Hindawi
Complexity
Volume 2018, Article ID 5676095, 8 pages
https://doi.org/10.1155/2018/5676095

Research Article

Multimodal Deep Feature Fusion (MMDFF) for RGB-D Tracking

Ming-xin Jiang[1,2] **Chao Deng**[3] **Ming-min Zhang**[4,5] **Jing-song Shan**[6] **and Haiyan Zhang**[6]

[1]*Jiangsu Laboratory of Lake Environment Remote Sensing Technologies, Huaiyin Institute of Technology, Huaian, 223003, China*
[2]*Faculty of Electronic information Engineering, Huaiyin Institute of Technology, Huaian, 223003, China*
[3]*School of Physics & Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, 454000, China*
[4]*School of Computer Science & Technology, Zhejiang University, 310058, China*
[5]*Institute of VR and Intelligent System, Hangzhou Normal University, 310012, China*
[6]*Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, 223003, China*

Correspondence should be addressed to Chao Deng; dengchao_hpu@163.com and Ming-min Zhang; zhangmm95@zju.edu.cn

Received 22 July 2018; Accepted 5 November 2018; Published 28 November 2018

Guest Editor: Li Zhang

Copyright © 2018 Ming-xin Jiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Visual tracking is still a challenging task due to occlusion, appearance changes, complex motion, etc. We propose a novel RGB-D tracker based on multimodal deep feature fusion (MMDFF) in this paper. MMDFF model consists of four deep Convolutional Neural Networks (CNNs): Motion-specific CNN, RGB- specific CNN, Depth-specific CNN, and RGB-Depth correlated CNN. The depth image is encoded into three channels which are sent into depth-specific CNN to extract deep depth features. The optical flow image is calculated for every frame and then is fed to motion-specific CNN to learn deep motion features. Deep RGB, depth, and motion information can be effectively fused at multiple layers via MMDFF model. Finally, multimodal fusion deep features are sent into the C-COT tracker to obtain the tracking result. For evaluation, experiments are conducted on two recent large-scale RGB-D datasets and results demonstrate that our proposed RGB-D tracking method achieves better performance than other state-of-art RGB-D trackers.

1. Introduction

As one of the most fundamental problems in computer vision, visual object tracking, which aims to estimate the trajectory of a object in a video, has been successfully addressed in numerous applications, including intelligent traffic control, artificial intelligence, and autonomous driving [1–3]. Despite the research on visual object tracking has made outstanding achievements, many challenges remain when seeking to track objects effectively in practice. For example, it is still quite difficult to track objects when occlusion occurs frequently, appearance changes, the motion of object is complex, and illumination varies [4–6].

The major drawback of the tracking methods only using RGB data is that they are not robust to appearance changes. Thanks to the availability of consumer-grade RGB-D sensors, such as Inter RealSense, Microsoft Kinect, and Asus Xtion, more accurate depth information of objects can be obtained to revisit the existing problems of tracking [7, 8]. Compared with RGB data, the RGB-D data can remarkably improve the performance of tracking due to the access to the depth information complementary to RGB [9]. The depth information is invariant to illumination or color variations [10, 11] and can provide geometrical cues and spatial structure information; thus it shows powerful benefits in visual object tracking. However, how to effectively utilize the depth data provided by RGB-D sensors is still a challenging issue.

However, RGB and depth only encode static information from a single frame so that tracking often fails when the motion of object is complex. Under these circumstances, deep motion features can provide high-level motion information to distinguish the target object [12–14]. The dynamic information can be captured by deep motion features, and that will be complementary to static features extracted from RGB and depth image.

Our motivation is to design a RGB-D object tracker based on multimodal deep feature fusion. Specific emphasis of this paper is placed on exploring three scientific questions: how to fuse deep motion features with static features provided by RGB and depth image; how to fuse the deep RGB and depth features sufficiently; how to effectively derive geometrical cues and spatial structure information from depth data. In summary, the key technical contributions of this study are three-fold:

(i) A novel MMDFF model is designed for RGB-D tracking, including four deep CNN: motion-specific CNN, RGB- specific CNN, Depth-specific CNN, and RGB-Depth correlated CNN. In MMDFF, we can fuse RGB, depth, and motion information at multiple layers via CNN effectively. The proposed method can separately extract RGB and depth features by using RGB-specific CNN and Depth-specific CNN and adequately exploit the correlated relationship between RGB and depth modality by using RGB-Depth correlated CNN. At the same time, Motion-specific CNN can provide an important high-level information about motion for tracking.

(ii) The depth image is encoded into three channels: horizontal disparity, height above ground, and angle with gravity. Then, the three channel images are sent into depth-specific CNN to extract deep depth features. The strong correlation between RGB and depth modality can be learnt by RGB-Depth correlated CNN. In contrast to only using depth information as one channel in many existing RGB-D trackers, we can obtain more useful information for tracking, such as geometrical features and spatial structure information, by encoding the depth image into three channels.

