Eigen Vectors based Rotation Invariant Multi-Object Deep Detector (EVRI-MODD)

1st Bharat Giddwani
Dept. of Electronics and Communication
National Institute of Technology, Raipur
Raipur, India
bharatgiddwani@gmail.com

2nd Dheeraj Varma
Dept. of Electronics and Communication
National Institute of Technology, Raipur
Raipur, India
dheeraj.vinukonda@gmail.com

3rd Mohana Murali
Dept. of Electrical Engineering
Indian Institute of Technology, Tirupati
Tirupati, India
ee18d00@iittp.ac.in

4th Rama Krishna Gorthi
Dept. of Electrical Engineering
Indian Institute of Technology, Tirupati
Tirupati, India
rkg@iittp.ac.in

***Abstract*—In this paper, we propose an accurate, scalable data-free approach based on eigenvectors and Convolutional Neural Networks (CNNs) for rotated object detection. Multiple object detection using CNN based algorithms such as Faster-RCNN, SSD, YOLOv3 have evolved as a robust and faster approach for extracting features from the training data. Detecting an arbitrarily diverted object poses a challenging problem, as features extracted by CNNs are variant to small changes in shift and scale. They lack in performance for images at orientation different from input data. Hence, we introduced a novel two-step architecture, which detects multiple objects at any angle in an image efficiently. We utilize eigenvector analysis on the input image based on bright pixel distribution. The vertical and horizontal vectors are used as a reference to detect the deviation of image from the original orientation. This analysis gives four orientations of the input image which, we pass through a pre-trained YOLOv3 with proposed unique decision criteria. Our approach referred to as "Eigen Vectors based Rotation Invariant Multi-Object Deep Detector" (EVRI-MODD), produces rotation invariant detection without any additional training on augmented data and also determines actual image orientation without any prior information. Our proposed network achieves high performance on Pascal-VOC 2012 dataset. For demonstration, we evaluate our model on three differently rotated angles, 90°, 180°and 270°, and achieves a significant gain in accuracy by 48%, 50% and 47% respectively, over YOLOv3.**

***Index Terms*—eigenvectors, object detection, YOLOv3, orientations, pre-trained, decision criteria, rotation invariant**

I. Introduction

In Deep Learning, a Convolutional Neural Network (CNN or ConvNet) is a class of deep, feed-forward neural networks used mostly in the analysis of visual imagery. One of the advantages of CNNs is its translation equivariant property provided by weight sharing. The feature space learned from the mapping function of CNN changes in the same way as the linear transformation in the image. However, it cannot deal with a rotation transformation of input images. Though the features extracted by CNN are equivariant to small changes in shift and scale, they are sensitive to rotations in the input image. To overcome this problem, the classical approach is data augmentation by including rotated images. However, it has some limitations, the performance and generalizability of Deep Neural Networks are largely dependent on the availability of data where we are not always provided with the same, cases such as defence and medical, where data privacy is required. It does also increase the training time by a factor proportional to the number of rotations introduced in training data.

All the pictures in visual imagery are not in the same orientation. During capture, they differ from each other in terms of alignment due to positional disturbance of the camera. Moreover, few scenes obtain significant view while capturing in "landscape", while some in "portrait". Finally, if all images pass through an object detector, then all of them must be in the same orientation, as CNNs are unable to disentangle planar rotation transformations. For accurate detection on rotated images, CNNs should be invariant to these planar rotations.

With the success of convolution networks many object detectors such as Faster-RCNN [13], SSD [9], YOLO [10] [11] [12]and its modifications are showing fast and accurate detections. We use YOLOv3 [12] as the base pre-trained object detector where YOLO stands for You Only Look Once, one of the faster object detection algorithms. This uses features learned by a deep CNN to detect an object.

YOLOv3 is an object detector, built on a network called Darknet-53, trained on images at normal orientation. It makes incorrect predictions for rotated images/objects. Performance of YOLOv3 for original and rotated images is shown in Fig.3. Thus, it indicates that YOLOv3 is failing in object detection when diverted images pass through it, as it is only trained on images at a specific orientation. To tackle object detection in rotated scenes, we need to train the CNN on rotations of training images. It is an immense task since it involves rotation and collection of ground truths for every image. The training duration also increases drastically, Hence, an efficient method is required, which could make object detectors perform better on rotated images.

