Unsupervised Multi-Constraint Deep Neural Network for Dense Image Matching

W. Yuan[1, 2, *], Z. Fan[2], X. Yuan [1], J. Gong [1], R. Shibasaki [2]

[1] Center for Spatial Information Science, University of Tokyo, 5-1-5, Kashiwa, Chiba, Japan – (miloyw,shiba,fanzipei)@csis.u-tokyo.ac.jp
[2] Wuhan University, School of Remote Sensing and Information Engineering, 129 Luoyu Road Wuhan, Hubei, China - (yuanxx, jygong)@whu.edu.cn

Commission II, WG II/2

KEYWORDS: Multi-Constraint, Unsupervised Learning, Dense Image Matching, Deep Neural Network, Photo Consistency, Matching Accuracy;

ABSTRACT:

Dense image matching is essential to photogrammetry applications, including Digital Surface Model (DSM) generation, three dimensional (3D) reconstruction, and object detection and recognition. The development of an efficient and robust method for dense image matching has been one of the technical challenges due to high variations in illumination and ground features of aerial images of large areas. Nowadays, due to the development of deep learning technology, deep neural network-based algorithms outperform traditional methods on a variety of tasks such as object detection, semantic segmentation and stereo matching. The proposed network includes cost-volume computation, cost-volume aggregation, and disparity prediction. It starts with a pre-trained VGG-16 network as a backend and using the U-net architecture with nine layers for feature map extraction and a correlation layer for cost volume calculation, after that a guided filter based cost aggregation is adopted for cost volume filtering and finally the soft Argmax function is utilized for disparity prediction. The experimental conducted on a UAV dataset demonstrated that the proposed method achieved the RMSE (root mean square error) of the reprojection error better than 1 pixel in image coordinate and in-ground positioning accuracy within 2.5 ground sample distance. The comparison experiments on KITTI 2015 dataset shows the proposed unsupervised method even comparably with other supervised methods.

1. INTRODUCTION

Dense image matching is essential to photogrammetry applications, including Digital Surface Model (DSM) generation, three dimensional (3D) reconstruction, and object detection and recognition (Xu et al, 2017). Given a pair of stereo images and its corresponding camera parameters, the goal of dense image matching is to generate the 3D point clouds of the overlap area between the stereo image pair (Szeliski, 2010). The development of an efficient and robust method for dense image matching has been one of the technical challenges due to high variations in illumination and ground features of aerial images of large areas. Nowadays, due to the development of deep learning technology, deep neural network-based algorithms outperform traditional methods on a variety of tasks such as object detection, semantic segmentation and stereo matching. On the stereo matching benchmark KITTI, the top 50 methods on the rank are deep learning-based, which show a significant advantage by utilizing deep neural networks for dense image matching tasks (Geiger, 2012). However, almost all the deep neural network-based methods are supervised learning methods, because of all the benchmarks provide accurate ground truth for validation. In real 3D data production, the accurate 3D points cloud ground truth is usually obtained by LiDAR system, which is really expensive for handling a large survey area (Yuan et al, 2019). Due to the difficulty of labeling ground truth depth, usable data for training a network is rather limited, making the supervised learning-based methods are difficult to apply to real applications (Zhou et al, 2017). To tackle the above problems, in this paper, we present an end-to-end unsupervised multi-constraint Deep Neural Network for aerial image-based dense image matching.

2. RELATED WORKS

The early stereo matching methods (Kong, 2014) defined the initial matching cost calculated using various metrics (such as the Euclidean distance of pixel values). It is also very popular to fit the hyperparameters of graphic models (Zhang, 2007) with real ground data. In recent work, the confidence of the estimated matching cost has been studied (Spyropoulos, 2014). These methods train a random forest classifier to combine several confidence measures or a Markov random domain.

In recent years, as deep neural networks show a great performance on object detection and classification tasts. Convolutional neural networks (CNN) have also been widely used in matching cost learning. Zbontar and LeCun (2015) describe patch based matching as a binary classification problem to determine the pixel-wise correspondence use deep neural networks. Later methods followed this work improved network architecture, such as MatchNet (Han, 2015).

In order to take full advantage of the limited size and matching range of stereo depth, the researchers established more specific architectures and losses. GCNet(Alex 2017) proposed a method of generating 3D cost by intensively comparing the features on the pixels of the reference image with all possible matching pixels on the target image. The network finds the best match through soft argmin operator. PSMNet (Chang, 2018) uses pyramid space pools and hourglass networks to utilize image context. Later work (Cheng, 2018) added a post-processing module, resulting in better recovery details. These network

* Corresponding author

architectures provide a solid foundation for developing unsupervised learning methods.