(iii) To evaluate the performance of our proposed RGB-D tracker, we conduct extensive experiments on the two recent challenging RGB-D benchmark datasets: the large-scale Princeton RGB-D Tracking Benchmark (PTB) Dataset [13] and the University of Birmingham RGB-D Tracking Benchmark (BTB) [15]. The experimental results show that our proposed approach is superior to the state-of-the-art RGB-D trackers.

The remainder of this paper is organized as follows. The related work is discussed in Section 2. Section 3 describes the overview and the details of our proposed method. In Section 4, we demonstrate experimental results to evaluate our proposed RGB-D tracker. We conclude our work in Section 5.

2. Related Work

2.1. RGB-D Object Tracking. With the emergence of RGB-D sensors, there has been great interest in visual object tracking using RGB-D data [15–17] to improve tracking performance since depth modality can provide useful information complementary to RGB modality.

A RGB-D tracking method using depth scaling kernelised correlation filters and occlusion handling was proposed in [18]. In [19], authors used Haar-like features and HOG features based on RGB and depth to form a boosting tracking approach. Zheng et al. [20] presented an object tracker based on sparse representation of depth image. Hannuna et al. proposed a RGB-D tracker built upon the KCF tracker, they exploit depth information to handle scale changes, occlusions, and shape changes.

Although the existing RGB-D trackers have made great contributions to promote RGB-D tracking, most of them used hand-crafted features and fused simply RGB and depth information, ignored the strong correlation between RGB and depth modality.

2.2. Deep RGB Features. Owing to the superiority in feature extraction, CNN has been increasingly used in RGB trackers [21, 22]. The CNN includes a number of convolutional layer, pooling layer, and fully connected layer; the features at different layers have different properties. The higher layers can capture the senmantic features and the lower layers can capture the spatial features, and they are both important in the tracking problem [23, 24].

In [25], Song et al. proposed the CREST algorithm, which treated the correlation filter as one convolutional layer and applied residual learning to capture appearance changes. C-COT was presented in [26], which employed an implicit interpolation model to solve the learning problem in the continuous spatial domain. Zhu et al. [27] proposed a tracker named UCT, which designed an end-to-end framework to learn the convolutional features and perform the tracking process simultaneously.

All these trackers only considered that RGB appearance deep features in current frame cannot benefit from geometrical features extracted from depth image and interframe dynamic information. Thus, it is important to get rid of this problem by fusing deep features from RGB-D data and deep motion features.

2.3. Deep Depth Features. In the recent years, deep depth features have received a lot of attention in object recognition [28, 29], object detection [30, 31], indoor semantic segmentation [32, 33], etc. Unfortunately, few existing RGB-D trackers use CNN to extract deep depth features to improve the tracking performance. In [34], Jiang et al. proposed a RGB-D tracker based on cross-modality deep learning, in which Gaussian-Bernoulli deep Boltzmann Machines (DBMs) were adopted to extract the features of RGB and depth image. As far as the drawbacks of DBMs are concerned, one of the most important ones is the high computational cost of inference.

In recent years, CNN has witnessed great success in computer vision. In this context, we will focus on how to use CNN to fuse deep depth feature with deep RGB and motion features effectively for RGB-D tracker.

2.4. Deep Motion Features. Deep motion feature has been successfully applied for action recognition [35] and video classification [36], but it is rarely applied to visual tracking.