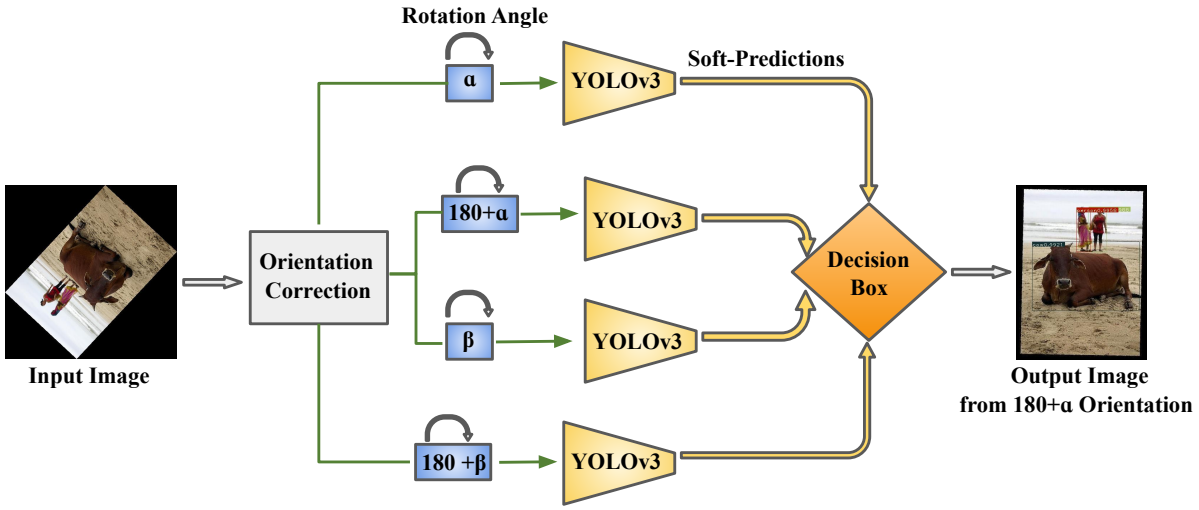


Fig. 1. End-to-End testing phase of EVRI-MODD.

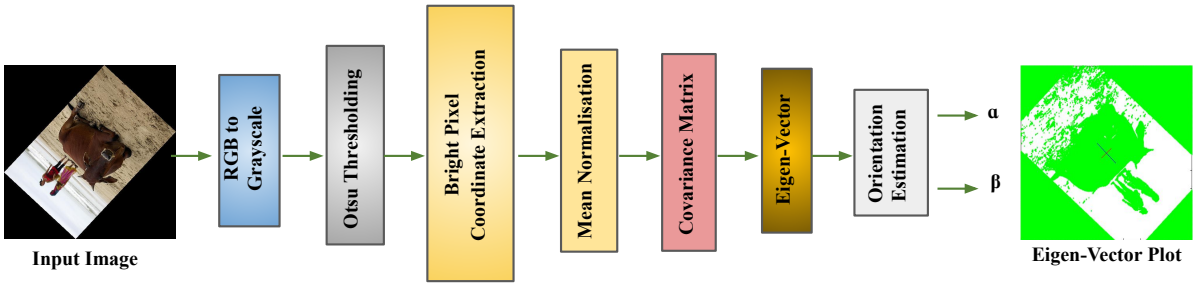


Fig. 2. Pre-processing steps for calculating orientation angles.

II. Related Works

Rotational invariant object detection is one of the prominent research topics in computer vision. In the past decade, researchers worked on this using deep learning and CNNs, to produce better accuracy over data augmentation methods. For images containing single objects, works such as Spatial Transformer Networks (STN) [5], TI-pooling [7], Oriented Response Networks (ORN) with Active Rotating Filters (ARF) [15], RIFD-CNN [2] are proposed. These techniques implement data augmentation in training phase or architectural modifications, which increases training complexity. Data augmentation is practically impossible for very large datasets with millions of training images, such as that provided in ILSVRC [14] and also in cases where data privacy is mandatory.

Fast rotation invariant object detection with gradient based detection models [3] introduces the method of training specific model multiple times, each with a different orientation. During testing, a rotation map containing information about orientations at a particular location from dominant orientation is obtained. They use SURF algorithm on Haar features to calculate the dominant orientation. This method claims high evaluation speed, but training complexity is similar to data augmentation as different orientations are used to train a specific model.

Fig. 3. YOLOv3 detection accuracy performance on original and rotated images.

III. Proposed Approach: Orientation Correction Analysis with YOLOv3

YOLOv3 makes very accurate predictions only when the image is in training orientation. Therefore, it is required that detection is precise on deviated images. Our proposed method is based on estimation of the approximate angle of rotation

Fig. 4. Decision Logic for rotation invariant image annotation.

