International Journal of Digital Crime and Forensics
Volume 14 • Issue 2

Monocular Depth Matching With Hybrid Sampling and Depth Label Propagation

Ye Hua, Hunan University of Arts and Science, China
Qu Xi Long, Hunan University of Finance and Economics, China
Li Zhen Jin, Hunan University of Finance and Economics, China*

ABSTRACT

This paper proposes a monocular depth label propagation model, which describes monocular images into depth label distribution for the target classification matching: 1) depth label propagation by hybrid sampling and salient region sifting, improve the discrimination of detection feature categories; 2) depth label mapping and spectrum clustering to classify target, define the depth of the sorting rules. The experimental results of motion recognition and 3D point cloud processing show that this method can approximately reach the performance of all previous monocular depth estimation methods. The neural network model black box training learning module is not used, which improves the interpretability of the proposed model.

KEYWORDS

Depth Label Matching, Hybrid Sampling, Interpretability Model, Monocular Depth Estimation

1. INTRODUCTION

The application of depth estimation based on monocular images has the advantages of simplified structure and low operation cost. It is one of the hot spots in academia and industry.

In the existing monocular depth estimation method, based on the relative depth methods (Ming et al.,2016)(Saxena et al.,2009)(Berman et al.,2014)(Melfi et al.,2013)(Zhang et al.,2014), the monocular image data is used to directly predict the depth value corresponding to each pixel in the image, resulting in the existing method usually requiring a large amount of depth annotation data. It usually requires higher acquisition costs. Deep learning-based methods (Xu et al.,2017)(Laina et al.,2016)(Tateno et al.,2017)(Tung et al.,2017)(Zhou et al.,2021) often use network architecture tuning to countervail complex deployment costs and huge computing costs. Based on the method of semantic segmentation information (Li et al.,2018)(Wang et al.,2018)(Liu et al.,2015)(Jiang et al.,2019), implicit geometric constraints are introduced in the training process. Through geometric transformation, view synthesis is used as a supervision signal to reduce the dependence on data. However, such methods still lack explicit geometric association constraints.

ORBSLAM descriptor using feature detection according to the matching feature point motion track difference (Tseng et al.,2020). Feature point extraction and optical flow tracking are not easy to maintain the global map, has a great impact on the illumination, and cannot be used in high-precision maps.

DOI: 10.4018/IJDCF.302879 *Corresponding Author

Thus researcher's motivation is to reduce the dependence on depth annotation data and the geometric structure constraints. The monocular depth estimation researchers proposed has the advantages of convenient model structure deployment and low operation cost. It has rich potential application scenarios in many fields, such as motion recognition, 3D reconstruction, real-time map positioning and navigation for unmanned vehicles, etc.

2. RELATED WORK

Depth estimation and 3D reconstruction of a single image are still very challenging now. Difficulty in identification of the motion of occluded targets can be decomposed into three sub-problems: occlusion target division and sparse depth estimation, overall image depth estimation and motion recognition. The following is the method proposed in this paper to solve the three sub propositions.

1) Hybrid sampling of salient region

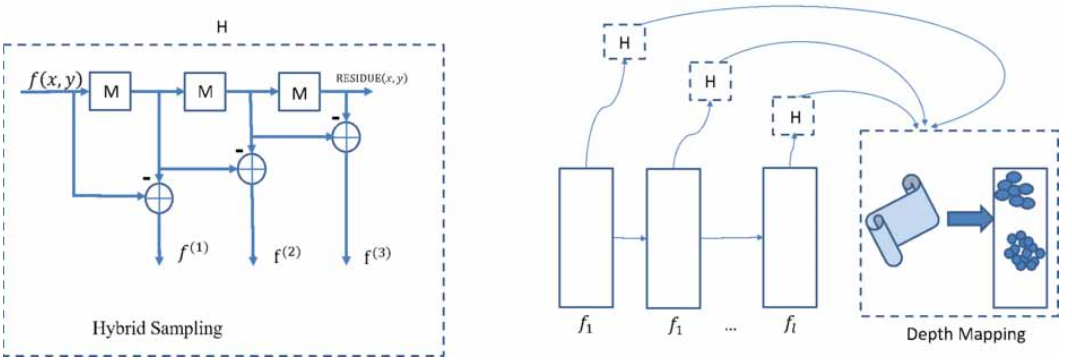
In section 3.1 sparse processing methods (Cai et al.,2011)(Ye et al.,2018)(Jiao et al.,2017)(Liu et al.,2013)(Wang et al.,2015)(Maaten et al.,2008) is applied to detect the salient region, using hybrid sampling for the image information in the scene space structure, calculating the ambiguity of the salient region.

2) Depth label mapping

In section 3.2 the depth of salient regions by section 3.1 is sorted with the DCT high-frequency coefficient distribution (Li et al.,2021)(Hou et al.,2016)(Jiao et al.,2017), and constraint relationship between data feature map manifold embedding and label map manifold embedding of single image is established. And temporal processing is performed on the motion features of the video sequence to obtain the features of depth accumulation map.

