17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Expandable Spherical Projection and Feature Fusion Methods for Object Detection from Fisheye Images

Songeun Kim
School of Electronic and Electrical Engineering
Kyungpook National University
akskdk4444@gmail.com

Soon-Yong Park
School of Electronics Engineering
Kyungpook National University
sypark@knu.ac.kr

Abstract

One of the key requirements for enhanced autonomous driving systems is accurate detection of the objects from a wide range of view. Large-angle images from a fisheye lens camera can be an effective solution for automotive applications. However, it comes with the cost of strong radial distortions. In particular, the fisheye camera has a photographic effect of exaggerating the size of objects in central regions of the image, while making objects near the marginal area appear smaller. Therefore, we propose the Expandable Spherical Projection that expands center or margin regions to produce straight edges of de-warped objects with less unwanted background in the bounding boxes. In addition to this, we analyze the influence of multi-scale feature fusion in a real-time object detector, which learns to extract more meaningful information for small objects. We present three different types of concatenated YOLOv3-SPP architectures. Moreover, we demonstrate the effectiveness of our proposed projection and feature-fusion using multiple fisheye lens datasets, which shows up to 4.7% AP improvement compared to fisheye images and baseline model.

1 Introduction

For advanced driver-assistance systems(ADAS), some of the important properties are to obtain comprehensive information about the environment and to cover a sufficiently wide range of view. In order to thoroughly understand the road scenes, it is necessary to detect all the relevant surrounding objects. Currently, deep-learning based methods show the most promising performance. This approach requires relatively large computational resource, but with modern hardware can easily be adapted to real-time detection.

Since most ADAS depend on visual information, a considerable number of studies are being conducted on the vision-based object detector. In a low-cost sensor setup, 2D cameras with high field of view(FOV) can effectively cover a large area around the vehicle and ensure the safety of the autonomous driving. However, this advantage comes at the cost of strong radial distortion. The resulting issues, such as curving and diagonal tilting of objects are increasingly severe towards the edges of the image. Another notable feature of wide FOV camera is that both the relative size and distance are exaggerated. When comparing near and distant objects, nearby objects appear much larger, while objects located far away appear much smaller than in the general camera. Consequently, the already poor performance of object detectors for small objects is further degraded.

To solve these problems, SphereNet [1] suggests the distortion-invariant neural network for the omnidirectional images, adapting the sampling grid locations of a convolutional kernel. Alternatively, a rotation-invariant model, which predicts object orientations was proposed by [2, 3]. However, these studies require complex computations, hindering the real-time performance required from one-stage object detectors.

Therefore, we propose a simple but effective spherical-based projection. In order to compensate for the weaknesses of the fisheye lens, our method simply expands the center or border areas of the image without the necessity for multiple complicated operations. Moreover, we suggest three variants of multi-scale feature-fusion method for the YOLOv3 [4] with Spatial Pyramid Pooling [5] (YOLOv3-SPP). Each of the variants incorporates a different feature concatenation scheme. Short-skip Concatenation (SCat) merges additional smaller scale feature maps from the neck part of the detector, while Long-skip Concatenation (LCat) draws the features from the backbone. Short-Long-skip Concatenation (SLCat) fuses the feature maps from both neck and backbone, effectively combining the core features of SCat and LCat. We evaluate our solution with several public datasets, as well as our new collection of images gathered with a 185° fisheye lens. The major contributions of this study are noted as follows.

- Introduction of a new front-view fisheye dataset consisting of 3K bounding box annotations.
- Proposal of simple but effective spherical projection on fisheye images,
- Analysis of feature fusion methods to reduce small objects detection issues in real-time object detector

2 Related Works

Fisheye Dataset for Urban Driving: Several authors have considered the fisheye cameras for the ego vehicles. Among them, some notable works are classification and tracking of cars and pedestrians using hybrid cameras [6], pedestrian detection using a combination of synthetic, and real images from a 360° horizontal FOV camera [7], as well as a multi-camera fisheye dataset for multi-task with 7 different object categories by WoodScape [8]. The deep-learning based models require large volumes of training data, but generating such a dataset is a very time-consuming task. Therefore, some authors suggest leveraging synthetic fisheye data generated from non-fisheye datasets [9, 10, 11, 12, 13].

