Orientation Estimation Network

Jie Sun, Wengang Zhou(✉), and Houqiang Li

Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei 230026, Anhui, China
zhwg@ustc.edu.cn

Abstract. We propose the Orientation Estimation Network (OEN) to predict the dominant orientation of the outdoor images and rotate the images to a canonical orientation which is visually comfortable. The OEN outputs the sine and cosine of the angle which are continuous in contrast to the angle. We collect a new dataset called the Outdoor Images dataset for this task. This dataset contains various kinds of outdoor images, such as buildings, landscape, persons and boats, and the orientation information has been manually annotated. We choose AlexNet, MobileNet and VGGNet to extract image features and regress to the angle of images. In our task, MobileNet achieves high performance while needing less resource, and can be applied to mobile and embedded vision applications. We compare our method with the hand-crafted methods on our dataset. In the evaluation, our learning based method significantly outperforms the hand-crafted methods in the task of outdoor images orientation estimation.

Keywords: Outdoor images · Orientation estimation · Deep learning

1 Introduction

In recent years, convolutional neural networks have been used in various tasks in computer vision, such as classification [4–7,26], detection [8–10,25], segmentation [11], image search [15,16], and have achieved state-of-the-art performance in these tasks. CNN has got unprecedented success ever since AlexNet [4] won the ImageNet Challenge in 2012 because of the power of hierarchical abstract representation, but it also has the limit when dealing with rotation invariance only by convolution and pooling.

To overcome such limit, the traditional solution is data augmentation. Training samples are rotated into multi-oriented versions. Although data augmentation improves the performance by extending the training data, the network tends to be more fitted to the training data and would lose some generalization capacity, and more training time is required. Another way is rotating the filters. Usually, one filter in the network can detect one specific pattern in the image, so we can rotate this filter to detect the same pattern with different orientations. It can alleviate the network to learn all different orientations and achieve rotation invariance. The motivation is straight-forward, but it has to change the way of

Y. Zhao et al. (Eds.): ICIG 2017, Part II, LNCS 10667, pp. 151–162, 2017.
https://doi.org/10.1007/978-3-319-71589-6_14

forward and backward propagation which is not convenient and the angle of the filters is discontinuous.

In this paper, we directly predict the image orientation, instead of making a rotate invariant representation, and then align the images according to the angle predicted by the OEN to achieve upright configuration of the visual content. This approach is inspired by hand-crafted features, such as SIFT [1]. SIFT calculates the domain orientation of the keypoint and aligns the patch by the domain orientation. After that, SIFT gets the descriptor for this key-point. Unlike SIFT, we predict the global orientation of images instead of local orientation. Here we choose the outdoor scene images as our target images because they usually have a clearly defined principal orientation in human perception.

In this paper, we make orientation estimation by regressing to the human annotated ground truth. In contrast to our method, Spatial Transformer Network [12] uses the classification information as ground truth to learn the transform indirectly. Spatial Transformer Layer learns the way to transform a feature map or a region of images. And the transform is forwarded to the next layer. But STN can only handle a limited range of orientation variance. Here we give the outdoor images a canonical orientation and we learn this orientation directly by OEN. So we can learn the images with arbitrary orientation and finally rotate images to the appropriate orientation. Figure 1 shows the different orientations of the same scene. The contributions of this paper are summarized as follows:

- We propose a new task of predicting the canonical orientation of outdoor images and present a solution for this task.
- We collect a new dataset called Outdoor Images for our task. The dataset is composed of outdoor images and the orientation information has been manually annotated.
- We compare our method with hand-crafted methods and study AlexNet, MobileNet [19] and VGGNet [5] in our task.

Fig. 1. Different orientations of same scene

2 Related Work

The related works to achieve rotation invariance: (1) the hand-crafted orientation estimation, (2) data augmentation and (3) Spatial Transformer Network.