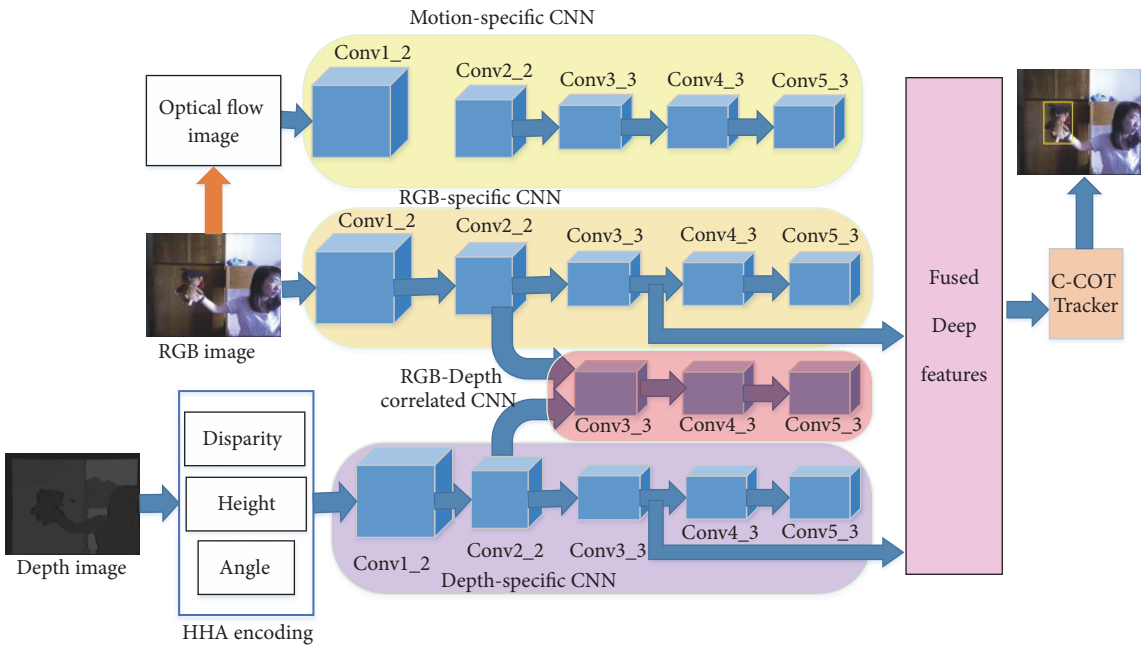


Figure 1: Our MMDFF model for RGB-D Tracking.

Most existing tracking methods only extract appearance features, ignore motion information. In [37], Zhu et al. proposed an end-to-end flow correlation tracker, which focuses on making use of the rich flow information in consecutive frames to improve the feature representation and the tracking accuracy. Danelljan [38] investigated the influence of deep motion features in a RGB tracker. But they have not taken depth information into account.

To the best of our knowledge, deep motion features are yet to be applied for RGB-D tracking. In this paper, we will discuss how to fuse deep motion features with RGB appearance features and geometrical features provided by depth image to improve the performance of RGB-D tracking.

3. RGB-D Tracking Based on MMDFF

3.1. Multimodal Deep Feature Fusion (MMDFF) Model. In this section, a novel MMDFF model is proposed for RGB-D tracking aiming at fusing deep motion features with static appearance and geometrical features provided by RGB and depth data. The overall architecture of our approach is illustrated as Figure 1, and our end-to-end MMDFF model is composed of four deep CNN: motion-specific CNN, RGB-specific CNN, depth-specific CNN, and RGB-Depth correlated CNN. A final fully connected (FC) fusion layer is proposed to effectively fuse deep RGB-D features and deep motion features; then the fused deep features are sent into the C-COT tracker, and the tracking result can be obtained at last.

As described in Section 1, CNN has been shown to significantly outperform traditional machine learning approaches in a wide range of computer vision tasks. It has been widely acknowledged that CNN features extracted from different layers play different important roles in the tracking. A lower layer captures spatial detailed features, which is helpful to precisely localize the target object; at the same time, a higher layer provides semantic features which is robust to occlusion and deformation. In our MMDFF model, we separately adopt hierarchical convolutional features extracted from RGB-specific CNN and depth-specific CNN. To be more specific, two independent CNN networks are adopted: the RGB-specific CNN is for RGB data and the depth-specific CNN is for depth features. And the features on Conv3-3 and Conv5-3 in the two CNNs are sent to the highest fusing FC layer in our experiments.

It is believed that more features extracted from different modalities can be helpful to accurately descript the objects and improve the tracking performance. As abovementioned, most of existing RGB-D trackers directly concatenate features extracted from RGB and depth modalities, not adequately exploiting the correlation between the two modalities. In our method, the strong correlation between RGB and depth modality can be learnt by RGB-Depth correlated CNN.

For human visual system, geometrical and spatial structure information plays an important role in tracking objects. In order to more explicitly derive geometrical and spatial structure information from depth data, we encode it into three channels: horizontal disparity, height above ground, and angle with gravity, using the encoding approach proposed in [39]. Then, the three channel image is sent into depth-specific CNN to extract deep depth features.

The optical flow image is calculated for every frame and then is fed to motion-specific CNN to learn deep motion features, which can capture high-level information about the movement of the object. A pretrained optical flow network provided by [13] is used as our motion-specific CNN,

Table 1: Comparison results: SR using different deep features fusions on the PTB dataset.

Deep Features	All SR	Target type			Target size		Movement		Occlusion		Motion type	
		human	animal	rigid	large	small	slow	fast	yes	no	passive	active
Only RGB	0.73	0.75	0.69	0.74	0.80	0.72	0.71	0.72	0.74	0.85	0.78	0.73
RGB+Depth	0.77	0.79	0.71	0.77	0.81	0.77	0.76	0.73	0.81	0.86	0.79	0.75
RGB+Depth + RGB-depth correlated	0.79	0.81	0.74	0.79	0.83	0.79	0.78	0.74	0.83	0.86	0.81	0.77
RGB+Depth + RGB-depth correlated +Motion	0.84	0.83	0.86	0.85	0.85	0.86	0.82	0.83	0.87	0.87	0.82	0.83

which is pretrained on the UCF101 dataset and includes five convolutional layers.

