EasyChair Preprint

№ 4586

A Comparative Study for Fisheye Image Classification: SVM or DNN

Zhen Chen and Anthimos Georgiadis

November 16, 2020

A Comparative Study for Fisheye Image Classification: SVM or DNN

Zhen Chen[1][0000-0002-7907-6547] and Anthimos Georgiadis[2][0000-0003-4092-8834]

[1 2] Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany
zhen.chen@leuphana.de

Abstract. The comparison between the feature-based method and the learning-based method is conducted in the training time, the accuracy and the generalization capacity, to address the optimisation for the multi-style fisheye imagery classification. We construct an srd-SIFT descriptor based SVM classifier to present the feature-based method for describing the influence of the dataset scale and the visual word scale on the classifier. The SVM classifier achieves 15.98% accuracy on the test set after 162 hours training, with the condition that includes 800 images per class in 12 classes and 1500 visual words. For the learning-based method, we propose to expand training samples' style variety, via style transformation, to facilitate the contemporary architecture retraining. Following this approach, we retrain the ResNet-50 by an artificial multi-style fisheye image dataset without complementing new training labels. The performance of the obtained ResNet classifier is evaluated on 6000 images collected in the real-world. The result shows that the retrained classifier has great generalization capacity and reaches 97.19% top-3 accuracy.

Keywords: Fisheye image, Super vector machine, DNN, Style expansion.

1 Introduction

In the last few years, cameras with the ultra-wide field of view are integrated into plenty of automation scenes, like automated inspection, unmanned driving vehicle and virtual reality [1]. Owing to the compact optical structure of the fisheye lens, the camera system provides flexible applications in limited workspace. However, the equirectangular image obtained from a fisheye camera suffers from high compression ratios that appear intensely at the edge [2] . The previous research concentrates on omnidirectional image unwrapping algorithms which encourage reusing mature achievements on the perspective image [3]. Due to the extra computing cost from the unwrapping process, research focus is shifting to process the raw fisheye image.

The feature-based method and the learning-based method are two mainstream research aspects of the classification issue. By the feature-based method, a feature point with scale, shift and rotation invariant becomes a unique feature descriptor on image's representation. The classifier can be trained with the aid of the Bag-of-Visual-Words (BoVW) model for restructuring an image's representation [4]. Different to the fea-

ture-based method, the learning-based method learns substantial characteristics of training samples by layers of artificially constructed neural units. In this paper, we compare two methods for the multi-style fisheye imagery classification. Contrast to the rectilinear image classification task and the regular omnidirectional image processing procedure, our work focuses on samples' geometric models that are an unknown or are mixed with a variety of geometric models. In tests, we adopt the SVM classifier and the Deep Residual Networks (ResNet) as representations of these two methods. The comparison focuses on their performance in terms of the classifier's training speed, the accuracy and the generalization capacity. The artificial fisheye image dataset implemented in experiments is from our previous work [5]. In addition, we give a further test of the achieved model on real-world fisheye images.

2 Related work

A hand-crafted feature descriptor determines the uniqueness of an image's description, which exhibits great performance for perspective images in the long-term development. Even though feature descriptors SIFT, SURF, BRIEF and ORB are widely used in the image recognition and matching algorithm, their performance on the omnidirectional image is inadequate [6]. For the omnidirectional image matching, Peter Hansen et al. proposed pSIFT that maps the omnidirectional image and spherical Gaussian function to an intermediate stereographic image for approximating the diffusion on the sphere [7]. Cruz-Mota et al. employed a spherical Gaussian filter in spherical SIFT to perform the Gaussian smoothing [8]. In 2012, the srd-SIFT was demonstrated with the radial distortion improvement on the original SIFT framework. Compared to other existing omnidirectional descriptors, the srd-SIFT relies less on camera's calibration parameters [9].