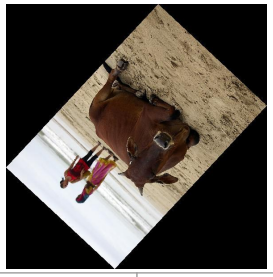
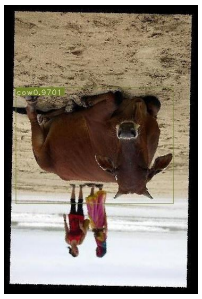
Input Image				
Angle	**Φ1= α°**	**Φ2= 180+α°**	**Φ3= β°**	**Φ4= 180+β°**
Rotated Image				
Detections	Cow : 0.970	Cow : 0.992 Person: 0.996, 0.988	Cow : 0.965	Horse: 0.742
Class Frequency	Cow : 3, Person: 1, Horse: 1			
Frequent	Cow			
Max. Score	Cow : 0.992 from **Φ2**			
Principal Image	**Φ2** (Since it has the maximum score for frequent class)			

Fig. 5. Decision Logic explained for finer rotation invariant image detection based on class frequency and their classified objectness score.

using eigenvectors to perform rotational invariant detection. Hence, this method is incorporated with YOLOv3 so that it aligns the input image close to the orientation of training images and makes accurate predictions.

YOLOv3 gives high objectness or prediction score for images oriented at an angle close to that of training images. When

TABLE I
COMPARISON BETWEEN OURS METHOD WITH YOLOV3 IN TERMS OF CLASSIFICATION ACCURACY.

Class	0		90		180		270	
	Ours	YOLOv3	Ours	YOLOv3	Ours	YOLOv3	Ours	YOLOv3
Aeroplane	98.11	99.86	98.13	64.03	98.37	83.75	98.10	66.23
Bicycle	97.43	99.91	99.60	62.86	97.86	51.03	98.53	63.87
Bird	98.69	98.37	98.90	65.80	98.76	67.40	98.12	73.98
Boat	69.47	97.09	75.05	04.79	74.85	08.04	70.15	02.83
Bottle	78.69	97.84	85.77	64.01	79.52	45.82	81.67	65.52
Bus	94.88	99.94	94.86	09.02	94.76	29.49	94.62	02.55
Car	99.84	99.27	99.64	05.98	99.23	34.80	99.15	09.60
Cat	94.71	95.98	94.65	90.45	94.85	42.40	93.84	76.69
Chair	78.85	95.23	75.08	17.04	77.84	06.84	76.43	12.08
Bike	99.91	99.94	98.63	56.85	98.22	41.50	98.35	58.31
Mean	91.04	98.34	**91.73**	43.78	**91.42**	41.11	**90.89**	43.16

rotated images pass through YOLOv3, it makes incorrect detections and even if it is correct, the objectness score is low as compared to that of proper oriented image. The nearly oriented images have a high score for each object, whereas large deviated images have misclassified objects or correctly classified with lower scores.

A. Training Phase

Training involves no data augmentation or rotation of images. It is trained in the same way as original, with only one specific orientation. No modifications are made in the architecture , and if pre-trained model is available, the training can be completely avoided.

B. Testing Phase

The proposed method is implemented during testing. Firstly, we convert the input image to grayscale and further apply Otsu's thresholding to binarize it. Then we extract the locations of bright pixels into a matrix from the segmented image and calculate the covariance matrix of these coordinates. Next, we perform singular value decomposition on the obtained covariance matrix. By this, we get two eigenvalues and corresponding eigenvectors, termed as principal components. And the eigenvector corresponding to largest eigenvalue is the first principal component. The above explained method is illustrated in the form of a flowchart in Fig.2. By observation, we note the following points:

- The direction of the first principal component of images of a certain class is nearly same.
- The first principal component is approximately horizontal for original images with aspect ratio width $>$ height and vertical for images with aspect ratio width $<$ height.
- The eigenvectors of an image and its 180°rotated version are in the same direction.

So, the estimated principal components may belong to two versions of an image. Next, we perform the orthogonal transformation [1] of these principal components so that they align in exactly horizontal and vertical positions and consider these as reference vectors. By calculating the angle between the first principal component of the input image and these reference vectors, we obtain two angles of rotation, say α°and β°. Then, with consideration of the above mentioned three observations, we rotate the input image by α°, 180 + α°, β°and 180 + β°. This is the end stage of the pre-processing method and hence, we obtain four images at different orientations. We feed these four images to YOLOv3, to obtain detections and their objectness scores. The principal image among them is selected based on a decision criterion which is discussed in the next section. The following Fig.1 demonstrates the flow of the process explained above.

C. Decision Criteria

Based on the behaviour of YOLOv3 for rotated images, a decision criterion based on class frequency and objectiveness score is developed for selecting the principal image among the four, which gives the correct annotation of the input image. Motivated from the work by [6], multiple orientations of the same image are fed to YOLOv3.

Observations from detections are as follows:

- The detections are accurate in the image having orientation closer to training images. So, the principal image will have good predictions and higher scores.
- The object which is not in the principal image might be detected in any of other three other orientations. But the score of that detection is less than that of at least one detection in the principal image.
- The objects present in the principal image can also be identified in other orientations with low scores. Thus, the number of times the true detections appear will be more than or equal to false detections.