Thus far, we have obtained multimodal deep features to represent the rich information of the object, including the RGB, horizontal disparity, height above ground, angle with gravity, and motion. Next, we attempt to explore how to fuse the multimodal deep features using the CNN. To solve this problem, we conduct extensive experiments to evaluate the performance using different fusion schemes, each experiment fuses the multimodal deep features at different layer. Inspired by the working mechanism of human visual cortex in the brain which indicates the features should be fused in the high level, so we test fusing at several relatively high layers, such as pool 5, fc 6 and fc 7. We find that fusing the multimodal deep features from fc 6 and fc 7 can obtain better performance.

Let p_i^j represent feature map of j modalities and i denotes the spatial position, $j = \{1, 2, \ldots, J\}$. In our paper, $J = 4$ as we adopt the feature maps from Conv3-3 and Conv5-3 in RGB, depth, RGB-depth correlated, and motion modality. The fusing feature map P_i^{fusion} is the weighted sum of feature maps for three levels in four modalities,

$$P_i^{fusion} = \sum_{j=1}^{J} w_i^j p_i^j \tag{1}$$

where the weigh w_i^j can be computed as follows:

$$w_i^j = \frac{\exp\left(\mathrm{p}_i^j\right)}{\sum_{k=1}^{J} \exp\left(\mathrm{p}_i^k\right)} \tag{2}$$

3.2. C-COT Tracker. The multimodal fusion deep features are sent into the C-COT tracker, which was proposed in [26]. A brief review of the C-COT tracker will be provided in the following of this section, and we will use the same symbols as in [33], for convenience, and more detailed description and proofs can be found in [26].

The C-COT transfers multimodal the fusion feature map to the continuous spatial domain $t \in [0, \mathrm{T})$ by defining the interpolation operator $J_d : \mathbb{R}^{N_d} \longrightarrow L^2(\mathrm{T})$ as follows:

$$J_d\left\{x^d\right\}(t) = \sum_{n=0}^{N_d-1} x^d[n]\, b_d\left(t - \frac{T}{N_d} n\right) \tag{3}$$

The convolution operator is defined as

$$S_f\{x\} = f * J\{x\} - \sum_{d=1}^{D} f^d * J_d\left\{x^d\right\} \tag{4}$$

The objective function is

$$E(f) = \sum_{j=1}^{M} \alpha_j \left\| S_f\left\{x_j\right\} - y_j \right\|_{L^2}^2 - \sum_{d=1}^{D} \left\| \omega f^d \right\|_{L^2}^2 \tag{5}$$

Equation (5) can be minimized in the Fourier domain to learn the filter. The following can be obtained by applying Parseval's formula to (5)

$$E(f) = \sum_{j=1}^{M} \alpha_j \left\| \sum_{d=1}^{D} \widehat{f}^d X_j^d \widehat{b}_d - \widehat{y}_j \right\|_{L^2}^2 - \sum_{d=1}^{D} \left\| \widehat{\omega} * \widehat{f}^d \right\|_{L^2}^2 \tag{6}$$

The desired convolution output $\widehat{y}_j$ can be provide by the following expression:

$$\widehat{y}_j[k] = \frac{\sqrt{2\pi\sigma^2}}{T} \exp\left(-2\sigma^2 \left(\frac{\pi k}{T}\right)^2 - i\frac{2\pi}{T} u_j k\right) \tag{7}$$

4. Experimental Results

The experiments are conducted on two challenging benchmark datasets: BTB dataset [17] with 36 videos and PTB dataset [7] with 100 videos to test the proposed RGB-D tracker using MATLAB R2016b platform with the Caffe toolbox [40] on a PC with an Intel(R) Core(TM) i7-4712MQ CPU@3.40GHz (with 16G memory) and TITAN GPU (12.00 GB memory).

4.1. Deep Features Comparison. The impacts of deep motion features are evaluated by conducting different experiments on PTB dataset. Table 1 indicates the comparison results of different deep features fusions using success rate (SR) as measurements. From Table 1, we find that SRs are lowest when only using deep RGB feautures. SRs increase by fusing deep depth features and RGB-depth correlated features, especially when occlusion occurs. Further, the performance is obviously improved when deep motion features are added, especially when the movement is fast, motion type is active, and target size is small.