The processing flow of depth prediction of monocular image sequence is shown in figure 1. It is divided into two sub-methods, first sampling of salient region by hybrid sampling, then calculation of depth marks under spectral constraints. The $f\left(x,y\right)$ is the source image(Ye et al.,2018), $f^{(1)}\left(x,y\right)$ is the first sifting image by hybrid sampling.

Figure 1. Flowchart of Depth Estimation of monocular image sequence



3. DEPTH MAP MATCHING ON HYBRID SAMPLING

3.1. Depth Label Prediction of Single Image

Depth feature map learning is a relative depth estimation method for images taken by traditional cameras.

Salient region of the image is divided according to the Level Set Method (Wang et al.,2016), and the feature is based on the texture geometric structure. Let $F_{j(x,y)}$ be the j image in the iBEMD subspace image pyramid(Ye et al.,2018), $j=\left(1,\ldots,L*N\right)$, which can be divided into k sets of multiple maximally stable regions R with different scales. The formula is as follows:

$$F_j\left(x\right)=\sum_{i=1}^{N}\sigma_i^k\mathrm{R}\left(x,\sigma_i\right) \tag{1}$$

Then, the image is divided into multiple sets of salient regions with different scales.

Take the image DCT, count the number of coefficients with non-zero altitude value of high-frequency feature coefficients to characterize the significance of the image area, that is, the ordering information of the depth of this region (Li et al.,2021). The label constraint value of the normalized depth of the pixel xi in the image is defined as,

$$Id_i=\frac{\left(\left(\frac{M_h}{M_b}\right)^m\right)_i}{max\left\{\left(\left(\frac{M_h}{M_b}\right)^m\right)_k,k=1,2,\cdots,N\right\}} \tag{2}$$

The depth marks obtained by applying DCT spectral features are discrete and cannot be directly used for the characterization of moving targets.

Let the pixel set of the image be *X*, and the image depth data be , which represents the true depth value of each sample point, and the depth value has *N* levels, and *S* represents the number of depth patches of the image. The feature vector describes the sample point category labels. Its value is obtained according to the formula (2) in Section 3.1, and is the number of category labels. is the depth prediction value, which is obtained by multiplying the sample point category label and the depth label index. Let *Z* be the model normalization parameter, the MRF model of the depth map is as follows,

$$P\left(d_i\#\ x\right)=\frac{1}{Z}\exp\left(-h\left(\frac{d_i-\hat{d}_i}{d_i}\right)-\sum_{j\in N(i)}\mathrm{h}\left(\frac{\hat{d}_i-\hat{d}_j}{max\left(\hat{d}_i,\hat{d}_j\right)}\right)\right)$$

$$\hat{d}_i=\varphi\left(x_i\right)\cdot Id_i,\ h\left(\tau\right)=\tau^2$$

$$\mathrm{Z}=\sum_{i=1}^{S}\exp\left(-h\left(\frac{d_i-\hat{d}_i}{d_i}\right)-\sum_{j\in N(i)}\mathrm{h}\left(\frac{\hat{d}_i-\hat{d}_j}{max\left(\hat{d}_i,\hat{d}_j\right)}\right)\right) \tag{3}$$

After generating the depth prediction value, the probability density function P of MRF model of (3) is maximized, and the maximum posterior estimated probability of the model is obtained, which is the continuous depth estimation value of the regions.

3.2. Depth Accumulation Map with Dual-Spectrum Embedding

Dual-spectrum embedding method (Jiao et al.,2017) is used to establish the constraint relationship between the data feature manifold and the label manifold. Data feature map is established by defining the graph $G^S = \left(\mathbf{V}^S, E^S, W^S\right)$, in which $\mathbf{V}^S$ is the vertex set of data feature map, representing the sample topology structure. E^S is the edge set representing the relationship betresearchersen the data features x_i and x_j, each edge is e_{ij}, each element of the researchersight matrix W_{ij} stands for the researchersight of the edge, representing the similarity of the sample data.

A similar method as data feature map G^S is used to build a category label map of features $G^T = \left(\mathbf{V}^T, E^T, W^T\right)$. The researcher sight matrix of the feature map is defined as,

$$\left[W^s\right]_{ij} = \begin{cases} 1, X_j \in N\left(X_i\right) \\ 0, \quad \text{others} \end{cases} \quad i, j = 1, \ldots, N \tag{4}$$

In which, $N\left(X_i\right)$ is the KNN (k-nearest neighbor) set of data point X_i. The Laplacian matrix of the feature map is as follows: $L_V = D^V - W^V$, D^V is a diagonal matrix, that is,

$$\left[D^V\right]_{ij} = \sum_j \left[W^V\right]_{ij} \tag{5}$$

Let $V = \left[v_1^T, \cdots, v_N^T\right]^T \in \Re^{N \times R}$ be the low-dimensional data representation to be sought, then the local manifold constraint of the data representation is as follows,

$$\begin{aligned} S_V\left(N\left(x_i\right)\right) &= \frac{1}{2}\sum_{i,j=1}^N v_i - v_j{}^2 W^V = \sum_{i=1}^N v_i v_i^T D_{ii}^V - \sum_{i,j=1}^N v_i v_j W_{ij}^V \\ &= \mathrm{Tr}\left(V^T D^T V\right) - \mathrm{Tr}\left(V^T W^V V\right) = \mathrm{Tr}\left(V^T L_V V\right) \end{aligned} \tag{6}$$

S function reflects the smoothness of the clustering partition of the local manifold with k-nearest neighbor. That is, the larger the feature value, the greater the point researchersight will be, the more likely it is to be assigned to a common category label.