The NN architecture designed for the fisheye image classification is limited. Jeon et al. introduced an active convolution unit to learn position parameters for defining more diverse forms of receptive fields [10]. Coors et al. proposed to dynamically wrap the convolutional kernel sampling pattern around the sphere's surface according to the sampling location [11]. It lessens image distortion by the perspective convolutional neural network (CNN), especially at borders and poles of the equirectangular projection. Khasanova et al. treated each omnidirectional image as a weighted graph signal representation, and extended convolutional neural network on the processing of graphs [12]. However, it was proved just with several specific terms of mapping projections. Above research builds end-to-end training for the omnidirectional image but focuses only on single-style image. Our research targets the classification problem under the multi-style fisheye image. Each object's images achieved from different distances and angles is treated as its non-single representations. If training samples cover variety of representations of an object, existing NN architectures can be reused.

The debate on adopting two different methods continues. KIM et al. showed that SVM classifier outperformed KNN classifier on their Caltech-4-Cropped dataset [13].

Liu et al. had a comparative study between the SVM and the Stacked Auto-Encoders (SAE) on the remote sensing image classification [14]. The comparison showed that SVM takes less time on the small-scale dataset, but SAE can be implemented in parallel clusters. SAE did not show competitive performance on accuracy and noise immunity. In this paper, a further study between two methods on the performance of the fisheye imagery is conducted.

3 Experimental Framework

3.1 srd-SIFT descriptor based SVM classifier

We implement a three-step process to build a feature-based classification architecture. First, selected representative feature descriptors construct feature vectors as an image's description. Then, feature vectors achieve further dimension scale-down and are also reorganized in the BoVW model. Finally, a classifier is trained with local descriptors from the BoVW model. In the prediction phase, feature vectors from a test image are assigned to the trained classifier for a classification output.

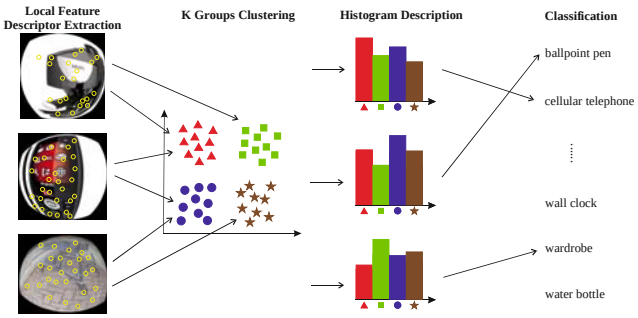

Fig. 1. The framework of the srd-SIFT descriptor based SVM classifier

The srd-SIFT enhances the SIFT algorithm in aspects of the repeatability and the effectiveness on the radial distortion for the omnidirectional image matching. The algorithm processes the original image plane, instead of resampling the image signal. In contrasts to pSIFT and mdBRIEF [15], srd-SIFT has lower accuracy requirement for camera's calibration parameters. Due to high dimensions of achieved srd-SIFT feature descriptors, K-means clustering is adopted to decrease the dimensions. The BoVW is implemented according to replace the concept 'words' by clusters achieved from the K-mean clustering. The SVM solves the multi-class classification problem by simplifying the problem into a set of binary classification problems. We perform one-versus-one SVM strategy by LIBSVM [16].

As shown in Figure 1, sample images are sent to an srd-SIFT feature extractor. Extracted feature descriptors buildup an Nx128 dimension feature vector. For decreasing the dimension of feature vectors, the vectors are clustered to K classes under the un-

supervised learning. To analyse clusters' spatial histogram distribution, feature words are constructed by a BoVW model. Each image can be described by M visual words. Then, we train the SVM classifier with each image's BoVW expression.