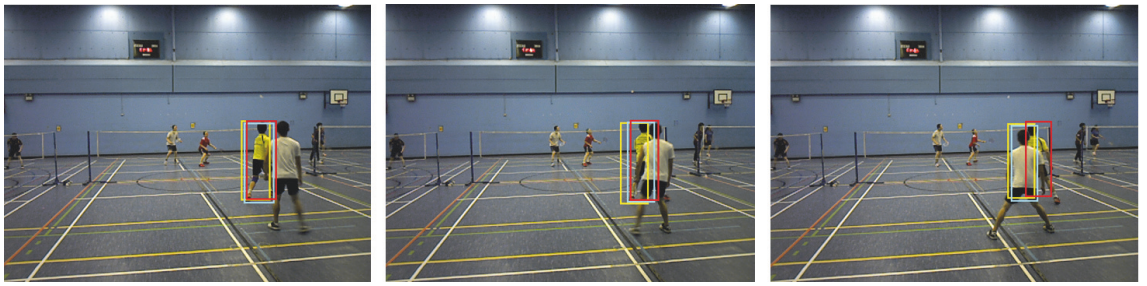


Figure 2: The comparison results on athlete_move video from BTB dataset.

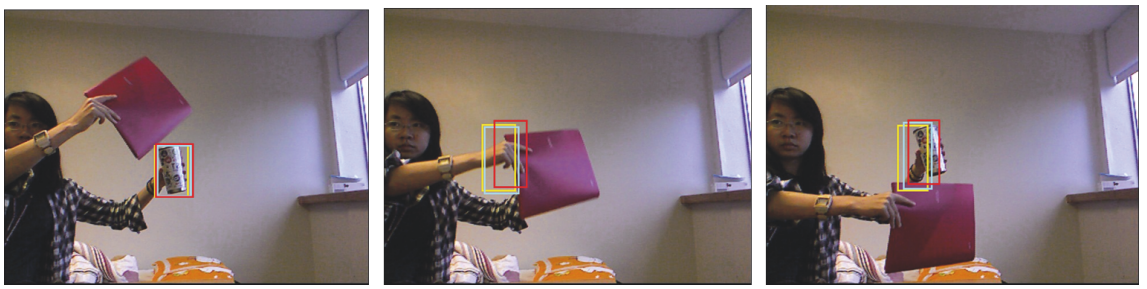


Figure 3: The comparison results on cup_book video from PTB dataset.

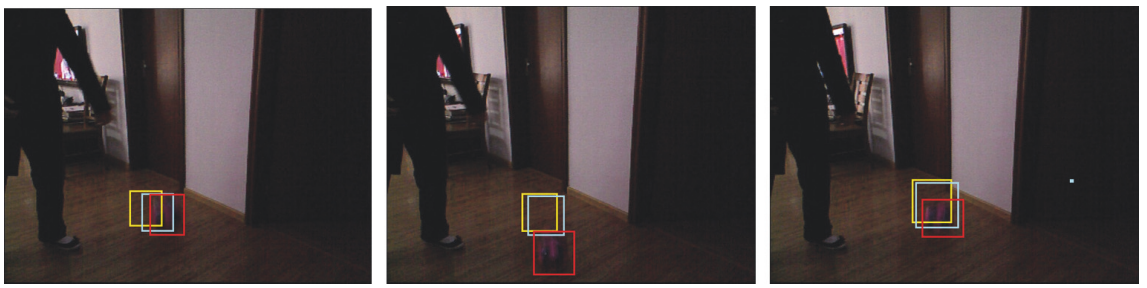


Figure 4: The comparison results on zballpat_no1 video from PTB dataset.

To intuitively show the contribution of fusing deep depth features and deep motion features for RGB-D tracking, in the following, we demonstrate part of experimental results on the BTB dataset and PTB dataset that only using deep RGB features obtains unsatisfactory performance. The tracking result only using deep RGB features is in yellow, fusing deep RGB and depth is in blue, adding deep motion features to RGB, and depth is in red.

In Figure 2, the major challenge of the athlete_move video is that the target object moves fast from left to right. As illustrated in Figure 3, the cup is fully occluded by the book. The zballpat_no1 sequence is challenging due to the change of moving direction (Figure 4). The deep motion features are able to improve the tracking performance as they can exploit the motion patterns of the target object.

4.2. Quantitative Evaluation. We present the results of comparing our proposed tracker with four state-of-the-art RGB-D trackers on the BTB dataset and PTB dataset: Prin Tracker [7] (2013), DS-KCF* Tracker [41] (2016), GBM Tracker [34] (2017), and Berming Tracker [17] (2018). We provide the results in terms of success rate (SR) and area-under-curve (AUC) as measurements.