The feature label map $G^T = \left(\mathrm{V}^T, E^T, W^T\right)$ is also established by using 0-1 researchersighting method. The vertex set of label map is the feature set $\left\{X_1^T, \cdots, X_M^T\right\}$, and the researchersight matrix of the feature label map is defined as,

$$\left[W^U\right]_{ij} = \begin{cases} 1, X_j^T \in N\left(X_i^T\right) \\ 0, \quad \text{others} \end{cases} \quad i, j = 1, \ldots, M$$

Its Laplacian matrix is $L_U = D^U - W^U$. Let $U = [u_1^T, \cdots, u_M^T]^T \in \Re^{M \times R}$ be the base dictionary to be decomposed[17] and then the local manifold constraint of the base dictionary is as follows,

$$S_U\left(N\left(x_i\right)\right) = \frac{1}{2}\sum_{i,j=1}^{M} u_i - u_j{}^2 W^u = \sum_{i=1}^{N} u_i u_i^T D_{ii}^u - \sum_{i,j=1}^{N} u_i u_j W_{ij}^V$$
$$= \mathrm{Tr}\left(U^T D^U U\right) - \mathrm{Tr}\left(U^T W^U U\right) = \mathrm{Tr}\left(U^T L_U U\right) \tag{8}$$

Based on the above feature maps and label maps, this paper proposes dual regularized spectral clustering (DRPC)(Jiao et al.,2017). The objective function is as follows,

$$J_{DRPC} = X - UDV^{T2}_F + \lambda Tr\left(V^T L_V V\right) + \mu Tr\left(U^T L_U U\right), \quad \text{s.t. } U \geq 0, V \geq 0 \tag{9}$$

The sample category label space T and the feature space Y share a similar local topology, that is, label space T and data space share a similar manifold topology, and the label propagation relationship is established, that is,

$$Q\left(T\right) = \mathrm{F}\left(T\right) + \mathrm{L}\left(T\right) = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\left(\frac{T_i}{\sqrt{D_{ii}}} - \frac{T_j}{\sqrt{D_{jj}}}\right)^2 + \alpha\sum_{i=1}^{n} T_i - Y_i^2$$
$$T = \left(t_i \#\ \ i = 1, 2, \ldots, N\right) \qquad (10)$$

Yi represents the indicator function of the depth value. The second term of the model is data item, which is used to evaluate the depth prediction value of the sample points in the image regions.

$$T^* = \arg\min_T Q\left(T\right) = \left(1-\lambda\right)\left(I - \lambda S\right)^{-1} Y \tag{11}$$

After generating the depth prediction value, only the (11) objective function needs to be maximized during testing, and the continuous depth function value can be obtained(Jiao et al.,2017).

Based on the depth map of the image obtained, the depth accumulation map feature of the video series is established, with time-sequence motion data *y* and the time-sequence dimension *m*, then the k[th] motion is described as, the accumulation depth feature d_{cum} and the accumulation mean depth features d_{med} is defined as in the formula (12),

$$d_{\mathrm{cum}}\left(i,j\right) = \sum_{k=1}^{t} \alpha_k \left| y_k\left(i,j\right) - y_1\left(i,j\right)\right|$$
$$d_{med}\left(i,j\right) = s\left(i,j\right) d_{cum}\left(i,j\right) \tag{12}$$

In the formula (12), k[th] image in the video series, reflecting the importance of the current motion and position, *s* the edge amplitude value, the accumulative mean depth feature d_{med} can reflect researchersak motion features and Get matched depth map.

4. EXPERIMENTS RESULTS AND DISCUSSION

The monocular depth estimation researchers proposed has the advantages of convenient model structure deployment and low operation cost. It has rich potential application scenarios in many fields, such as motion recognition, 3D reconstruction, real-time map positioning and navigation for unmanned vehicles, etc.

4.1. Experiment Parameter Settings on Motion Recognition

In section 3.3 t-SNE method (Eyiyurekli et al.,2017) is used to reduce the dimensionality of the depth features, cluster the formed spectrum of propagation label of video sequence.