2.1 Hand-Crafted Orientation Estimation

Orientation information is important for hand-crafted feature to align local patches to achieve rotation invariance. SIFT detector calculates the dominant orientation of key-point by statistics of local gradient direction of image intensities. ORB detector [2] uses the moment of a patch to find the offset between the patch's intensity and its center and then this offset vector is used as the orientation of key-point. BRISK detector [3] and FREAK detector [13] sample the neighborhood of the key-point by using a pattern. The long distance point pairs are used to calculate the orientation of key-point.

In contrast to hand-crafted orientation estimation, our learning-based method automatically predicts the orientation without hand-crafted feature detectors. The power of CNN to extract feature is more effective than hand-crafted method. In recent years, traditional methods have been replaced gradually and usually are used as baselines.

2.2 Data Augmentation

Deep convolution neural network have the ability of dealing with transitions, scale changes, and limited rotations. And the capability comes from rich convolutional filters, and pooling. Data augmentation is used to achieve local or global transform invariance with rich convolutional filters [14]. Therefore, data augmentation can improve performance for many tasks. However, the network tends to be more fitted to the training data and would loss some generalization capacity, and more training time is required.

In contrast to data augmentation, our method based on network predict the canonical orientation directly. Then the learned orientation can been used to other tasks to achieve higher performance.

2.3 Spatial Transformer Network

The Spatial Transformer Layer is proposed by Jaderberg et al. [12], and learns the way to transform a feature map or a region of images. The transform is forwarded to the next layer. A general framework for spatial transform comes out by STN, but the problem about how the complex transform parameters use CNN to precisely estimate has not been well solved. Most recent work [17,18,20,21,23,24] have tried rotating conventional filters to achieve rotation invariance. But they have to change the way of forward and backward propagation.

In contract to STN, our method learns the orientation directly using the ground truth of the explicit orientation. And STN learns the orientation implicitly using the ground truth of the task, such as classification information. STN

has the limit to significant change of orientation. So we solve this problem by learning orientation directly in our method.

3 Orientation Estimation Network

In this work we focus on predicting the canonical orientation for outdoor images and making images visually comfortable. The Orientation Estimation Network takes outdoor images as input, and outputs the angle of images which is consistent with human perception. Then we take the predicted angle outputs to rotate the images to a canonical orientation.

In the following, we introduce the detail of Orientation Estimation Network by three parts: (1) Learning stage, (2) Fine-tuning stage and (3) Predicting stage.

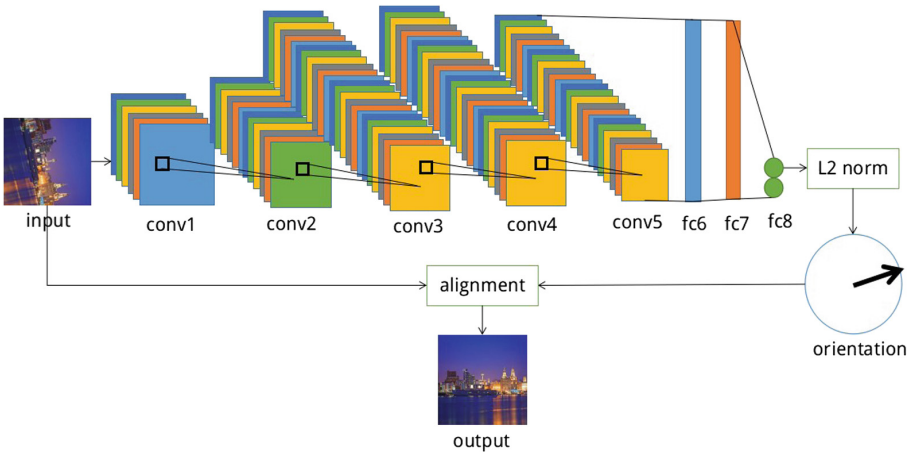

Fig. 2. The framework of our method based on AlexNet

3.1 Learning Stage

In this stage, we take the images as the input, through a classic network, such as AlexNet, the champion of ImageNet 2012. Then we get the CNN features extracted from images. We combine the features and output two values which are defined as sine and cosine of the angle. We choose sine and cosine of angle, because sine and cosine are continuous with respect to the angle. So it is easy to optimizes and train. Besides, we add a normalization layer which normalizes the output vector to unit-norm to ensure the validity of sine and cosine. We use L2 loss between the predicted values and the ground truth in training. We calculate the angle by the arctangent function. Figure 2 shows the framework of our method.

