Where can distinguishing features be extracted in an image for visibility estimate?

han wang[1], Jia Li Liu[1], Ke Cheng Shen[1], and Quan Shi[1]

[1]Nantong University

February 22, 2024

Abstract

Standard convolution is difficult to provide an effective fog feature for visibility estimate tasks due to the fixed grid kernel structure. In this paper, a multiscale deformable convolution model (MDCM) is proposed to extract features that make effectively sampling discriminating features from the atmospheric region in foggy image. Moreover, to enhance performance we use RGB-IR image pair as observations and design a multimodal visibility range classification network based on the MDCM. Experimental results show that both the robustness and accuracy of visibility estimate performance are raised beyond 30% compared to standard convolutional neural networks (CNNs).

Where can distinguishing features be extracted in an image for visibility estimate?

Han Wang[1], JiaLi Liu[1], KeCheng Shen[2], Quan Shi[1]

[1] *School of Transportation and Civil Engineering, Nantong University, China*
[2] *School of Information Science and Technology, Nantong University, China*

Email: hanwang@ntu.edu.cn

Standard convolution is difficult to provide an effective fog feature for visibility estimate tasks due to the fixed grid kernel structure. In this paper, a multiscale deformable convolution model (MDCM) is proposed to extract features that make effectively sampling discriminating features from the atmospheric region in foggy image. Moreover, to enhance performance we use RGB-IR image pair as observations and design a multimodal visibility range classification network based on the MDCM. Experimental results show that both the robustness and accuracy of visibility estimate performance are raised beyond 30% compared to standard convolutional neural networks (CNNs).

Introduction: Image visibility estimation is a method to determine the visibility range according to the atmospheric transparency in the image scene. Because of the strong ability of CNNs to describe image features, image visibility estimate based on deep learning has become a research hotspot. For example, Li, et al. [1] used AlexNet [2] to extract features and designed a GRNN for visibility evaluation based on these deep learning features. You, et al. [3] used CNN-RNN to feature extraction and made a relative visibility evaluation by relative SVM. Palvanov, et al. [4] proposed multi-branch parallel network to enhance the accuracy of visibility estimate. Song, et al. [5] proposed a deep label distribution model to realize effective and efficient visibility estimation. The kernel used in the above methods were fixed regular grid structures.

Fig.1 shows visualization [6] results under different convolution models including standard convolution and proposed multiscale deformable convolution model as Fig.1(a) and Fig.1(b) shows respectively. We can find that the key features of standard kernel are concentrated in non-atmospheric areas, such as buildings and trees, like Fig.1(d) shown. In this paper, to extract effective features directly from the atmospheric region in the image, a channel attention model is utilized to adaptively merge the offset regions of different scale deformable convolution kernel [7] for completing multiscale deformable convolution model (MDCM), like Fig.1(b) shown. As a result, MDCM can directly use the shape and transparency of the atmospheric area to estimate and classify the visibility range like Fig.1(e) shows.

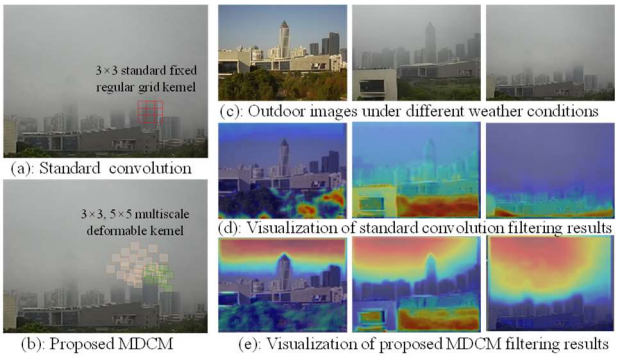

Fig 1 Feature visualization results of various convolution kernels.

Multimodal visibility estimation based on MDCM: Fig.2 shows an illustration of multiscale deformable convolution, the grid $R_{3\times3}$ and $R_{5\times5}$ respectively defines the two different field sizes. $R_{3\times3}=\{(-1,-1),(-1,0),\ldots,(1,1)\}$ defines a 3×3 kernel with dilation 1 and $R_{5\times5}=\{(-2,-2),(-2,-1),\ldots,(2,2)\}$ defines a 5×5 kernel with dilation 2. The output feature $y_{3\times3}(P_0)$ on the top 3×3 kernel convolution branch is

$$\boldsymbol{y}_{3\times3}(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in\boldsymbol{R}_{3\times3}} w_{3\times3}\ (\boldsymbol{P}_n)\cdot\mathbf{x}(\boldsymbol{P}_0+\boldsymbol{P}_n+\Delta\boldsymbol{P}_n^{3\times3}) \quad (1)$$