Figure 5 shows the comparison results of SR of different trackers on the PTB dataset. The results illustrates that the overall SR of our tracker is 87%, SR is 86% when the object moves fast, and SR is 84% when the motion type is active. These values are higher than other trackers obviously, especially when the object moves fast.

Figure 6 indicates the AUC comparison results on the BTB dataset. The overall AUC of our tracker is 9.30, the AUC is 9.84 when the camera is stationary, and the AUC is 8.27 when the camera is moving.

From Figures 5-6, we can see that our tracker obtains the best performance, especially when the object is moving fast or the camera is moving. These results show that deep motion feature is helpful to improve the tracking performance.

5. Conclusion

We study the problems in the existing visual object tracking algorithms and find that existing trackers cannot benefit

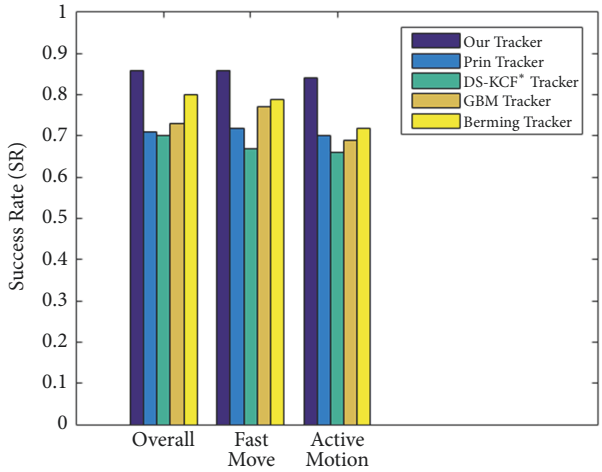


Figure 5: The comparison results of SR on the PTB dataset.

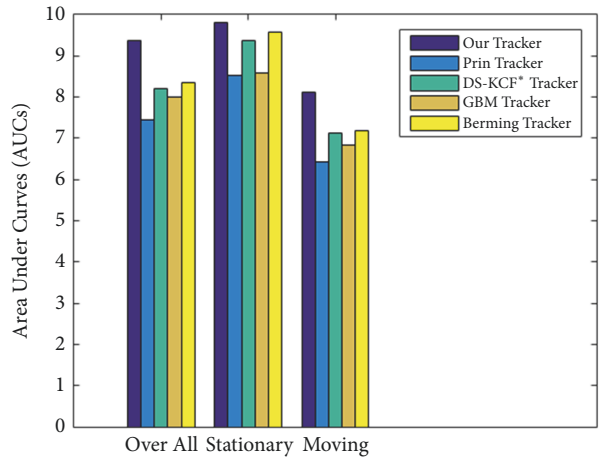


Figure 6: The comparison results of AUC on the BTB dataset.

from geometrical features extracted from depth image and interframe dynamic information by fusing deep RGB, depth, and motion information. We propose MMDFF model to solve these problems. In this model, RGB, depth and motion information are fused at multiple layers via CNN, the correlated relationship between RGB and depth modality exploited by using RGB-Depth correlated CNN. The experimental results show that deep depth feature and deep motion feature provide complementary information to RGB data and the fusing strategy promotes the tracking performance significantly, especially when occlusion occurs, the movement is fast, motion type is active, and target size is small.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D project under Grant no. 2017YFB1002803, Major Program of Natural Science Foundation of the Higher Education Institutions of Jiangsu Province under Grant no. 18KJA520002, Project funded by the Jiangsu Laboratory of Lake Environment Remote Sensing Technologies under Grant no. JSLERS-2018-005, Six-Talent Peak Project in Jiangsu Province under Grant no. 2016XYDXXJS-012, the Natural Science Foundation of Jiangsu Province under Grant no. BK20171267, 533 Talents Engineering Project in Huaian under Grant no. HAA201738, project funded by Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-aged Teachers and Presidents, and the fifth issue 333 high-level talent training project of Jiangsu province.

References

[1] S. Duffner and C. Garcia, "Using Discriminative Motion Context for Online Visual Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2215–2225, 2016.

[2] X. Zhang, G.-S. Xia, Q. Lu, W. Shen, and L. Zhang, "Visual object tracking by correlation filters and online learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017.

[3] E. Di Marco, S. P. Gray, and K. Jandeleit-Dahm, "Diabetes alters activation and repression of pro- and anti-inflammatory signaling pathways in the vasculature," *Frontiers in Endocrinology*, vol. 4, p. 68, 2013.

[4] M. Jiang, Z. Tang, and L. Chen, "Tracking multiple targets based on min-cost network flows with detection in RGB-D data," *International Journal of Computational Sciences and Engineering*, vol. 15, no. 3-4, pp. 330–339, 2017.