Unsupervised learning is becoming more and more popular in deep learning researches. In moving object prediction, autoencoders and visual representations, a lot of researcher choose to using unsupervised learning based methods. Unsupervised learning also works well with edge detection and optical flow calculation. Li et al. (2015) utilize the correlation between moving boundaries for object edges detections. The alternation between motion estimation and edge detection forms the basic steps of edge learning. Yu et al. (2016) defines a loss function with data and smoothing terms, similar to the objective function in the energy minimization strategy. Therefore, this method is utilizing the loss function to determing the optimal solution of whole task. In contrast, most of the dense image matching methods defines matching as a classification problem, which can be better solved in the occlusion area.

3. METHOD

3.1 Dense Matching Learning Network

Our Dense Matching Learning Network is an end-to-end trainable framework taking stereo image pairs as input and output the predicted disparity map. The whole network can be mainly divided into three parts, feature extraction parts, cost-volume computation parts, and disparity prediction part. At first, the stereo image patches are utilized as the input for feature extractor, the feature extract layer are mainly conducted with 9 layers, after the feature extract, the extracted features are input into a correlation layer for cost volume calculation. Since the pixel based cost volume always affected by the pixel intensity noise and brightness changes, we employ a guided image filter to represent the cost aggreation to enhance the robustness of the estimated costs. At last, a soft Argmax operator is utilized for disparity prediction.The detailed explanation in what follows.

3.1.1 Feature extraction

We first employ a U-net (Ronneberger, 2015) like architecture for feature extraction, because common CNNs usually utilizing the sliced image patches as inputs and in our case we want using the whole raw image as inputs, further more U-net has a strong performance in object detection tasks, which makes it fulfil our requirements. The utilized U-net architecture contained with nine layers, the second to fifth layer are downsampled layer with PReLU activation function, and the fifth to ninth layer are the upsample layer with PReLU function and batch normalization.

3.1.2 Cost-Volume Computation

After the Unet liked feature extraction layer, a correlation layer same as Mayer (Mayer et al, 2016) proposed in their paper are employed for the cost-volume computation, at the meantime, an edge preserved guided image filter (He et al, 2013) base on the left image is employed to joint filter the cost-volume and the cost aggregation step is finished.

3.1.3 Disparity Prediction

After getting the filtered cost volume, the winner-takes-all strategy is utilized for disparity prediction. As the usually utilized Argmax operator is not derivable in the backpropagation. Here we use the soft Argmax operator to solve the problem (Chapelle et al., 2010).
With all the three components, the proposed networks can learn the dense image matching in an end-to-end manner.

3.2 Multi-Constraint Loss Function

The common supervised learning methods can use a large amount of ground truth data for loss function determination; however, for unsupervised learning, only a reasonable loss function can lead to good prediction results. To determine the loss function in our unsupervised learning networks, we chose a three-term based multi-constraints for optimization of the prediction results.

3.2.1 Reconstruction Appearance Loss

Through the learning networks the input left image I^l can be easily wrapped from input right image I^r with the predicted disparity map d^l, then we use the appearance loss function to encourage the predicted disparity and the right image can reconstruct the same image as the original left image. In order to justify the similarity between two images, we use the structure similarity function proposed by Wang (2004) and the L1 loss to build our appearance loss. The loss can be described as

$$L_{ap}^{l} = \frac{1}{N}\sum_{p\in\Phi}\alpha\frac{1-\sigma(I^l,\hat{I}^l)}{2}+(1-\alpha)\left\|I^l(p)-\hat{I}^l(p)\right\| \quad (1)$$

Where Φ is a 5×5 neighborhood of pixel p and the $\sigma(I^l,\hat{I}^l)$ is the image structure similarity function from Wang's paper, α is a weighted factor, and $\hat{I}^l$ is the wrapped image based on the predicted left disparity map and the original right image.

3.2.2 Left-right Consistency Loss

In the traditional dense image matching methods, L-R check is an important step to eliminate the mismatched correspondence, thereby, we determine a same Left-right consistency loss to regularize the consistency of left disparity map and the right disparity map. The Left-right Consistency Loss can be described as

$$L_{LR}^{l} = \frac{1}{N}\sum_{p\in\Phi}\left|d^l(p)-d^r(p-d^l(p))\right| \quad (2)$$

Where Φ is a 5×5 neighborhood of pixel p, d^l represent the left disparity map, and d^r represent the right disparity map.

3.2.3 Smoothness Loss

The smoothness loss term is utilized to enforce the disparity smoothness, and it can be described as equation 3.

$$L_{Smooth}^{l} = \frac{1}{N}\sum_{p\in\Phi}\left|\partial_x d^l\right|e^{-\left\|\partial_x I^l\right\|}+\left|\partial_y d^l\right|e^{-\left\|\partial_y I^l\right\|} \quad (3)$$

Where ∂d^l and ∂I^l are the gradient value of d^l and I^l.