3.2 Fine-Tuning Stage

It is not enough robust only based on learning stage, where the network learns one image at one time. And the orientation can be transformed by rotating the image. So two images at same scene with different orientations can help network to learn the orientations of them by each other. In this stage, we learn the rotated angle between two images at same scene to help improve the performance of orientation estimation.

In the fine-tuning stage, we extend our framework to a Siamese architecture. The input images are the same scene with two different orientations. The ground truth is the difference between the orientations of input images. The predicted angle is the difference between the orientations which are outputted by the two sub-networks for input images. The loss is the difference between the ground truth angle and the predicted angle. After the single network learning stage, we have got a good performance for prediction. Based on that, we improve the performance by fine-tuning. We fine-tune the fully-connected layers for AlexNet and VGGNet, and last convlutional layer for MobileNet. Figure 3 shows the framework of the fine-tuning stage.

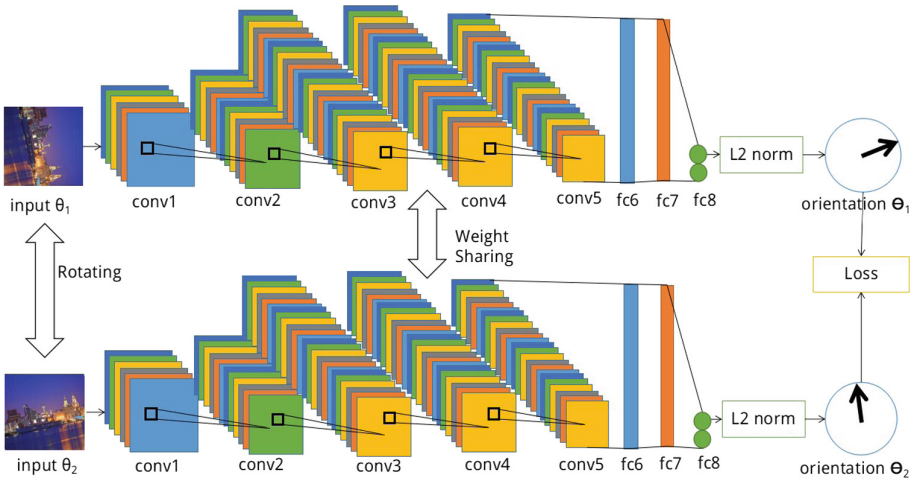

Fig. 3. The framework of the fine-tuning stage based on AlexNet. Two input images are same scene with different orientations, which are θ_1 and θ_2. Two output orientations are Θ_1 and Θ_2. And the loss is $||\theta_1 - \theta_2| - |\Theta_1 - \Theta_2||^2$.

3.3 Predicting Stage

In the predicting stage, we take the image as the input and output the orientation. We use a sub-network from the Siamese architecture to predict the angle. Then we rotate the image using the predicted orientation. Table 1 reports the detail of OEN architecture based on AlexNet.

Table 1. The architecture of the orientation estimation network based on AlexNet

Type	Size
Input	227 * 227 * 3
Conv1+Pool1	27 * 27 * 96
Conv2+Pool2	13 * 13 * 256
Conv3	13 * 13 * 384
Conv4	13 * 13 * 384
Conv5+Pool5	6 * 6 * 256
FC6	1 * 1 * 4096
FC7	1 * 1 * 4096
FC8	1 * 1 * 2
L2 normalization	1 * 1 * 2
Output	1 * 1 * 1

4 Experiment

We introduce the experiments by three parts. First, we introduce our new dataset, Outdoor Images dataset. Second, we compare the performance of three classic networks which are applied to Oriented Estimation Network and we compare our method with the hand-crafted methods. Third, we analyze the result of the experiments.