[5] K. Chen, W. Tao, and S. Han, "Visual object tracking via enhanced structural correlation filter," *Information Sciences*, vol. 394-395, pp. 232–245, 2017.

[6] M. Jiang, D. Wang, and T. Qiu, "Multi-person detecting and tracking based on RGB-D sensor for a robot vision system," *International Journal of Embedded Systems*, vol. 9, no. 1, pp. 54–60, 2017.

[7] S. Song and J. Xiao, "Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines," in *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 233–240, Sydney, Australia, December 2013.

[8] N. An, X.-G. Zhao, and Z.-G. Hou, "Online RGB-D tracking via detection-learning-segmentation," in *Proceedings of the 23rd International Conference on Pattern Recognition, ICPR 2016*, pp. 1231–1236, Mexico, December 2016.

[9] D. Michel, A. Qammaz, and A. A. Argyros, "Markerless 3D human pose estimation and tracking based on RGBD cameras: An experimental evaluation," in *Proceedings of the 10th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2017*, pp. 115–122, Greece, June 2017.

[10] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Modality and component aware feature fusion for RGB-D scene classification," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 5995–6004, USA, July 2016.

[11] S. Hou, Z. Wang, and F. Wu, "Object detection via deeply exploiting depth information," *Neurocomputing*, vol. 286, pp. 58–66, 2018.

[12] A. Dosovitskiy, P. Fischery, E. Ilg et al., "FlowNet: Learning optical flow with convolutional networks," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 2758–2766, Chile, December 2015.

[13] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 759–768, USA, June 2015.

[14] S. Karen and Z. Andrew, "Two-Stream Convolutional Networks for Action Recognition in Videos [C]," *Andrew Z. Two-Stream Convolutional Networks for Action Recognition in Videos [C]. NIPS*, 2014.

[15] L. Spinello, M. Luber, and K. O. Arras, "Tracking people in 3D using a bottom-up top-down detector," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1304–1310, Shanghai, China, May 2011.

[16] K. Meshgi, S.-I. Maeda, S. Oba, H. Skibbe, Y.-Z. Li, and S. Ishii, "An occlusion-aware particle filter tracker to handle complex and persistent occlusions," *Computer Vision and Image Understanding*, vol. 150, pp. 81–94, 2016.

[17] J. Xiao, R. Stolkin, Y. Gao, and A. Leonardis, "Robust Fusion of Color and Depth Data for RGB-D Target Tracking Using Adaptive Range-Invariant Depth Models and Spatio-Temporal Consistency Constraints," *IEEE Transactions on Cybernetics*, 2017.

[18] M. Camplani, S. Hannuna, M. Mirmehdi et al., "Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling," in *Proceedings of the British Machine Vision Conference 2015*, pp. 145.1-145.11, Swansea.

[19] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with on-line boosted target models," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems: Celebrating 50 Years of Robotics, IROS'11*, pp. 3844–3849, USA, September 2011.

[20] W.-L. Zheng, S.-C. Shen, and B.-L. Lu, "Online Depth Image-Based Object Tracking with Sparse Representation and Object Detection," *Neural Processing Letters*, vol. 45, no. 3, pp. 745–758, 2017.

[21] H. Li, Y. Li, and F. Porikli, "DeepTrack: learning discriminative feature representations online for robust visual tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1834–1848, 2016.

[22] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9914, pp. 850–865, 2016.

[23] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 3119–3127, Chile, December 2015.

[24] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 3074–3082, Chile, December 2015.

[25] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional Residual Learning for Visual Tracking," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 2574–2583, Italy, October 2017.

[26] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking," in *Computer Vision – ECCV 2016*, vol. 9909 of *Lecture Notes in Computer Science*, pp. 472–488, Springer International Publishing, Cham, 2016.

[27] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang, "UCT: Learning Unified Convolutional Networks for Real-Time Visual Tracking," in *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 1973–1982, Venice, October 2017.

[28] Y. Cheng, X. Zhao, K. Huang, and T. Tan, "Semi-supervised learning and feature evaluation for RGB-D object recognition," *Computer Vision and Image Understanding*, vol. 139, pp. 149–160, 2015.

[29] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015*, pp. 681–687, Germany, October 2015.

[30] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, "Cross-modal adaptation for RGB-D detection," in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation, ICRA 2016*, pp. 5032–5039, Sweden, May 2016.

[31] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for RGB-D object detection," *Pattern Recognition*, vol. 72, pp. 300–313, 2017.

[32] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 1475–1483, USA, July 2017.

[33] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 4990–4999, Italy, October 2017.

[34] M. Jiang, Z. Pan, and Z. Tang, "Visual object tracking based on cross-modality Gaussian-Bernoulli deep Boltzmann machines with RGB-D sensors," *Sensors*, vol. 17, no. 1, 2017.