Set y as the feature vector of the depth accumulation map with time-sequence motions, and M as the time-sequence dimension of the sequence. Then, researchers can obtain $Y = \left\{y_1\left(i,j\right), y_2\left(i,j\right),\ldots, y_m\left(i,j\right)\right\}$, in which, l^{th} motion vector is described as $y_l\left(i,j\right)$, the vector dimension is K. The transition probability of the signal at time t_i to t_j is proportional to the Gaussian kernel function between their distance d, and further the depth stack map feature is t-SNE embedded. The Gaussian kernel function relationship between the transition probability of the signal at time t_i to t_j and their spatial distance,

$$p_{ij} \propto \exp(-d\left(t_i, t_j\right)^2 / 2\sigma_i^2\right) \tag{13}$$

Spatial distance function D selects the Kullback–Leibler distance to measure the difference in feature vectors. Let P_{ij} be the feature vector at time t_i, and Q_{ij} as the feature vector at time t_j, then,

$$d\left(P,Q\right) = D_{KL}(PQ) = \sum\nolimits_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{14}$$

Perform a frequency-domain transformation on the feature vector, and use S(t_i) to represent the frequency-domain vector corresponding to the feature vector at time t_i. S(t_j) is the frequency-domain vector at time t_j. Then, the distance function is as follows:

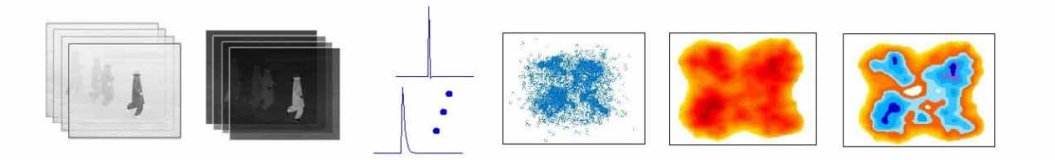
$$D(t_i t_j) = \sum\nolimits_{f,k} \hat{S}\left(k,f;t\right) \log\left[\frac{\hat{S}\left(k,f;t_i\right)}{\hat{S}\left(k,f;t_j\right)}\right] \tag{15}$$

$$\hat{S}\left(k,f;t\right) = \frac{S\left(k,f;t\right)}{\sum k',f' S\left(k',f';t\right)}, \tag{16}$$

In which, f is a frequency domain transformation function of the motion feature vector y, and a certain wavelet function can be selected, $S\left(k,f;t\right)$ is composed of wavelet amplitude, k represents the motion primitives, and $\widehat{S}\left(k,f;t\right)$ normalized wavelet frequency domain features. The clustering distance $D(t_1,t_2)$ of the spectral features is used as an evaluation index for motion classification to complete human motion recognition.

Figure 2 shows the processing effect of monocular images on feature accumulation and depth estimation.

Figure 2. Schematic Diagram of Monocular Depth Estimation for motion recognition



The experimental selection was performed in a typical motion recognition and pose recognition library, including KTH video library(Zhou et al.,2017), Weizmann dataset(Weizmann, 2021), UCF Sports Motion Dataset video library(UCF101, 2021), UCF 101 dataset and HDMB dataset(Wishart et al.,2007). All videos are divided into 10 groups each time, and the training data and test data are randomly divided for each group according to the 10-fold cross-check method. The final experimental result is the average of 10 test results.

In the use experiment, the number of EEMD decomposition layers is set as 4, the number of iterations is set as 10, the features are extracted by using the BEMD-MTS model method, the stability region parameter threshold region direction d is set as 90%, and the flattening ratio q is set as 50%. Let the sample feature vector be denoted as $X = \left(x_1, x_2, \ldots, x_n\right)$, f is the frequency-domain transformation function of the motion feature vector x. The Gabor wavelet function is selected according to Reference[4]. There are 25 wavelet components, $G\left(k, f; t\right)$ are composed of a series of wavelet amplitudes. K represents the number of motion primitives, taking 50, $\hat{G}\left(k, f; t\right)$ is the normalized wavelet frequency domain feature. The perplexity is generally set to 5-50. The objective function defined by minimized distance formula (15), and thereby the classification result of the motion can be obtained.

In the experiment of the KTH dataset, the default settings are 30 perplexities, 1000 iterations, and 1000 learning rate. The best t-SNE embedded operation effect map shall be selected. Figure 3 is an embedded map of a typical t-SNE method without depth feature constraints, and Figure 4 is an embedded map result of the t-SNE method with depth feature constraints proposed in this paper. It is feasible to use t-SNE embedding method to identify motion features. After adding deep feature constraints, the motion category has still higher reparability.