Multi-scale Feature Fusion for Small Objects: MS COCO [14] defines small objects as having a bounding-box size less than 32^2 pixels. Detecting these small objects is crucial in many applications, such as autonomous vehicle and unmanned aerial vehicle(UAV). One of the representative methods is multi-scale feature learning [15]. FPN [16] has been used to improve the accuracy of small object detection with the connection between shallow and deep layers. Moreover, additional concatenation of low-level features achieves higher detection results on small objects in [17, 18]. We make use of similar approach.

Ultra-Wide Angle Projection: Ultra-wide angle lens can cover large areas with more than 100° horizontal and vertical FOV, but results in considerable distortions. Consequently, several rendering methods have been studied to minimize this. The most representative projections are spherical and cylindrical methods [3, 19]. Some works have successfully demonstrated in 2D object detection in spherical images with non-standard convolutions [1, 2]. Using the cylindrical projection, several authors estimate the depth from wide angle cameras [19, 20]. In this paper, we use spherical-based projection since it can support high FOV of fisheye lens in horizontal and vertical line and simply render the images. Spherical projection, also called as equidistant cylindrical projection, maps the longitude and latitude(θ, ϕ) are linearly to horizontal and vertical coordinates(x, y) [21]: $x = \theta$, $y = \phi$

3 Proposed Method

3.1 Training Data

We use synthetic as well as real fisheye datasets. The synthetic fisheye images are generated from CityScape [22] and KITTI [23] datasets following the method in ERFNet [13]. Additionally, we convert segmentation annotations of CityScape to bounding boxes.

We collected our own fisheye camera images using the Fujinon FE185C057HA-1 lens with 1.8mm focal length and 185 ° field of view. Circular fisheye images were captured with a 2/3 inch sensor. Since a front view is the most important vision of the autonomous driving, we installed the camera to the front window of the vehicle as illustrated in Fig. 1. The data, consisting of 21k road scenes in Daegu, South Korea, were gathered during cloudy and drizzling weather. Bounding boxes were manually annotated to 3k images. The full set of annotated images will be completed in future work.

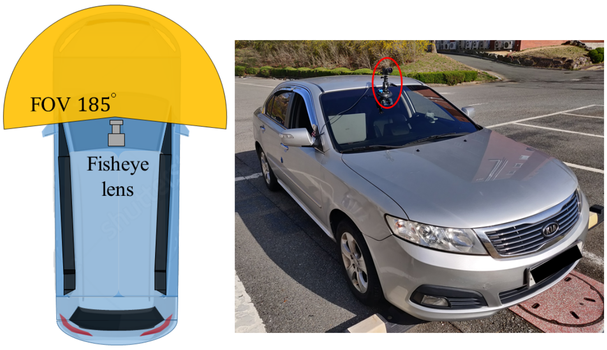

Figure 1. Fisheye lens camera installed on a front window of a test car

3.2 Expansion Weight for the Projection

Expandable Spherical projection is based on the equidistant projection with an additional parameter w, which is a non-negative expansion weight for increasing the marginal or central area of the image. Instead of using longitude θ, we propose $\theta_{proposed}$ which is multiplication of w and θ, as shown in Eq. 1. The weight w consists of scale factor α for determining the expansion, and β for balancing the effect of the edge areas, as illustrated in Eq. 2.

$$\theta_{proposed} = w\theta \tag{1}$$

$$w = \alpha + \beta \frac{|\theta|}{\theta_{max}} \tag{2}$$

Fisheye lens camera commonly captures a circular image in a rectangular image sensor. Since the width-to-height ratio is greater than 1, sides of the image contain blank regions. Using this margin and general location of large and small objects, we decide whether to expand center or periphery area, manually setting the value of α and β.