4.1 Dataset

In our experiment, we test our method in the Outdoor Images dataset collected by ourselves, where the orientations of images have been manually annotated. The dataset is composed of several kinds outdoor scene, such as buildings, landscape, persons, boats, and has been divided into the training set and test set. The images in our dataset are selected from the Flickr1M dataset [22]. Figure 4 shows some images in our Outdoor Images dataset.

The images have been preprocessed to keep the information in the circle of center while dropping outside. The pixels out of circle will go to the outside of images when we rotate the images, so these pixels are abandoned. They are colored as black in case of influencing the experiment result.

4.2 Experiment Setup

In our experiment, we first compare the performance of AlexNet, MobileNet and VGGNet for predicting the orientation of images. AlexNet contains five convolutional layers and three fully-connected layers. The fully-connected layers almost take up 90% parameters of AlexNet, and MobileNet drops the fully-connected layers to compress model. MobileNet takes many 3×3 depthwise convolutional

Fig. 4. Sample images in the outdoor images dataset

filters and 1×1 pointwise convolutional filters to reduce a large number of parameters. MobileNet only has 4.2 million parameters while 60 million for AlexNet. And VGGNet has 16 layers and 138 million parameters. So MobileNet has the advantage to be applied to mobile and embedded vision applications.

We also compare our CNN method with the hand-crafted methods. In our experiment, SIFT calculates the dominant orientation of the whole image by statistics of global gradient direction of image intensities, instead of local gradient information. ORB uses the moment of the whole image to find the offset between the image's intensity and its center and then this offset vector is used as the orientation of image. BRISK and FREAK sample the neighborhood of the whole image by using a pattern. The long distance point pairs are used to calculate the orientation of the image. We use the average error of orientation as the criteria for evaluation.

4.3 Results

In our experiment, we use L2 loss for training. And we set initial learning rate as 0.0001 and every epoch drops to 0.96 of the last learning rate. We set the batch size as 128 and the size of input image as 227×227.

Table 2 shows the average error of AlexNet, MobileNet and VGGNet in our task for predicting the orientation of images. We set the same learning rate and other experiment parameters for them. The results show that MobileNet has better performance than AlexNet and has comparable performance with VGGNet.

But MobileNet needs few resources. So it is suitable to choose MobileNet in our task to predict the orientation.

Table 2. Average error of AlexNet, MobileNet and VGGNet for predicting the orientation of images

Network	AlexNet	MobileNet	VGGNet
Average error (degree)	25.85	23.63	23.61

Table 3 shows the average error of the OEN and the hand-crafted methods for predicting the global orientation of images. The results show that our method significantly outperforms the hand-crafted methods. Because SIFT and ORB which use the information of intensities to decide the orientation have no relationship with the global orientation which is visually comfortable. BRISK and FREAK which use the long distance point pairs have the same reason with SIFT and they are easy to change orientation after moving a few pixels.

Table 3. Average error of our method and the hand-crafted methods for predicting the global orientation of images

Method	Our method	SIFT	ORB	BRISK	FREAK
Average error (degree)	25.85	40.08	52.76	70.98	68.47

Figure 5 shows the angle error histogram of our method. The error of the angle is mostly below 20°. Therefore, our method is stable to predict the orientation of images. It is reliable to rotate the images by the predicted orientation through our method.

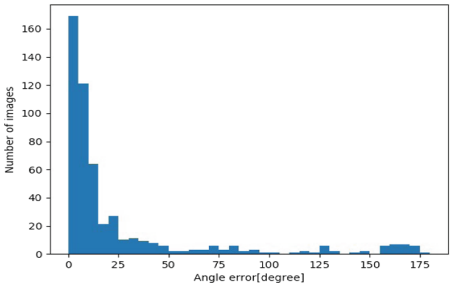

Fig. 5. The angle error histogram of our method

Figure 6 shows examples of predicting the orientation of images with little error. The left column is the ground truth images, the middle column is the

Fig. 6. Examples of predicting the orientation of images with little error. The left column is the ground truth images, the middle column is the input images for prediction and the right column is the results of the images rotated by the predicted orientation.