[35] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 3218–3226, Chile, December 2015.

[36] A. Chadha, A. Abbas, and Y. Andreopoulos, "Video Classification With CNNs: Using The Codec As A Spatio-Temporal Activity Sensor," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

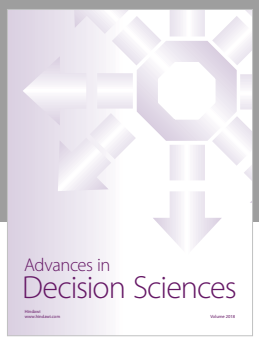
[37] Z. Zhu, W. Wu, and W. Zou, "End-to-end Flow Correlation Tracking with Spatial-temporal Attention[J]," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

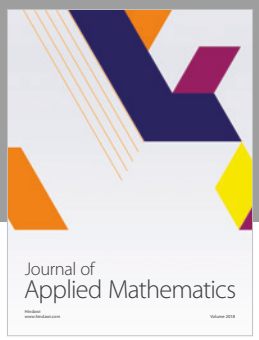
[38] M. Danelljan, G. Bhat, S. Gladh, F. S. Khan, and M. Felsberg, "Deep motion and appearance cues for visual tracking," *Pattern Recognition Letters*, 2018.

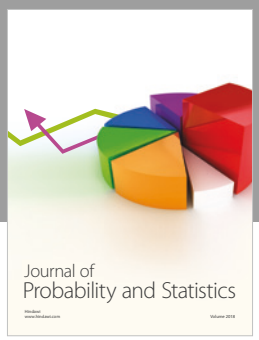
[39] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8695, no. 7, pp. 345–360, 2014.

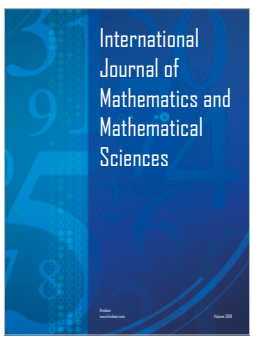
[40] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM Conference on Multimedia (MM '14)*, pp. 675–678, ACM, Orlando, Fla, USA, November 2014.

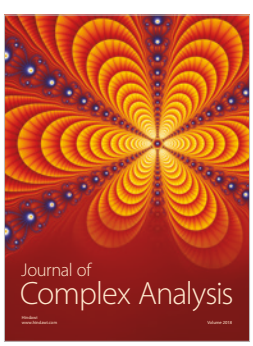
[41] S. Hannuna, M. Camplani, J. Hall et al., "DS-KCF: a real-time tracker for RGB-D data," *Journal of Real-Time Image Processing*, pp. 1–20, 2016.

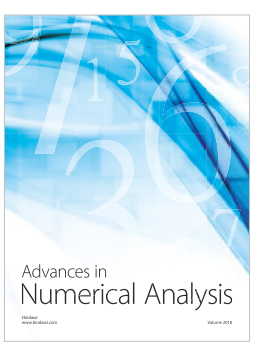


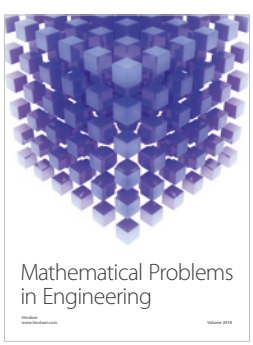


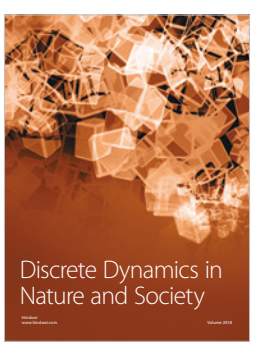


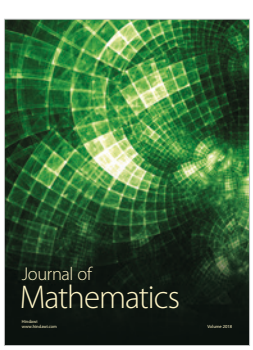


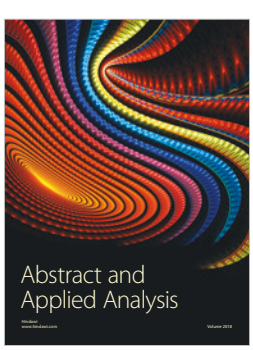


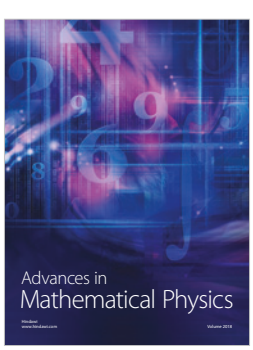












Hindawi

Submit your manuscripts at
www.hindawi.com