Figure 3. Embedded Diagram of Motion Recognition of a Typical t-SNE Method

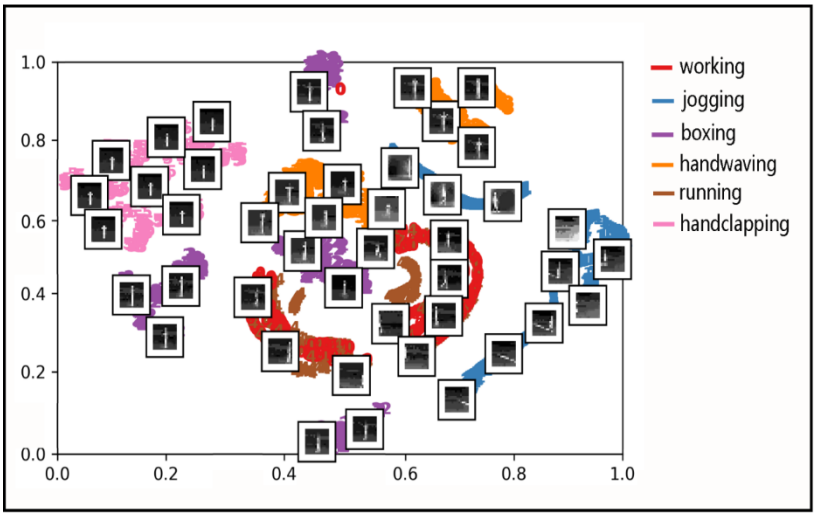
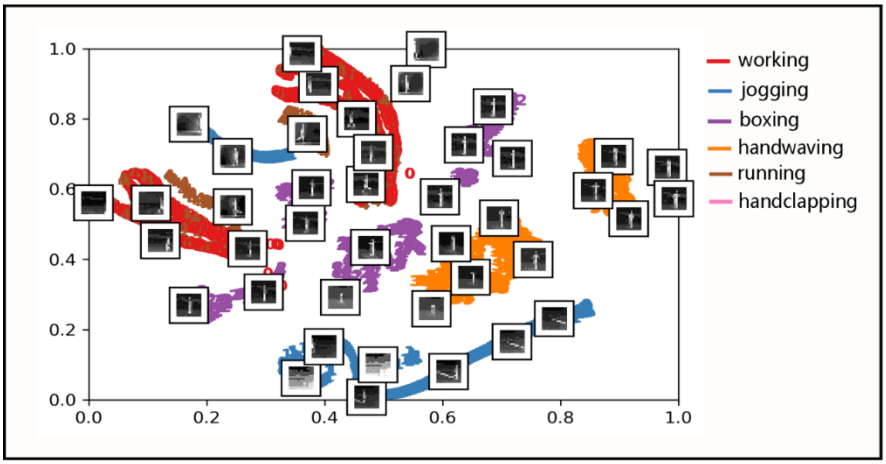


Figure 4. Embedded Diagram of Motion Recognition of t-SNE with Depth Value Constraint



4.2. Results on Motion Recognition Experiment

The results of comparison with the recognition rate of the existing methods are as shown in Tables 1. Compared with the experimental results in Reference (Liu et al.,2011), the depth accumulation map features extracted in this paper are used to accumulate the depth features of each frame from the reverse order of the time sequence.

For the motion recognition effect of simple background, Table 1 is displayed in the KTH video sequence. This processing strategy makes the distinction betresearchersen jog and run motion features increase. The average recognition rate in the KTH dataset reaches 98.5%, and the average recognition rate in the Researchersizmann dataset reaches 98.3%.

The method in this paper has achieved considerable improvement in the recognition performance in the UCF Sports dataset and the HMDB dataset. The proposed method has an average recognition rate of 83.2% in the UCF Sport dataset. The HMDB data set is a large motion recognition data set. As shown in Table 1, the recognition rate of all methods is generally low. Unlike the reference (Liu et al.,2011), the method in this paper does not need to track and detect the moving human body in the video. Using multi-source data to establish 3D features has become a research focus and development trend in the field of motion recognition.

Table 1. Average Recognition Rate in Human Motion Dataset

Method	Recognition rate in KTH(%)	Method	Recognition rate in Researchersizmann (%)	Method	Recognition rate in UCF Sports(%)	Method	Recognition rate in HDMB(%)
HOG/HOF (Liu et al.,2011)	88	Goudelis (Goudelis et al.,2013)	95.4	Feature combination (Liu et al.,2011)	71.2	Ballas (Ballas et al.,2013)	51.8
ISA(Laptev et al.,2008)	91.4	Gorelick (Gorelick et al.,2005)	97.5	ISA(Laptev et al.,2008)	75.8	Wang (Wang et al.,2016)	53.9
Overcomplete ICA(Golestaneh et al.,2017)	93.8	Melfi (Melfi et al.,2013)	99.1	overcomplete ICA(Golestaneh et al.,2017)	82.8	Wang (Wang et al.,2015)	57.2
Proposed Method	98.5	Proposed Method	98.8	Proposed Method	83.2	Proposed Method	55.4

4.3. Evaluate

Use M and N to indicate the height and width of the image, native feature images with the complexity as O(MN).

Complexity of image depth estimation: 1) detect the maximum stable region MSERs, sort the pixels and extract the extreme value regions, with the complexity as O(MN) and O(MNlg (lg (MN))). 2) perform DCT transformation of regional block, with the complexity not exceeding that in the first step. 3) depth sorting, which is obtained by performing comparison calculations less than M times. The algorithm complexity of depth estimation is about 5×O(MN).

Deep label propagation complexity: 1) the complexity of manifold learning is $O(n^2+t(d^2n+mn^2+dmn+dn^2))$(Jiao et al.,2017). Let the size of the decomposition matrix be n * d, and the number of iteration times is t. Neighbor depth assignment is obtained by performing less than M times of comparisons.