Fig. 7. Examples of predicting the orientation of images with large error. The left column is the ground truth images, the middle column is the input images for prediction and the right column is the results of the images rotated by the predicted orientation.

input images for predicting and the right column is the results of the images rotated according to the predicted orientation.

The results show the examples with little error and they catch the information of background for outdoor images. These images are typical ones mostly occupied by background. So it can be predicted well.

Figure 7 shows examples with larger error than examples in Fig. 6. In the images which contain buildings, the height of buildings are not same. So the line for top of buildings is not parallel with the ground. It is confusing for the network to predict the orientation of images and it causes error. In the images which contain persons, it has large error for predicting orientation. The reason we thought is that persons occupy too much space and have a little background while other kinds images have more background information. So the images contain persons are not predicted well. A solution to address this problem is to train for the images contain persons alone, if the task is predicting an orientation only for images with persons.

5 Conclusion

We have presented a new task of calculating the holistic dominant angle for outdoor images, and aligning the images to be visually comfortable. We compare CNN method with hand-crafted methods and show the advantage of convlutional neural network. Experiment on AlexNet, MobileNet and VGGNet demonstrates the performance for predicting the canonical orientation of outdoor images. And it turns out that MobileNet is more suitable for this work with less average error and less resource.

Acknowledgements. This work was supported by NSFC under contract No. 61472378 and No. 61632019, the Fundamental Research Funds for the Central Universities, and Young Elite Scientists Sponsorship Program By CAST (2016QNRC001).

References

1. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (1999)
2. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: ICCV, pp. 2564–2571 (2011)
3. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: binary robust invariant scalable keypoints. In: ICCV, pp. 2548–2555 (2011)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR, pp. 1–9 (2015)

8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
9. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv preprint arXiv:1703.06870 (2017)
12. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: NIPS, pp. 2017–2025 (2015)
13. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: fast retina keypoint. In: CVPR, pp. 510–517 (2012)
14. Van Dyk, D.A., Meng, X.L.: The art of data augmentation. J. Comput. Graph. Stat. **10**(1), 1–50 (2001)
15. Liu, Z., Li, H., Zhou, W., Tian, Q.: Uniting keypoints: local visual information fusion for large scale image search. IEEE Trans. Multimed. **17**(4), 538–548 (2015)
16. Zhou, W., Lu, Y., Li, H., Song, Y., Tian, Q.: Spatial coding for large scale partial-duplicate web image search. In: ACM MM, pp. 131–140 (2010)
17. Wu, F., Hu, P., Kong, D.: Flip-rotate-pooling convolution and split dropout on convolution neural networks for image classification. arXiv preprint arXiv:1507.08754 (2015)
18. Marcos, D., Volpi, M., Tuia, D.: Learning rotation invariant convolutional filters for texture classification. arXiv preprint arXiv:1604.06720 (2016)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
20. Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Oriented response networks. arXiv preprint arXiv:1701.01833 (2017)
21. Marcos, D., Volpi, M., Komodakis, N., Tuia, D.: Rotation equivariant vector field networks. arXiv preprint arXiv:1612.09346 (2016)
22. http://press.liacs.nl/mirflickr/
23. Cohen, T.S., Welling, M.: Steerable CNNs. arXiv preprint arXiv:1612.08498 (2016)
24. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Harmonic networks: deep translation and rotation equivariance. arXiv preprint arXiv:1612.04642 (2016)
25. Mao, J., Li, H., Zhou, W., Yan, S., Tian, Q.: Scale-based region growing for scene text detection. In: ACM MM (2013)
26. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: CVPR, pp. 1134–1142 (2016)