where $P_n^{3\times3}$ enumerates the locations in $R_{3\times3}$ and $\Delta P_n^{3\times3}$ is the corresponding offset, n∈[1,9]. x(P) is P position pixel value and $w_{3\times3}$ (P_n) is kernel weight value. Similarly, the down branch 5×5 kernel convolution feature $y_{5\times5}(P_0)$ is

$$\boldsymbol{y}_{5\times5}(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in\boldsymbol{R}_{5\times5}} w_{5\times5}\ (\boldsymbol{P}_n)\cdot\mathbf{x}(\boldsymbol{P}_0+\boldsymbol{P}_n+\Delta\boldsymbol{P}_n^{5\times5}) \quad (2)$$

where $P_n^{5\times5}$ enumerates the locations in $R_{5\times5}$ and $\Delta P_n^{5\times5}$ is the corresponding offset, n∈[1,25]. x(P) is P position pixel value and $w_{5\times5}$ (P_n) is 5×5 kernel weight value. Therefore, the output of MDCD is the combination of two-scale kernel deformable convolution results.

$$\boldsymbol{y}_m(\boldsymbol{P}_0)=w(\boldsymbol{k}_{3\times3}^{\boldsymbol{m}})\cdot\boldsymbol{y}_{3\times3}(\boldsymbol{P}_0)+w(\boldsymbol{k}_{5\times5}^{\boldsymbol{m}})\cdot\boldsymbol{y}_{5\times5}(\boldsymbol{P}_0) \quad (3)$$

where, $w(k_{3\times3}^m)$ is the channel important weight of the m-th 5×5 convolution kernel and $w(k_{5\times5}^m)$ is the channel important weight of the m-th 5×5 convolution kernel. They are both assigned from a block attention module [8].

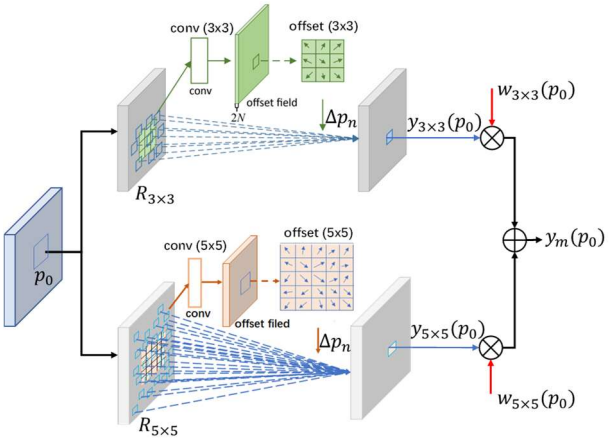

Fig 2 Illustration of multiscale deformable convolution.

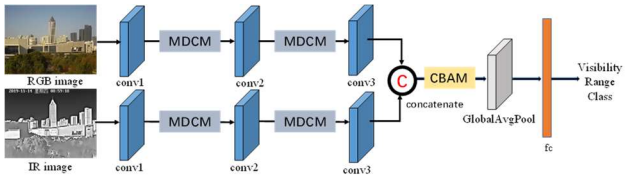

Fig 3 The structure of multimodal visibility range estimate network.

The main structure of the proposed multimodal visibility deep learning model as Fig.3 shows. The feature extraction streams consist of infrared feature stream and visible feature stream, which are connected in parallel to ensure the accuracy and robustness of the parallel network as whole during classification. It receives an input visible-infrared image pair, classifies them, and produces a visibility range.

The structures of the infrared and visible feature streams are identical, but their corresponding inputs are infrared and visible images, respectively. The two streams both consist of three convolutional layers and two MDCMs. Then feature maps of two streams are concatenated and forwarded to a convolutional block attention module (CBAM) [8]. Finally, the CBAM weighted feature map is aggregated by global average pooling and forwarded to a fully connected layer then fed into the final classification layer, in which the softmax function is used.

Experimental Results: To evaluate and verify the proposed method using real-world images, a Hikvision binocular camera is used to build the experimental dataset, with a raw image pair consisting of a visible (1092×1080) and an infrared image (384×288). We then use a visibility meter to create corresponding visibility range labels for all of the experimental image pairs as Fig.4 shows. As a result, a dataset consisting of 5208 RGB-IR image pairs with a 384×288 resolution is collected as our dataset that covers ranges from 0 km to 10 km, presented with 7 classes. With the number of classes increase the visibility becomes worse.