3.2 DNN classifier

DNN architecture is mainly consisted of the convolutional layer, the pooling layer, the fully-connected layer and a series of active functions. The convolutional layer is located on the top layer of a CNN, which extracts features from an input image based on shared weights. It builds the connection between the input volume and the output feature map when the filter constructs convolution operations with the receptive field on each stride. The pooling layer is normally inserted between successive convolutional layers, to reduce the spatial dimension of the representation. It decreases the quantity of parameters in feature maps, but maintains the same depth between the input channel and the output channel. The fully-connected layer is the last output layer of the DNN architecture, which classifies the input image into various classes as pre-set in training dataset. Each neuron in a fully-connected layer has global activations from the previous layer for receiving signals from entire feature maps. High-level features from convolutional layers and pooling layers are reconstructed and activated by the softmax activation function in the output layer.

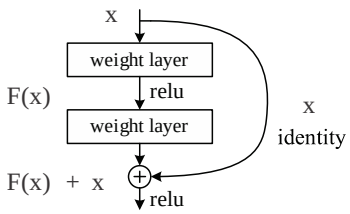

Fig. 2. Residual block building

Residual block

ResNet lets the stacked layers fit a residual mapping instead of a desired underlying mapping, which is assumed to be easier to optimise than the original mapping [17]. The expression of the short connection in a residual mapping is shown in Figure 2. Considering x as the inputs of the first layer and $H(x)$ as an underlying mapping by a few stacked layers, the residual function can be defined as:

$$F(x)=H(x)-x \tag{1}$$

In which the input and the output are the same dimensions. Then a residual unit is defined as:

$$y=q(F(x, w_i)+w_s x) \tag{2}$$

Where y is the output of the residual unit, w_i is a set of weights, w_s is a linear projection matrix for matching dimensions, $F(x, w_i)$ represents the residual mapping to be

learned, and q denotes ReLU. For a residual unit that contains two convolutional layers, it is simplified as:

$$F(x, W)=W_2\, q(W_1\, x) \tag{3}$$

4 Tests and Results

4.1 Training Environment

Tests are executed on Dell Precision T5500 workstation which is equipped with dual Intel Xeon Processor E5645, 23GB RAM and an MSI Geforce GTX 1080 graphic card. The SVM classifier and the ResNet classifier are trained by MATLAB 2018a on CPU and Deep Learning Library Caffe [18] on GPU separately. The artificial fisheye image dataset includes 167250 synthetic fisheye images in 12 classes, which are transformed by the principle of the equidistant projection.

4.2 SVM Classifier Training and Results

The artificial dataset owns over 10K images on each class. Using the whole dataset to train a 12 classes SVM classifier is extremely expensive. Therefore, we prepare three sample groups which have 300, 500 and 800 random images on each class separately for training and test. Valid images for each sample group reach 3600, 6000 and 9600 where the ratio between the training set and the test set is 8: 2. To understand the influence of the visual word scale on the BoVW model, 600, 1000 and 1500 different quantities of visual words are test for each dataset. Here the visual word scale has the same value as the K value in clustering. Test results are listed as Table 1, 2 and 3.

Table 1. The prediction accuracy for training sets

	the visual word scale		
img/class	600	1000	1500
300	79.65%	87.56%	93.33%
500	61.75%	78.83%	84.47%
800	53.54%	67.16%	77.73%

Table 1 reflects whether the trained model fits training samples. As shown, the models' accuracy increases when the visual word scale increases, and receives the best achievement at the visual word level 1500. It achieves 93.33% on the training set that includes 300 images per class, with accuracy boosts of 5.77% and 13.68% on the word level 1000 and 600. However, concerning only the word level 1500, the SVM classifier's prediction accuracy drops dramatically to 77.73% when the image quantity increases to 800 images per class.

Table 2. time cost for each classifier's training

	the visual word scale		
img/class	600	1000	1500
300	7 h	12 h	20 h
500	20 h	37 h	59 h
800	59 h	102 h	162 h

With the increase in the number of training samples and visual words, training a classifier becomes more and more expensive, rising from 7 hours to 162 hours (as shown in Table 2). We also give classifiers' test results. As shown in Table 3, classifiers have a slight improvement when a model is trained under a larger dataset at the same word scale level. In contrast, for the same training set, the enlargement of the visual word scale weakens the model's generalization capacity.