Thus, we count the number of times an object gets detected in all the four orientations which gives the class frequency or likelihood. Then, the score of the object can be captured in each orientation. Even if the objects in the principal image are not detected in the other three, one of them will have the highest score among all detections given on all four orientations. So, the detection with maximum objectness score or statistical mode of classes, forms the base for selecting the principal image. For instance, as shown in Fig.4, the bicycle is detected in all four orientations, it is the statistical mode class.

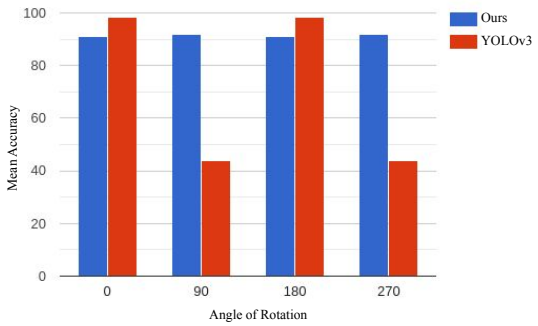


Fig. 6. Bar chart representing mean accuracies.

The prediction score of bicycle is found to be highest in the third rotated image. Thus, this is our required principal image. The method also works for finer rotations in the image other than some discrete rotations. For example, as shown in Fig. 5 the input image is in a random rotated version. By applying our method, final four orientations with predictions are obtained. Among those, cow is the statistical mode class, with highest class frequency. So, we find the maximum objectness score among the detections of cow to discern the principal image. Hence, the second image is the principal image based on class with maximum likelihood and its prediction score.

In brief, we find the class which occurs the maximum number of times in all four detections. Then, we find the scores of that particular class in every image, if detected. We find the maximum of those scores and the image having the detection with that score is chosen as principal image. If there is no repetition, select the detection with the highest score in all four images and the respective image is principal image. Further, if two or more classes have maximum likelihood, then select the one with maximum objectness score [6].

IV. Results

The dataset provided by PASCAL VOC challenge 2012 (consists of 20 classes) [4] is used for test analysis. YOLOv3 is pre-trained on MS COCO dataset [8] with 80 classes. The test images are rotated at random angles (specifically, 0°, 90°, 180°and 270°) to create rotated test data. The original image is rotated around 137°to obtain that input image in Fig 5. Our method is compared with YOLOv3 in terms of accuracy. The accuracy is calculated by summing the detection scores and dividing the sum by the number of images of that class. The comparison is displayed in the Table I. The angles mentioned at the top of the table in Table I are the rotation angles through which original (good) images are rotated. This is done for experimental purposes. Then, the proposed method is applied to the rotated images and we obtain the results. The results are accurate for finer rotations too as illustrated in Fig. 5. The mean accuracies calculated in the Table I are represented in a column-chart in Fig.6.

V. Conclusion

Observing the results obtained from our method and YOLOv3, we can say that the proposed detection method is elegant and robust, which makes accurate predictions irrespective of the sensitive behaviour of CNNs to rotation. The prominent features of proposed method are training simplicity, cost, computational efficiency and no architectural modification. This method also works for finer degrees of rotation in the image acquired by the camera, by passing only four resulting orientations into the network. In the method of incorporating rotational invariance to the CNN using multiple instances of network [6], the image is fed into the network for N times, where N depends on the degree of rotation. It can go beyond four for much finer degrees of rotation. Meanwhile, the proposed method uses only four instances every time, even for smaller and finer rotations. Thus, the proposed method is much simplified and time-efficient.

References

[1] Abass, H.H., Al-Salbi, F.M.M.: Rotation and scaling image using pca. Computer and Information Science **5**(1), 97 (2012)
[2] Cheng, G., Zhou, P., Han, J.: Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2884–2893 (2016)
[3] De Smedt, F., Goedemé, T.: Fast rotation invariant object detection with gradient based detection models. In: Proceedings of the 10th International Conference on Computer Vision Theory and Applications-(Volume 2)(VISIGRAPP 2015). vol. 2, pp. 400–407. VISIGRAPP; Setubal (2015)
[4] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
[5] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
[6] Kandi, H., Jain, A., Chathoth, S.V., Mishra, D., Subrahmanyam, G.R.S.: Incorporating rotational invariance in convolutional neural network architecture. Pattern Analysis and Applications **22**(3), 935–948 (2019)
[7] Laptev, D., Savinov, N., Buhmann, J.M., Pollefeys, M.: Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 289–297 (2016)
[8] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
[9] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
[10] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
[11] Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
[12] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
[13] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
[14] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
[15] Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Oriented response networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 519–528 (2017)