Complexity of t-SNE recognition: 1) creating a standard deviation σ_i for high-dimensional point *I,* limiting the perplexity level of each point, and calculating the similarity matrix. 2) KL distance optimization, and the total calculation amount of t-SNE is the temporal and spatial square of the number of data points. This algorithm is on the same level of complexity as $O(n^2)$.

This shows that the model calculation process proposed in this paper can be explained, the calculation cost is low, and there is no memory consumption and calculation consumption for deep network training.

4.4. Experiment Results on Dynamic Point Cloud Processing

In order to verify effectiveness of the depth prediction model in 3D applications (Eigen et al.,2014)(Tschopp et al.,2016)(Tseng et al.,2020)[23](Mahjourian et al.,2018)[41][42](Yin et al.,2018)[44], researchers verified in the following two comparison groups: (1) using the data set monocular data to compare with the existing unsupervised algorithm; (2) using self-built data to evaluate on obtaining 3D structural information and path planning.

(1) Experiment Using KITTI Image Dataset

The KITTI dataset contains 86,000 frames of images (Tseng et al.,2020), it is the current dataset of computer vision algorithm evaluation in automatic driving scenario. According to Nvidia and the MIT (Tschopp et al.,2016)(Tseng et al.,2019,2020)(Tateno et al.,2017)(Mahjourian et al.,2018), used indicators to measure the difference between ground truth and estimated images as shown in Table 2. All the other algorithm experimental results come from literature published by Tseng, Zhang and Zhu(Tseng et al.,2020).

Compared with the existing depth prediction algorithm (Eigen et al.,2014)(Tseng et al.,2020)(Mahjourian et al.,2018)(Yin et al.,2018), the unsupervised depth prediction algorithm can reduce

Table 2. Comparison of improved unsupervised algorithms and mainstream algorithms

	RMSE	Rel	δ_1	δ_2	δ_3
Eigen et al. (Eigen et al.,2014)	7.216	0.228	67.9	89.7	96.7
Mahjourian et al. (Mahjourian et al.,2018)	6.22	0.25	76.2	91.6	96.8
Tseng et al.(Tseng et al.,2020)	4.211	0.124	86.2	94.3	97.4
Yin et al. (Yin et al.,2018)	5.737	0.232	84.6	93.4	97.2
Researchers	4.235	0.125	85.1	93.6	96.3

the root mean square error (RMSE) close to 4.3%, and the depth prediction accuracy is increased as shown. From the experimental result, it can be seen that the unsupervised algorithm greatly improves the prediction accuracy. Compared with the current mainstream supervised algorithm, the hybrid sampling and sparse structure mapping makes the model prediction closer to the true depth label distribution. This verifies that the mapped sparse structure feature is effective to improve the accuracy of the final depth estimation.

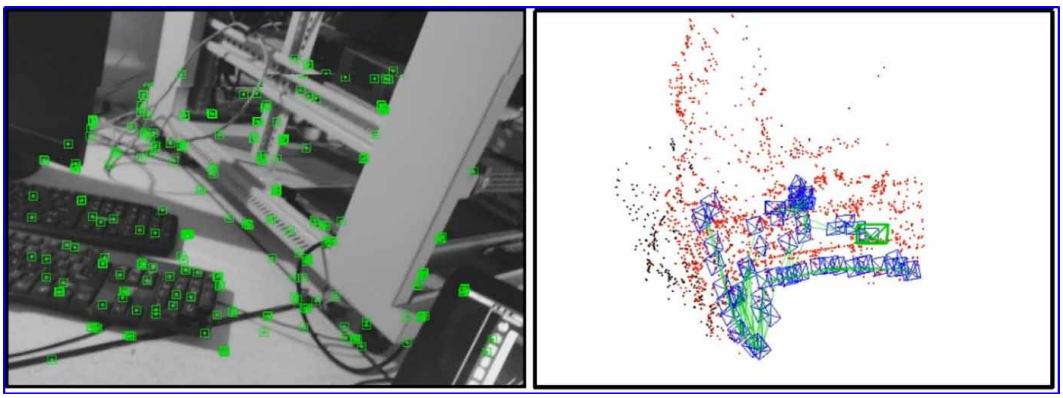
(2) Experiment Using Self-Built Data

To further evaluate the depth prediction performance of the model, we designed a second set of comparative experiments on the unsupervised depth prediction model.This research work is also suitable for supporting unmanned vehicle navigation projects(Golestaneh et al.,2017).

Based on the depth prediction algorithm, researchers obtain 3D structural information for path planning. The depth prediction results and path planning diagram are shown in Figure 5, In Figure 5, a) the feature point distribution of the unmanned vehicle navigation visual image; b) the real-time map feature point distribution and the 3D trajectory map constructed by the algorithm.

Experiments show that the unmanned vehicle module based on proposed method provides real-time location information and feature maps of unknown environments. It has satisfactory real-time performance and low error rate in unmanned vehicle navigation system.