After all the final unsupervised loss function is determined as equation (4).

$$L = \alpha(L_{ap}^{l} + L_{ap}^{r}) + \beta(L_{LR}^{l} + L_{LR}^{r}) + \gamma(L_{Smooth}^{l} + L_{Smooth}^{r}) \quad (4)$$

Where α , β and γ are the weighted factor to balance the influence of a different kind of the loss.

4. EXPERIEMT AND ANALYSIS

4.1 Datasets

The evaluative experiments were performed using sets of unmanned aerial vehicle images. The UAV image dataset is contained with 44 images in near Beijing area, and mainly covered with farmland and small buildings. The detailed parameters of the dataset is shown in table 1. In order to evaluate the actual positioning accuracy, we use the high accurate pass points as the checkpoints for actual positioning accuracy evaluation. And the comparison experiment is conducted on the KITTI 2015 dataset.

Table 1. Technical parameter of the test images

Item	Beijing
Aerial craft	Unmanned Aerial Vehicle (UAV)
Camera	PhaseOne IXU-1000
Principal distance (mm)	51.21293
Format (pixels)	11608 × 8708
Pixel size (μm)	4.6
Ground sample distance (GSD) (cm)	7
Relative flying height (m)	779
Longitudinal overlap (%)	60
Lateral overlap (%)	30
Number of mapping strips	4
Number of control strips	4
Number of images	88
Number of ground control points	21
Number of pass points	55701
Block area (km2)	2.8 × 2.8
Maximum topographic relief (m)	54
Average terrestrial height (m)	508

4.2 Quality assessment

The exact position of the flying strips and the ground control points are shown in Fig.1. To quantitatively analyze the accuracy of the proposed method, the root mean square error(RMSE) of the reprojection error is utilized to determine the matching accuracy in image coordinate. The calculation of the reprojection error is shown as equation (5).

$$m_0 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((x_i - \hat{x}_l)^2 + (x_i' - \hat{x}_l')^2)} \quad (5)$$

where m_0 is the matching accuracy in the image; n is the number of the dense image matching points; x_i, $\hat{x}_i$, x_i', and $\hat{x}_i'$ are the coordinate vectors.

The actual height accuracy in the ground is calculated by using the pass points as the checkpoints and calculate the height displacement between the checkpoints and the dense image matching generated 3D points. The actual height accuracy can be calculated through equation (6):

$$\mu = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta h_i^2} \quad (6)$$

where n is the number of checkpoints and Δh_i is the height error of the ith checkpoint.

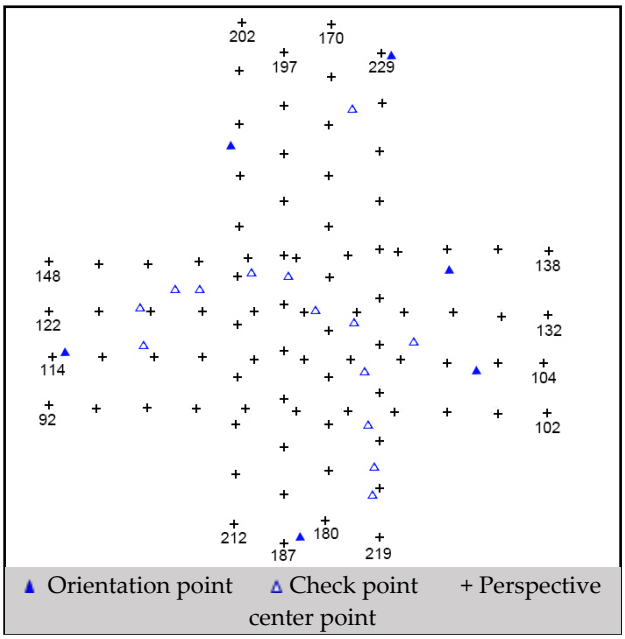


Figure 1. The Distribution of the Beijing UAV datasets

4.3 Analysis

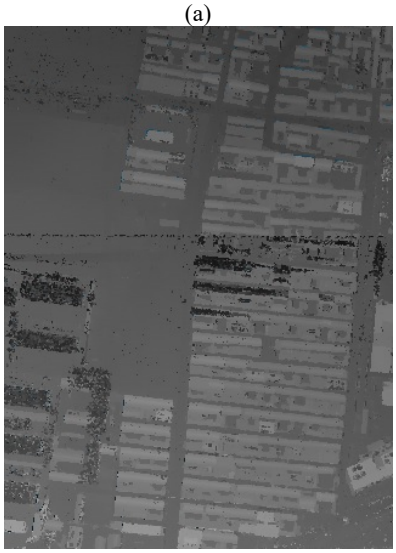
Figure 2 shows the original input left image and the predict disparity map. From the disparity map we can find the overall completeness is high, but the disparity map still contained some noise. By using the camera parameter, we calculate the overall RMSE of the projection error among the 87 stereo pair is better than 1 pixel, and the actual positioning accuracy achieved 18 cm, which is better than 2.5 GSD.

