the corresponding feature points in the two graphs are found.

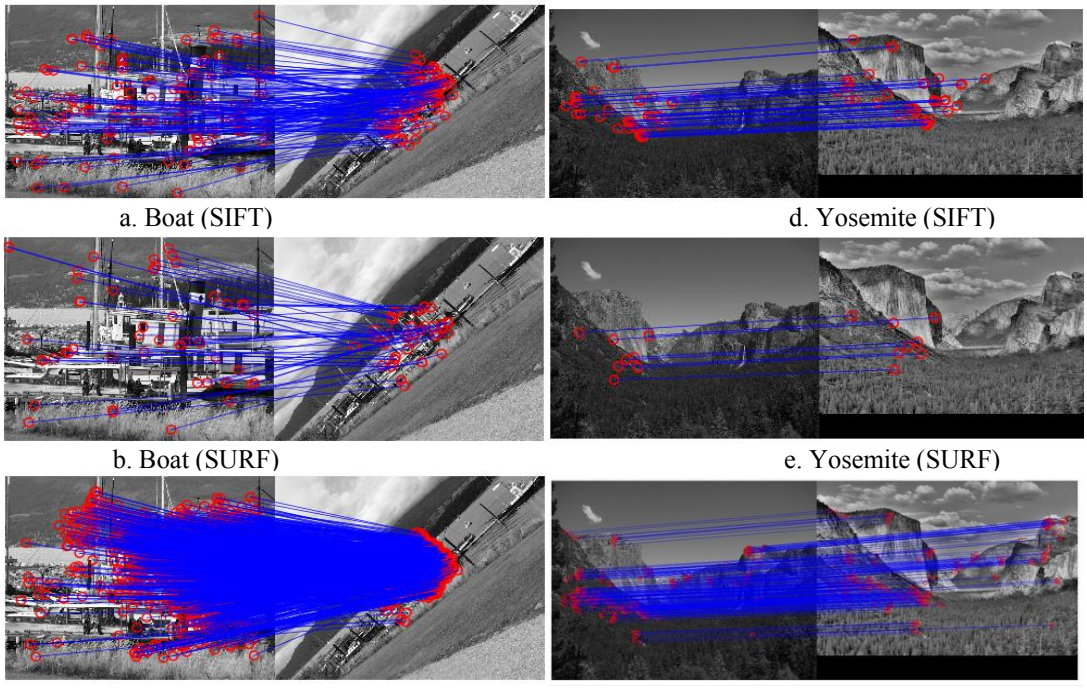

Figure 3. Two examples of feature-based Image match with three different methods. The red circle is the feature point, and the blue line is the corresponding points connection in the image pair. Top row: SIFT, second row: SURF, third row: Our method.

There is some noise in corresponding feature points of the image pair. The algorithm of a random sampling consensus (RANSAC) is used to reject abnormal data. When the match of the images is performed, it will generate some mismatched points. The algorithm can eliminate these mismatched points and optimize the correct matching effect.

The overall algorithm flow of our method can be see in Table I

TABLE I. THE ALGORITHM FLOW

Algorithm: ***Image Feature Matching based on deep Learning***
Input : One pair of images (I1, I2)
1: Extracting feature points sets (A1, A2) from (I1, I2)
2: Centering on feature points to Capture image patch sets (B1, B2), which patch size is 32*32*1
3: Through the model to obtain a 128-dimensional feature description set (C1, C2).
4: Use KD-Tree to find the corresponding feature points, represented by matrix G
5: Remove mismatch points from G by RANSAC, G'
Output: G'

IV. EXPERIMENT

We used UBC dataset to train the model and through the W1BS dataset to test. Compared with our method, we used two methods that are SIFT and SURF.

A. Datasets

UBC Travel Photo dataset [21] collected by Winder et al., which contains three scenes: The Statue of Liberty, Half Dome in Yosemite, Notre Dame. It is a standard dataset and the total number of the image patches are more than 1.5 million, which are suitable for discriminating descriptor and evaluation. These three datasets contain 100k, 200k and 500k label pairs, which correspond to 3D points.