Table 3. prediction accuracy for test sets

	the visual word scale		
img/class	600	1000	1500
300	15.41%	13.61%	12.5%
500	19.16%	17.16%	15.41%
800	19.89%	18.33%	15.98%

4.3 DNN Classifier Training and Results

We employ 119040 images for the training, 14880 images for the validation and 14880 images for the test separately. Since the fisheye image dataset is originated from the perspective dataset that sharing same objects, weights from the pre-trained model are transferred to new model using transfer learning technology.

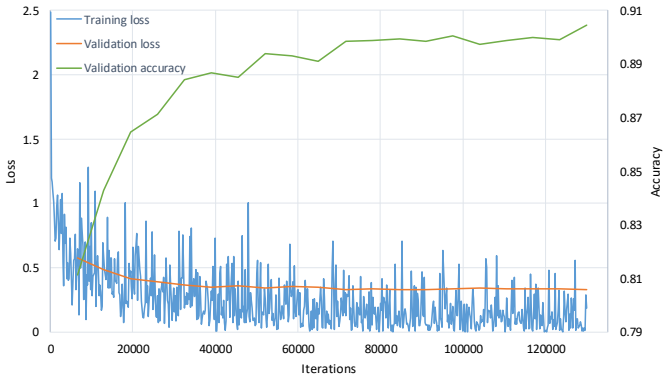

Fig. 3. The fine-tuning model process

In practice, we implement the Stochastic Gradient Descent optimizer with 0.0001 basic learning rate. The multi-step strategy is used with the gamma value 0.5. The training process is shown in Figure 3. After around 12 hours fine-tuning, the new model achieves 90.34% accuracy on the validation set and 89.98% accuracy on the test set, as shown in Table 5. It presents consistent accuracy between the validation set and the test set. The great performance on the test set shows the model's generalization capacity on the unknown data.

Table 4. The prediction accuracy of the CNN classifier on the validation set and the test set

Group number	Validation accuracy	Correct samples	Test accuracy	Correct samples
ballpoint pen	92.66%	(1149/1240)	92.18%	(1143/1240)
cellular telephone	88.06%	(1092/1240)	86.77%	(1076/1240)
desktop computer	85.48%	(1060/1240)	86.13%	(1068/1240)
espresso maker	92.98%	(1152/1240)	93.63%	(1161/1240)
printer	86.45%	(1072/1240)	84.84%	(1052/1240)
projector	85.40%	(1059/1240)	84.35%	(1046/1240)
shopping cart	92.90%	(1152/1240)	93.95%	(1165/1240)
stone wall	98.39%	(1220/1240)	97.90%	(1214/1240)
television	86.53%	(1073/1240)	86.53%	(1073/1240)
wall clock	92.82%	(1151/1240)	92.02%	(1141/1240)
wardrobe	94.84%	(1176/1240)	94.44%	(1171/1240)
water bottle	87.66%	(1087/1240)	87.02%	(1079/1240)
Overall	90.34%	(13443/ 14880)	89.98%	(13389/ 14880)

4.4 Physical World Image Test

Although the structural design of a fisheye lens concerns different optical parameters, from the view of the human vision, the distinction of different fisheye lenses reflects in the field of view (FOV) and projected images' distortion status. Therefore, we test the performance of the achieved model on real-world images.

In this test, 12 classes objects images are captured by the fisheye camera system which is equiped with Fujifilm Fujinon F-FE185C057HA-1 lens (focusing range: 0.1 m to infinity; viewing angle: 185 degrees on the 2/3-inch sensor) and Sony Alpha 6000 camera. Two-step strategies are adopted to prepare samples for the evaluation. Firstly, we shoot a video of an object along the vertical, the horizontal and the radial direction on the resolution 1920 x 1080 with 24 frames per second in AVCHD format. Secondly, we split each video into separate frames and crop out the black border. As a

result, images that contain an object's continuous deformation can be obtained as shown in Figure 4. The resolution of the cropped image is 500x500 pixels.