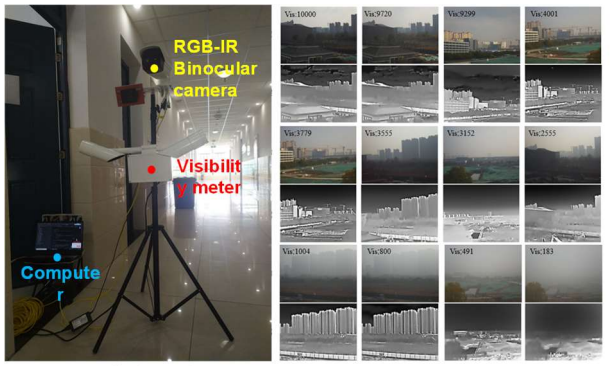

Fig 4 Experimental devices and an example of the raw data.

To measure accuracy, we employ the average accuracy of 5 times performance and a variance score is utilized as a robustness measure. The comparison experiments are set up as below. First, the experimental dataset is randomly divided into training and test data by 50%-50%. Then, all comparison models are trained with a random initial network parameter based on the training data and the trained models are employed in visibility classification and their accuracy is recorded. Finally, we repeat this procedure for 5 times and evaluate the performance using the pre-defined two measures.

To fairly evaluate the performance of the proposed MDCM and other conventional methods, we use two layers backbone network and apply three different convolution kernels including MDCM, DCN [7] and standard fixed grid kernel [9] as the comparison object to visibility range classification task with our dataset. The experimental results are shown in Table 1. It clear that MDCM improves the accuracy by 36% compared with standard convolution.

Table 1. Performance for visibility estimate with different kernel.

Kernel type	RGB	IR
Standard convolution [9]	59.4%	65.57%
Deformable convolution [7]	90.1%	91.0%
Proposed MDCM	**96.4%**	**96.8%**

Why can proposed MDCM greatly improve the visibility rage classification? Fig.5 shows the visualization results of the convolution features for three kinds of kernels under RGB-IR input image pairs. By comparison, the key features of fixed grid convolution kernel are all concentrated in the non-atmospheric area of the image, such as buildings or trees. For deformable convolution kernel, the key feature points of the first and the third input images are extracted from the atmospheric region, but the feature extraction regions of rest input images are the same as the standard convolution. Obviously proposed MDCM can effectively use CBAM [8] to adjust the position of each kernel element, so that most of the key feature extraction areas are concentrated in the atmospheric area. That is, the visibility feature can be described directly by using the atmospheric color, transparency and shape information.

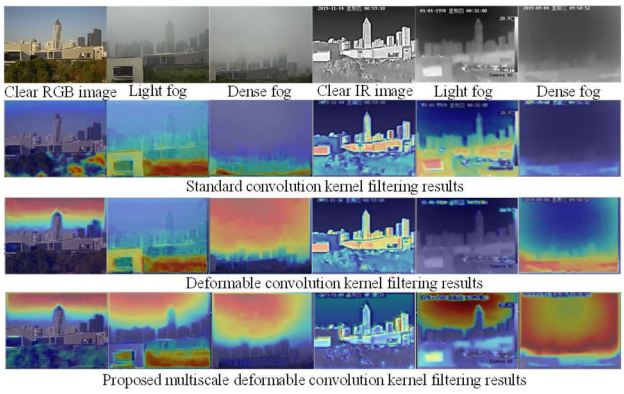

Fig 5 Feature visualization with different convolution type kernel based on visible and infrared image observations.

To illustrate the advantage of multimodal observations for visibility range classification, Fig.6 shows we apply five different models and evaluate their performance of five training result with random initial network parameters.

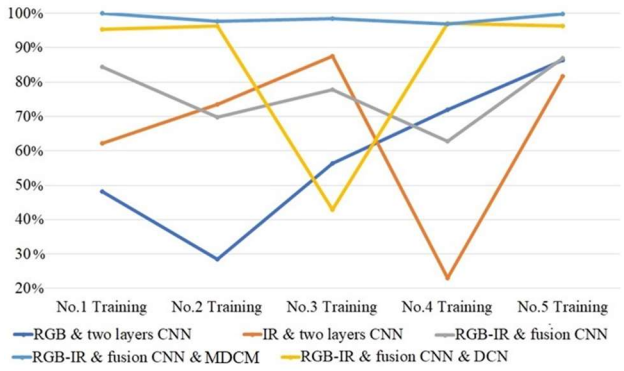

Fig 6 Robustness performance from 5 training runs for the different models.