Figure 5. Depth prediction results and path planning diagram



5. CONCLUSION

This paper proposes a simple and effective monocular depth estimation model, namely the data depth label distribution problem and the target classification matching problem. In order to better solve these two sub-problems, the geometric structure of the target is explicitly transformed into a label distribution to improve the model's ability to express the classification of the target. The neural network model black box training learning module is not used, which improves the interpretability of the model. The experimental results show that this method can approximately reach the performance of all previous monocular depth estimation methods. Compared with these mainstream algorithms, they are all complex network structures based on deep learning. Researchers algorithm structure is concise and interpretable. Applying this method to the sampling module and depth mapping module of the deep learning network can promote deep learning 3d applications more robust.

CONFLICT OF INTEREST

Researchers all declare that researchers have no conflict of interest in this paper.

FUNDING

Support provided by Hunan Provincial Key Laboratory for Control Technology of Distributed Electric Propulsion Aircraft; Hunan Provincial Key Laboratory of Finance & Economics Big Data Science and Technology; Key Science and Technology Project of Hunan Provincial Department of Education under Grant Numbers 21A0592, 20A081 and 19A077; Changsha Municipal Natural Science Foundation under Grant Numbers kq2014063.

REFERENCES

Ballas, N., Yang, Y., & Lan, Z. Z. (2013). Space-time robust representation for motion recognition. In *IEEE International Conference on Computer Vision*. IEEE Computer Society.

Berman, G. J., Choi, D. M., Bialek, W., & Shaevitz, J. W. (2014). Mapping the stereotyped behaviours of freely moving fruit flies. *Journal of the Royal Society, Interface*, *11*(99), 20140672. doi:10.1098/rsif.2014.0672 PMID:25142523

Cai, D., He, X., & Han, J. (2011). Graph Regularized Non-Negative Matrix Factorization for Data Representation. *IEEE Transmotions on Pattern Analysis & Machine Intelligence*, *33*(8), 1548–1560. doi:10.1109/TPAMI.2010.231 PMID:21173440

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. Neural Information Processing Systems, 2366–2374.

Eyiyurekli, M., & Breen, D. E. (2017). Detail-Preserving Level Set Surface Editing and Geometric Texture Transfer. *Graphical Models*, *93*, 39–52. doi:10.1016/j.gmod.2017.08.002

Gao, Z., Chen, M. Y., & Hauptmann, A. G. (2010). *Comparing Evaluation Protocols on the KTH Dataset*. Human Motion Understanding. DBLP. doi:10.1007/978-3-642-14715-9_10

Garg, , BG, & Carneiro. (2016). Unsupervised CNN for single view depth estimation: geometry to the rescue. *European Conference on Computer Vision*, 740–756.

Golestaneh, S., & Karam, L. (2017). Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes. *IEEE Conference on Computer Vision and Pattern Recognition*, 5800-5809. doi:10.1109/CVPR.2017.71

Gorelick, L., Blank, M., & Shechtman, E. (2005). Motions as space-time shapes. *IEEE International Conference on Computer Vision*, 1395-1402.

Goudelis, G., Karpouzis, K., & Kollias, S. (2013). Exploring trace transform for robust human motion recognition. *Pattern Recognition*, *46*(12), 3238–3248. doi:10.1016/j.patcog.2013.06.006

Hou, P., Geng, X., & Zhang, M. L. (2016). Multi-Label Manifold Learning. In *Thirtith AAAI Conference on Artificial Intelligence*. AAAI Press.

Jiang, G., Jin, S., Ou, Y., & Zhou, S. (2019). Depth Estimation of a Deformable Object via a Monocular Camera. *Applied Sciences (Basel, Switzerland)*, *9*(7), 1366. doi:10.3390/app9071366

Jiao, Sun, & Hou. (2017). *Sparse learning, classification and recognition*. Science Press.

Laina, Rupprecht, & Belagiannis. (2016). *Deeper Depth Prediction with Fully Convolutional Residual Networks*. Academic Press.

Laptev, I., Marszalek, M., & Schmid, C. (2008). *Learning realistic human motions from movies*. Computer Vision and Pattern Recognition.

Li, B., Ding, Y. C., & He, M. Y. (2018). Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference. *Pattern Recognition*, *83*, 328–339. doi:10.1016/j.patcog.2018.05.029

Li, G., He, Z. S., Tang, J. T., Deng, J. Z., Liu, X. Q., & Zhu, H. J. (2021). Dictionary learning and shift-invariant sparse coding denoising for controlled-source electromagnetic data combined with complementary ensemble empirical mode decomposition. *Geophysics*, *86*(3), 185–198. doi:10.1190/geo2020-0246.1

Liu, F., Shen, C., Lin, G., & Reid, I. (2015). Learning Depth from Single Monocular Images Using Deep Convolutional, Neural Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2024–2039. doi:10.1109/TPAMI.2015.2505283 PMID:26660697

Liu, J., Kuipers, B., & Savarese, S. (2011). *Recognizing human motions by attributes. Computer Vision and Pattern Recognition*. IEEE.

Liu, Lin, & Yan. (2013). Robust Recovery of Subspace Structures by Low-Rank Representation. *IEEE Transmotions on Pattern Analysis & Machine Intelligence, 35*(1), 171-184.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(2605), 2579–2605.