The generated disparity map is the raw output with out any post-processing. We believe that using those outputs by utilizing some filtering methods, the visualization of the generated disparity map will be more smooth and reliable.

From figure 2 (b), we can find in the boundary of the continual building area, the generated raw disparity map will occure some noise and holes, the reason is the texture information in these areas are highly repeated, which makes the correlation layer can not generate reliable cost volume. Which makes the generated disparity are noisy and with some unmatched pixels.

Overall, the building outlines and the small object such as cars and the road can be detected through the raw disparity map.

(a)



(b)

Figure 2. Matching results, (a) input left image,(b) output disparity map

4.4 Comparison experiments

The comparison experiments is conducted on the KITTI 2015 benchmark, we selected 3 supervised deep learning methods for comparison, namely MC-CNN (Zbontar, 2015), Deep Enbed (Zhou, 2015) and Dispnet (Mayer, 2016). The comparison results is shown in table 2.

Table 2. The comparison results

Methods	>2 pixel		>3 pixel	
	NOC	ALL	NOC	ALL
MC-CNN	13.20	15.83	11.35	13.21
Deep Enbed	9.81	11.26	7.29	8.51
Dispnet	9.56	10.74	7.19	8.23
Ours	11.27	13.03	8.35	9.41

From table 2, we can find that our methods can achieve the even accuracy results than supervised learning methods, and better than the MC-CNN method.

5. CONCLUSION

In this paper we proposed a unsupervised matching methods for dense image matching, The experimental results showed that the proposed method achieved the RMSE (root mean square error) of the reprojection error better than 1 pixel in image coordinates, an in-ground positioning accuracy within ±2.5 GSD (Ground Sampling Distance).The comparison experiments on KITTI show the proposed unsupervised learning method even comparably with other supervised methods. In the future, we expect to enhance the networks and the loss functions further. The structural similarity loss in the unsupervised loss is illumination sensitive, could also be improved.

.

REFERENCES

Alex K., Hayk M., Saumitro D., Peter H., Ryan K., Abraham Ba., Adam B., 2017. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, 66–75.

Chang J., Chen Y., 2018. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5410– 5418.

Chapelle O. and Wu M., 2010. Gradient descent optimization of smoothed information retrieval metrics. Inf. Retr., 13(3):216–235.

Chen Z., Sun X., Wang L., Yu Y., and Huang C., 2015. A deep visual correspondence embedding model for stereo matching costs. In ICCV, pages 972–980.

Cheng X., Wang P., Yang R., 2018. Depth estimation via affinity learned with convolutional spatial propagation network. In European Conference on Computer Vision, 108–125.

Geiger A., Lenz P., and Urtasun R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In CVPR, 3354–3361.

He K, Sun J, Tang X, 2013. Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(6):1397-1409.

Han X., Leung T., Jia Y., SukthankaR. r, Berg A. C., 2015. Matchnet: Unifying feature and metric learning for patchbased matching. In CVPR, 3279–3286,

Kong, D, Hai T, 2004. A method for learning matching errors for stereo computation. BMVC. Vol. 1

Li Y., Paluri M., Rehg J. M., Doll´ar P., 2015. Unsupervised learning of edges. CoRR, abs/1511.04166,.

Mayer N., Ilg E., H¨ausser P., Fischer P., Cremers D., 2016. A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In CVPR.

Ronneberger O, Fischer P, Brox T, 2015. U-net: Convolutional networks for biomedical image segmentation.International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 234-241.

Szeliski R, 2010, Computer Vision: Algorithms and Applications. Springer: Berlin, Germany.

Spyropoulos A., Komodakis N., Mordohai P, 2014. Learning to detect ground control points for improving the accuracy of stereo matching. In CVPR, 1621–1628.

Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; and Sebe, N. 2017. Learning cross-modal deep representations for robust pedestrian detection. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

YUAN X., YUAN W., XU S., JI Y., 2019. Research developments and prospects on dense image matching in photogrammetry. Acta Geodaetica et Cartographica Sinica, 48(12): 1542-1550.

Yu J., Harley A. W., Derpanis K. G, 2016. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. CoRR, abs/1608.05842.

Zbontar J. and LeCun Y., 2015. Computing the stereo matching cost with a convolutional neural network. In CVPR, pages 1592–1599.

Zhou C., Zhang H., Shen X., 2017. Unsupervised learning of stereo matching.Proceedings of the IEEE International Conference on Computer Vision. 1567-1575.

Zhang L., Seitz S. M., 2007. Estimating optimal parameters for MRF stereo from a single image pair. IEEE Trans. Pattern Anal. Mach. Intell., 29(2):331–342.