θ is a longitude from the spherical coordinate, and θ_{max} is half the field of view. When warping the image,

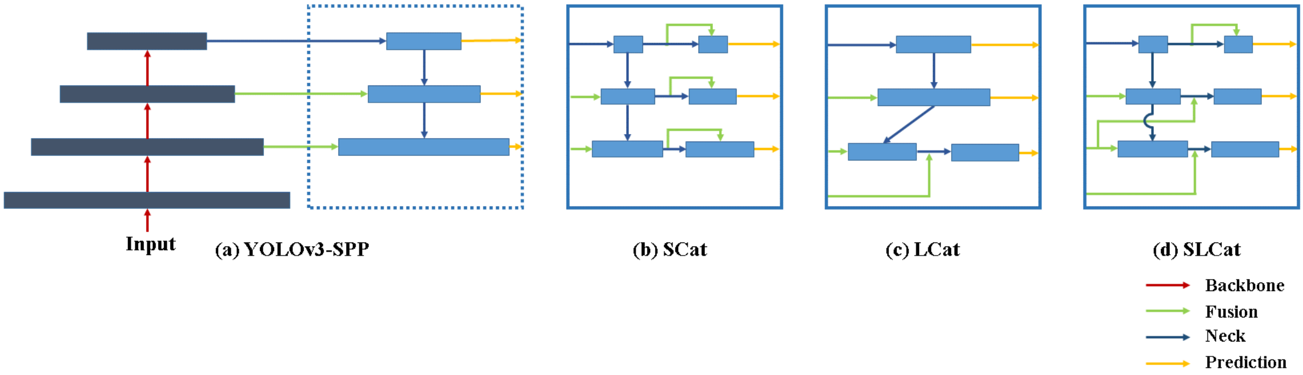

Figure 2. Different feature fusion strategies. (a) YOLOv3 with Spatial Pyramid Pooling module (b) Short-skip Concatenation (c) Long-skip Concatenation (d) Short-Long-skip Concatenation

we set the center of the image as $(0, 0)$ in the coordinate system. Then, $\frac{|\theta|}{\theta_{max}}$ represents whether a projected point is placed near the middle or boundary regions.

When $\theta \rightarrow 0$, the location is near the center of the image with $w \cong \alpha$. $\theta \rightarrow \theta_{max}$ indicates the margin area and $w \cong (\alpha + \beta)$. The area of θ is stretched more as $w \rightarrow 0$, while $w > 1$ projects the specific area narrower. In synthetic datasets where most of large objects located near the middle part of the image, while small objects frequently appears around the edge, α and β is set as 1.2 and -0.3 respectively. In case of real fisheye image, smaller objects are frequently around the center area. Therefore we extend more on the center area with $\alpha = 0.7$ and $\beta = 0.17$. These parameters guarantee the projected pixels within the bound of the image.

3.3 Concatenated YOLOv3-SPP

We employ YOLOv3 architecture with one SPP block, which consists of 4 parallel max-pooling layers with 4 different kernel sizes. Since it can extract deep features with increased receptive fields with a comparable speed, we select YOLOv3-SPP for testing object detection from fisheye and undistorted images.

To efficiently detect small objects for both synthetic and real datasets, we suggest additional concatenation modules to YOLOv3-SPP with three variants for extracting more meaningful information about small objects. Fig. 2a is a baseline model which predicts bounding boxes at three scales. SCat(Fig. 2b) is for concatenating with short skip-connection on the neck part of the object detector, following five convolutional layers for each feature fusion module. LCat merges feature maps with longer skip-connection with adding more local features from the backbone at the bottom prediction. Finally SLCat uses the fusion methods to neck and backbone, combining core features of previous two approaches. Details of the proposed architectures will be in Fig.1 of the supplement material.

4 Experimental Results

4.1 Experiments on Fisheye-CityScape

Implementation details of training will be provided in the supplemental material. In Table 4.1, the proposed projection outperforms by more than 3% in AP at SCat and LCat model, while the accuracy from basic spherical projection is similar as fisheye images. In addition, both projections increase the results on small objects as well.

Compared with feature fusion methods, SCat increases the accuracy by 1.5% in expandable spherical images. For SLCat model, it consistently achieves higher AP and AP_S than YOLOv3-SPP. In addition, SLCat with our projection increases 1.9% in AP_S compared to the baseline model with fisheye image dataset. From our undistorted datasets, LCat shows less improvement in AP.