Mahjourian, R., Wicke, M., & Angelova, A. (2018). Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5667–5675. doi:10.1109/CVPR.2018.00594

Melfi, R., Kondra, S., & Petrosino, A. (2013). Human activity modeling by spatio temporal textural appearance. *Pattern Recognition Letters*, *34*(15), 1990–1994. doi:10.1016/j.patrec.2013.04.025

Ming, A., Wu, T., Ma, J., Sun, F., & Zhou, Y. (2016). Monocular Depth Ordering Reasoning with Occlusion Edge Detection and Couple Layers Inference. *IEEE Intelligent Systems*, *31*(2), 54–65. doi:10.1109/MIS.2015.94

Nounou, M. N., Bakshi, B. R., Goel, P. K., & Shen, X. (2002). Bayesian Principal Component Analysis. *Journal of Chemometrics*, *16*(11), 576–595. doi:10.1002/cem.759

Saxena, A., Sun, M., & Ng, A. Y. (2009). *Make3d: Learning 3d scene structure from a single still image*. PAMI.

Tateno, K., Tombari, F., & Laina, I. (2017). CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction. In *IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society. doi:10.1109/CVPR.2017.695

Tschopp, F., Martel, J. N. P., & Turaga, S. C. (2016). Efficient convolutional neural networks for pixelwise classification on heterogeneous hardware systems. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI 2016)*. IEEE. doi:10.1109/ISBI.2016.7493487

Tseng, K. K., Li, J., & Chang, Y. C. (2019). A new architecture for simultaneous localization and mapping: An application of a planetary rover. *Enterprise Information Systems*. Advance online publication. doi:10.1080/17517575.2019.1698772

Tseng, K. K., Zhang, Y., Zhu, Q., Yung, K. L., & Ip, W. H. (2020). Semi-supervised image depth prediction with deep learning and binocular algorithms. *Applied Soft Computing*, *92*, 106272. doi:10.1016/j.asoc.2020.106272

Tung, S. S., & Hwang, W. L. (2017). Multiple depth layers and all-in-focus image generations by blurring and deblurring operations. *Pattern Recognition*, *69*, 184–198. doi:10.1016/j.patcog.2017.03.035

Wang, C., Buenaposada, J. M., & Zhu, R. (2018). Learning Depth from Monocular Videos Using Direct Methods. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/CVPR.2018.00216

Wang, J. H., Xu, Z., Jiang, P. Z., & Luo, L. Y. (2016). M L. Action recognition based on spatio-temporal information and nonnegative component representation. *Journal of Southwest University (Natural Science Edition)*, *46*(4), 675–680.

Wang, L. L., Qiao, Y., & Tang, X. (2015). Motion recognition with trajectory-pooled deep-convolutional descriptors. *IEEE Conference on Computer Vision and Pattern Recognition*, 4305-4314.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. doi:10.1109/TIP.2003.819861 PMID:15376593

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., & Querengesser, L. et al. (2007). HMDB: The Human Metabolome Database. *Nucleic Acids Research*, *35*(Database issue), 521–526. doi:10.1093/nar/gkl923 PMID:17202168

Xu, D., Ricci, E., & Ouyang, W. (2017). Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/CVPR.2017.25

Ye, H., Tan, G. Z., & Hu, C. K. (2018). Curvature filter-empirical mode decomposition on moving human target detection preprocessing. *Hongwai Yu Jiguang Gongcheng*, *47*(2), 259–264.

Yin, Z., & Shi, J. (2018). Geonet: Unsupervised learning of dense depth, optical flow and camera pose. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983–1992. doi:10.1109/CVPR.2018.00212

Zhang, S. P., Yao, H. X., Sun, X., Wang, K., Zhang, J., Lu, X., & Zhang, Y. (2014). Motion recognition based on overcomplete independent components analysis. *Information Sciences*, *281*(10), 635–647. doi:10.1016/j.ins.2013.12.052

Zhou, Brown, & Snavely. (2017). *Unsupervised Learning of Depth and Ego-Motion from Video.* Academic Press.

Zhou, X. M., Hu, Y. G., Liu, W. J., & Sun, R. G. (2021). Research on urban function recognition based on multi-modal and multi-level data fusion method. *Computer Science*, *48*(09), 50–58.

Hua Ye, female, Ph.D. degree in engineering, born in 1977, received a doctorate degree from Central South University in 2018. The main research interests include computer vision perception, machine learning, compressed sensing, and signal processing. She has published many papers, including 3 papers included in SCI.

Xilong Qu, male, Ph.D., professor, born in 1978, since August 2016, serves as the Dean of the School of Information Technology and Management. He was rated as a young backbone teacher in Hunan Province and a "121" talent in Hunan Province. He presided over the completion of 3 projects of the Self-Technology Fund of Hunan Province and won 1 Provincial Science and Technology Progress Award. Published 4 academic works and 35 publicly published papers, of which 5 were included in SCI and 30 were included in EI.