The comparison shows that the accuracy of the single input model is easily affected by the initial weight value. Therefore, the curve shows an obvious oscillation phenomenon just like "RGB & two layers CNN" and "IR & two layers CNN" shows. The multi-modal input significantly improves the amplitude of oscillation, just like "RGB-IR & fusion CNN" and "RGB-IR & fusion CNN & DCN". Especially with the help of MDCD, the accuracy of the visibility model is little affected by the initial weight of the network, and its curve is relatively flat like "RGB-IR & fusion CNN & MDCD" shows.

We compare our proposed multimodal network with 7 different deep learning models for the task of visibility range classification including a classical two layers CNN model with different inputs (RGB, IR), AlexNet [2], ResNet [11], CNN-RNN [3], VisNet [4] and a feature level fusion model FusionNet [10] with RGB-IR input.

Table 2. Performance of visibility range classification model.

Model	Accuracy	Robustness
RGB & CNN (two layers)	59.4%	0.0491
IR & CNN (two layers)	65.5%	0.0657
RGB & AlexNet [3]	83.24%	0.0212
RGB & Resnet [11]	85.39%	0.0091
RGB & CNN-RNN [3]	82.90%	0.0199
RGB & VisNet [4]	86.53%	0.0072
RGB-IR & FusionNet [10]	95.70%	0.00238
Proposed model	**99.6%**	**0.00001**

Table 2 shows that the visibility model based on single input image and simple network structure has the lowest accuracy. The complex network structure can improve the accuracy of visibility model by 17% - 20%. It also shows that the model based on multimodal input image combined with feature fusion network can improve the accuracy by more than 30%. The proposed multimodal visibility model is shown to achieve the highest accuracy and robustness compared to other conventional methods.

Conclusion: In this paper, the influence of convolution kernel pattern on visibility range classification was deeply analyzed and discussed. The visibility model with traditional fixed grid convolution kernel found the sampling area of key feature points was concentrated in the non-atmospheric area of the image, such as buildings and trees. That is, those visible and texture rich areas in the scene were used as the basis for visibility representation. However, atmospheric transparency, color and boundary shape were found as the direct feature to describe visibility. Hence, a multiscale deformable convolution kernel was established. Experiments showed that the proposed MDCM can effectively concentrate most of the sampling areas of key features in the atmospheric area of the image. Moreover, we built and demonstrated a multimodal visibility model based on MDCM and RGB-IR input, which significantly improved the accuracy and robustness of visibility range classification.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (NSFC) Grant 61872425

Received: 12 June 2022 *Accepted:* xx March 2022
doi: 10.1049/ell2.10001

References

1. Li, S.Y., Fu, H., Lo, W.L.: 'Meteorological Visibility Evaluation on Webcam Weather Image Using Deep Learning Features', International Journal of Computer Theory and Engineering, 2017, 9, (6), pp. 455–461
2. Alex, K., Ilya, S., Geoffrey, E. H.: 'ImageNet Classification with Deep Convolutional Neural Networks', Conf. NIPS., NY, USA, 2012, pp. 1097–1105
3. Yang, Y., et al.: 'Relative CNN-RNN: Learning Relative Atmospheric Visibility from Images', IEEE Trans. Image Process, 2019, 28, (1), pp.45–55
4. Palvanov, A., and Cho, Y. I.: 'VisNet: Deep Convolutional Neural Networks for Forecasting Atmospheric Visibility', Sensors, 2019, 19, (1343), pp.1–34
5. Song, M.F., et al.: 'Visibility estimation via deep label distribution learning in cloud environment', Journal of Cloud Computing: Advances, Systems and Applications, 2021, 10, (46), pp.1–14
6. Matthew, D., and Zeiler, R.: 'Visualizing and Understanding Convolutional Networks', Conf. ECCV., Zurich, Switzerland, 2014, pp. 818–833
7. Dai, J.F., et al.: 'Deformable Convolutional Networks', Conf. CVPR, Honolulu, HI, USA, 2017, pp. 174–181
8. Woo, S.Y., et al.: 'CBAM: Convolutional Block Attention Module', Conf. ECCV, Munich, Germany, 2018, pp. 3–19
9. Gao, L.G., Chen, P.Y., Yu, S.M.: 'Demonstration of convolution kernel operation on resistive cross-point array', IEEE Electron Device Letters, 2016, 37, (7), pp.870–873
10. Eitel, A., et al.: 'Multimodal deep learning for robust RGB-D object recognition', Conf. IROS, Hamburg, Germany, 2015, pp. 681–687.
11. He, K. M., et al.: 'Deep Residual Learning for Image Recognition', Conf. CVPR, Las Vegas, NV, USA 2016, pp. 770–778.