Table 1. Accuracy of different projections and feature- fused models on synthetic Fisheye-CityScape

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
			Fisheye Image			
Baseline	22.4	42.5	20.7	5.6	28.0	58.7
SCat	22.4	42.3	20.3	5.9	29.0	57.6
LCat	21.0	39.7	19.9	4.7	26.0	61.1
SLCat	22.9	44.0	21.5	6.1	28.1	60.9
			Spherical Projection			
Baseline	21.2	40.0	18.1	6.2	28.2	56.8
SCat	22.9	42.9	20.5	6.2	30.8	63.0
LCat	22.4	41.7	21.0	6.2	30.0	62.3
SLCat	22.8	42.5	21.3	6.9	31.0	58.4
			Expandable Spherical Projection			
Baseline	24.3	45.5	22.3	6.9	31.0	66.9
SCat	25.8	47.2	23.9	7.3	34.3	58.5
LCat	24.0	45.0	21.7	6.5	31.6	57.2
SLCat	24.8	46.0	23.0	7.5	31.0	66.3

4.2 Experiments on Fisheye-KITTI

Compared to fisheye images with baseline model, our proposed projection achieves the best performance in AP, significantly improving the accuracy up to 4.7%. On average, the detection result increased by 4.45% in AP.

For the feature-fused models, SLCat shows better accuracy in AP from projected images, and obtains 0.6% higher result in AP_S at the fisheye images. In this KITTI dataset, our concatenation methods in spherical-based projections show less effective in AP_S.

Table 2. Accuracy of different projections and feature- fused models on synthetic Fisheye-KITTI

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Fisheye Image without any projection						
Baseline	56.9	85.7	63.8	48.2	66.6	73.9
SCat	57.2	86.6	64.2	49.1	66.8	74.7
LCat	57.0	86.4	63.1	48.7	66.4	76.9
SLCat	56.9	85.5	63.7	48.8	66.4	76.2
Spherical Projection						
Baseline	59.1	86.5	66.2	46.5	64.6	75.2
SCat	58.8	86.9	66.5	47.3	64.5	75.7
LCat	58.8	86.4	65.1	44.5	65.0	76.9
SLCat	59.5	86.0	67.6	46.9	65.4	75.4
Expandable Spherical Projection						
Baseline	61.3	88.3	70.7	48.2	65.7	76.4
SCat	61.5	88.8	69.6	48.2	66.0	76.3
LCat	61.4	88.9	70.0	47.8	65.6	75.6
SLCat	61.6	88.2	71.0	48.0	65.9	76.0

4.3 Experiments on Fisheye-Dongseongno

Table 3. Accuracy of different projections and feature- fused models on real Fisheye-Dongseongno

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Fisheye Image						
Baseline	38.2	75.5	31.8	12.3	38.2	52.3
SCat	38.7	77.0	33.4	12.9	38.6	56.2
LCat	39.7	76.4	35.8	14.3	39.2	54.8
SLCat	40.1	78.6	35.7	13.5	39.6	56.1
Spherical Projection						
Baseline	38.4	76.5	32.9	26.4	46.5	57.0
SCat	38.1	76.1	32.3	26.8	45.1	58.2
LCat	39.0	75.8	32.5	28.0	46.4	61.8
SLCat	38.2	76.9	32.4	27.3	44.8	57.8
Expandable Spherical Projection						
Baseline	40.8	80.2	36.3	28.5	46.5	61.5
SCat	40.4	80.3	34.2	28.8	46.3	56.8
LCat	39.3	77.8	34.9	27.2	45.2	59.6
SLCat	40.8	79.6	37.6	31.0	45.6	56.6

From the experimental results in Table 4.3, the proposed expandable projection shows higher detection result in AP from all models except LCat. At YOLOv3-SPP model, our method achieves 2.6% improvement in AP. Moreover, this projection successfully improves AP_S up to 17.5% at SLCat model.

Compared with the feature concatenation methods, SLCat obtains best performance on the expandable spherical images, and achieves 1.9% higher AP than the baseline on the fisheye image dataset. On the other hand, AP in LCat decreased than the baseline from our projection images, same as Fisheye-CityScape. We assume merging the features with too low-level details can hinder the network from correctly extracting relevant features.