W1BS (Wide baseline stereo) dataset [22], which contains 40 pairs of images, each pair of images comes from the following four ways: appearance, geometry, illumination, sensors. The image pairs difference because of seasonal or weather change, occlusions; the position of scale, camera and object; intensity, wavelength of light source; sensors, etc.

B. Results

Figure 3 shows the example of image matching results for SIFT (Figure 3 a, d), SURF (Figure 3 b, e) and our method (Figure 3 c, f). The image pairs are different from geometry and appearance. As expected, our method has more correct correspondences across the image pair. The method of SURF has the worst performance because it cost the accurate to speed the calculation.

We analyze our result with the number of successfully matched image pairs and against the two methods of SIFT and SURF (see Tab. II). All of these methods have good performance in the difference of illumination. The methods of SIFT and SURF are sensitive to the geometry and appearance. The SIFT method has slightly better than SURF. But our method is insensitive to the geometry and appearance. All of three methods are not good at distinguish the images form sensors. Overall, our approach is better.

V. CONCLUSION

In this paper, we add the deep CNN model in feature-based image matching process that can learn feature which have strong express. In the end, we compare two feature-based image matching approaches to show that our method is insensitive to the illumination, geometry, appearance. There are some insufficient of the method, such as matching between images from different sensors. Also, the model is the part of the image matching, we can improve in the feature.

TABLE II. THE NUMBER OF SUCCESSFULLY MATCHED PAIRS

Methods/num	**SIFT**	***SURF***	***OURS***
Illumination(L)/11	9	8	10
Geometry(G)/10	4	3	9
Appearance(A)/6	5	3	6
Sensor(S)/7	3	1	4
Total/34	21	15	29

REFERENCES

[1] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints// International Journal of Computer Vision. 2004:91-110.

[2] Bay H, Tuytelaars T, Gool L V. SURF: speeded up robust features// European Conference on Computer Vision. Springer-Verlag, 2006:404-417.

[3] Noh H, Araujo A, Sim J, et al. Large-Scale Image Retrieval with Attentive Deep Local Features// IEEE International Conference on Computer Vision. IEEE Computer Society, 2017:3476-3485.

[4] Mishkin D, Matas J, Perdoch M, et al. WxBS: Wide Baseline Stereo Generalizations. 2015.

[5] Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization// Computer Vision and Pattern Recognition. IEEE, 2016:2921-2929.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks// International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[8] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. 2015:770-778.

[10] Verdie Y, Yi K M, Fua P, et al. TILDE: A Temporally Invariant Learned DEtector// Computer Vision and Pattern Recognition. IEEE, 2015:5279-5288.

[11] Han X, Leung T, Jia Y, et al. MatchNet: Unifying feature and metric learning for patch-based matching// Computer Vision and Pattern Recognition. IEEE, 2015:3279-3286.

[12] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:4353-4361.

[13] Simo-Serra E, Trulls E, Ferraz L, et al. Discriminative Learning of Deep Convolutional Feature Point Descriptors// IEEE International Conference on Computer Vision. IEEE Computer Society, 2015:118-126.

[14] Balntas V, Johns E, Tang L, et al. PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors. 2016.

[15] Balntas V, Riba E, Ponsa D, et al. Learning local feature descriptors with triplets and shallow convolutional neural networks// British Machine Vision Conference. 2016:119.1-119.11.

[16] Markuš N, Pandžić I S, Ahlberg J. Learning local descriptors by optimizing the keypoint-correspondence criterion// International Conference on Pattern Recognition. IEEE, 2017:2380-2385.

[17] Yi K M, Trulls E, Lepetit V, et al. LIFT: Learned Invariant Feature Transform. 2016:467-483.

[18] Tian Y, Fan B, Wu F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:6128-6136.

[19] Mishchuk A, Mishkin D, Radenovic F, et al. Working hard to know your neighbor's margins: Local descriptor learning loss. 2017.

[20] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. 2015:815-823.

[21] UBC dataset :http://phototour.cs.washington.edu/patches/default.htm

[22] Lin G, Milan A, Shen C, et al. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:5168-5177.