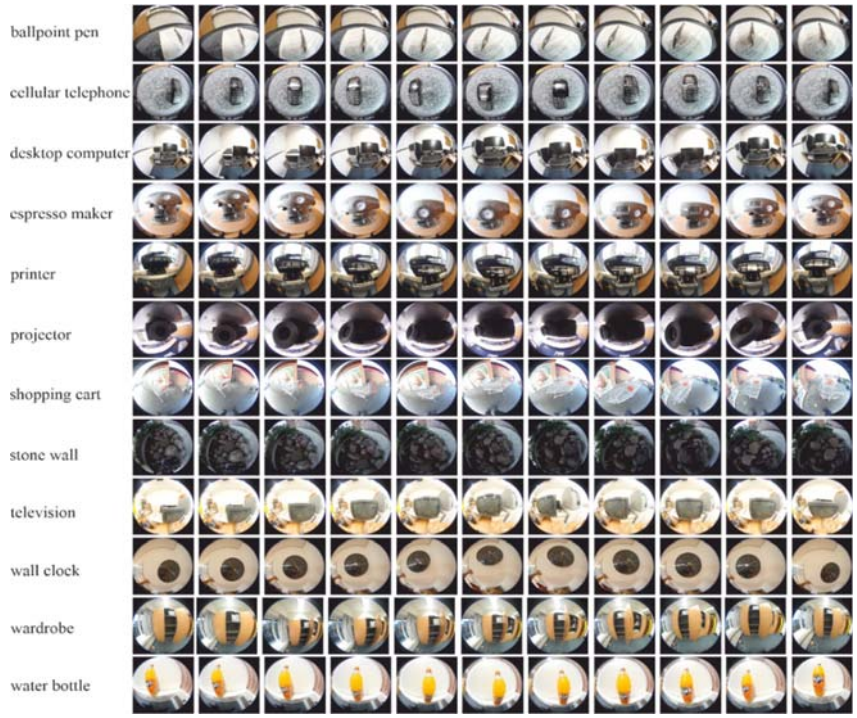

Fig. 4. Parts of collected real-world images for each class

During image collecting procedure, there is no proper scene for the stone wall. Instead, we replace this scene with the cobbled decoration which presents similar morphological feature as the stone wall in this situation. Finally, we combine 500 images for each class to form a new test set (6000 images totally). The dataset covers continuous object deformation on different angles and distances. All obtained images are labelled and sent to the trained classifier. The prediction results are shown in Figure 5. As illustrated in the confusion matrix, most of the test classes reach 100% prediction accuracy, except ballpoint pen (87.92%), printer (87.17%) and television (92.26%). The overall accuracy of the test set is 97.19%. The result proves that enlarging the style of training set can effectively improve a model's generalization capacity.