4.4 Ablation Experiments

Computational time of Projection: We obtain same computational time from both projections. In Table 4, time for generating rectification map is presented as 0.256 second, and de-warping per image takes 0.0155 second.

Inference time of the models: Table 5 shows the inference time in Titan V GPU with the input size 512x512 and 640x640.

Table 4. Computational time [sec] of the expandable spherical method. Rectification map is generated only once and dewarping is implemented per image.

Projection	Rectification map	Dewarping
Spherical	0.256	0.0155

Table 5. Inference time [sec] of the network by input size in Titan V GPU

Model	512	640
Baseline	0.210	0.233
SCat	0.249	0.248
LCat	0.213	0.233
SLCat	0.234	0.237

5 Conclusion

We presented the effective spherical projection with expansion weight on real-time object detection task. Using two scale parameters, center or margin areas of spherical images are expanded for reducing the distortions. We also analyzed the effect of feature fusion methods on small object detection. Finally, we provide a raw fisheye dataset from one front view camera for autonomous driving with 2D bounding box annotation files.

References

[1] B. Coors, A. P. Condurache, and A. Geiger, “Spherenet: Learning spherical representations for detection and classification in omnidirectional images,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 518–533.

[2] B. Arsenali, P. Viswanath, and J. Novosel, “Rotinvmtl: Rotation invariant multinet on fisheye images for autonomous driving applications,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[3] Z. Chen and A. Georgiadis, “Learning rotation sensitive neural network for deformed objects' detection in fisheye images,” in *2019 4th International Conference on Robotics and Automation Engineering (ICRAE)*. IEEE, 2019, pp. 125–129.

[4] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[6] I. Baris and Y. Bastanlar, “Classification and tracking of traffic scene objects with hybrid camera systems,” in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[7] I. Cinaroglu and Y. Bastanlar, “A direct approach for human detection with catadioptric omnidirectional cameras,” in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2014, pp. 2275–2279.

[8] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende *et al.*, “Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9308–9318.

[9] L. Deng, M. Yang, Y. Qian, C. Wang, and B. Wang, “Cnn based semantic segmentation for urban traffic scenes using fisheye camera,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 231–236.

[10] G. Blott, M. Takami, and C. Heipke, “Semantic segmentation of fisheye images,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[11] J. Fu, I. V. Bajić, and R. G. Vaughan, “Datasets for face and object detection in fisheye images,” *Data in brief*, vol. 27, p. 104752, 2019.

[12] A. Sáez, L. M. Bergasa, E. Romeral, E. López, R. Barea, and R. Sanz, “Cnn-based fisheye image real-time semantic segmentation,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1039–1044.

[13] Á. Sáez, L. M. Bergasa, E. López-Guillén, E. Romera, M. Tradacete, C. Gómez-Huélamo, and J. Del Egido, “Real-time semantic segmentation for fisheye urban driving images based on erfnet,” *Sensors*, vol. 19, no. 3, p. 503, 2019.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[15] K. Tong, Y. Wu, and F. Zhou, “Recent advances in small object detection based on deep learning: A review,” *Image and Vision Computing*, vol. 97, p. 103910, 2020.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[17] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, “Feature-fused ssd: Fast detection for small objects,” in *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, vol. 10615. International Society for Optics and Photonics, 2018, p. 106151E.

[18] P.-Y. Chen, J.-W. Hsieh, M. Gochoo, C.-Y. Wang, and H.-Y. M. Liao, “Smaller object detection for real-time embedded traffic flow estimation using fish-eye cameras,” in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2956–2960.

[19] E. Plaut, E. B. Yaacov, and B. E. Shlomo, “Monocular 3d object detection in cylindrical images from fisheye cameras,” *arXiv preprint arXiv:2003.03759*, 2020.

[20] A. Sharma and J. Ventura, “Unsupervised learning of depth and ego-motion from cylindrical panoramic video,” in *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019, pp. 58–587.

[21] H. Houshiar, J. Elseberg, D. Borrmann, and A. Nüchter, “A study of projections for key point based registration of panoramic terrestrial 3d laser scan,” *Geo-spatial Information Science*, vol. 18, no. 1, pp. 11–31, 2015.

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[23] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.