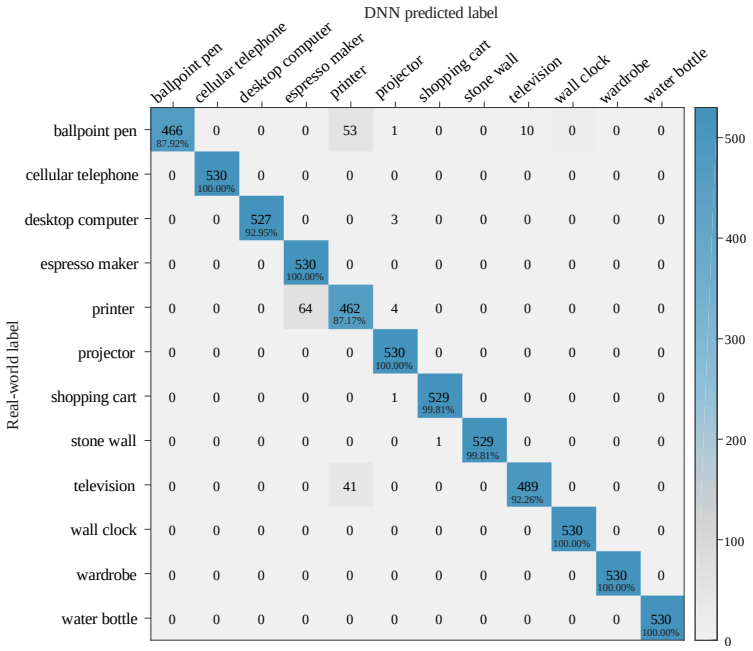

Fig. 5. Confusion matrix for the predictions of the DNN versus the real-world label

5 Conclusion

In this paper, we compare the performance of the srd-SIFT descriptor based SVM classifier and the ResNet on the multi-style fisheye imagery classification. The ResNet shows a great advantage on the training time and the accuracy in the large-scale dataset processing. The visual word scale of the BoVW model deeply influences the SVM classifier's accuracy, especially on the small-scale dataset. From the aspect of the generalization, the diverse simulated camera parameters lead to the fail of the SVM classifiers on the test set. However, the ResNet classifier demonstrates high prediction accuracy on the test set and real-world images.

References

1. Davies, E.R.: Machine vision: theory, algorithms, practicalities. Elsevier (2004).
2. Sarkar, M., Brown, M.H.: Graphical fisheye views of graphs. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 83–91 (1992).
3. Jung, H.G., Kim, D.S., Yoon, P.J., Kim, J.: Structure analysis based parking slot marking recognition for semi-automatic parking system. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). pp. 384–393. Springer (2006).

4. Yang, J., Jiang, Y.-G., Hauptmann, A.G., Ngo, C.-W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Workshop on multimedia information retrieval. pp. 197–206 (2007).
5. Zhen, C., Georgiadis, A.: Parameterized Synthetic Image Data Set for Fisheye Lens. In: 2018 5th International Conference on Information Science and Control Engineering (ICISCE). pp. 370–374. IEEE (2018).
6. Karami, E., Prasad, S., Shehata, M.: Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. arXiv Prepr. arXiv1710.02726. (2017).
7. Hansen, P., Boles, W., Corke, P.: Spherical diffusion for scale-invariant keypoint detection in wide-angle images. In: 2008 Digital Image Computing: Techniques and Applications. pp. 525–532. IEEE (2008).
8. Cruz-Mota, J., Bogdanova, I., Paquier, B., Bierlaire, M., Thiran, J.-P.: Scale invariant feature transform on the sphere: Theory and applications. Int. J. Comput. Vis. 98, 217–241 (2012).
9. Lourenco, M., Barreto, J.P., Vasconcelos, F.: sRD-SIFT: Keypoint detection and matching in images with radial distortion. IEEE Trans. Robot. 28, 752–760 (2012).
10. Jeon, Y., Kim, J.: Active convolution: Learning the shape of convolution for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4201–4209 (2017).
11. Coors, B., Paul Condurache, A., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 518–533 (2018).
12. Khasanova, R., Frossard, P.: Graph-based classification of omnidirectional images. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 869–878 (2017).
13. KIM[1], J., Kim, B.S., Savarese, S.: Comparing image classification methods: K-nearest-neighbor and support-vector-machines. In: Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics. pp. 42122–48109 (2012).
14. Liu, P., Choo, K.-K.R., Wang, L., Huang, F.: SVM or deep learning? A comparative study on remote sensing image classification. Soft Comput. 21, 7053–7065 (2017).
15. Urban, S., Weinmann, M., Hinz, S.: mdBRIEF-a fast online-adaptable, distorted binary descriptor for real-time applications using calibrated wide-angle or fisheye cameras. Comput. Vis. Image Underst. 162, 71–86 (2017).
16. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27 (2011).
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016).
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678 (2014).