Depth information aided constrained correlation filter for visual tracking

Depth information aided constrained correlation filter for visual tracking

Guanqun Li[1], Lei Huang[1], Peichang Zhang[1,3], Qiang Li[1] and YongKai Huo[2]

[1]Collage of Information Engineering, Shenzhen University, NanHai Ave.3688, NanShan Dist, Shenzhen, China.
[2]Collage of Computer Science and Software Engineering, Shenzhen University, NanHai Ave.3688, NanShan Dist, Shenzhen, China.
Email: pzhang@szu.edu.com

Abstract. In this paper, we proposed a novel visual tracking system by constructing the Constrained Correlation Filter (CCF) with Depth Information. More specifically, in our proposed system, to avoid the boundary effects and the fixed shape assumption of conventional Discriminative Correlation Filter (DCF), the shape of target is extracted from depth image provided by RGB-D sensor to construct the CCF, which may prevent the filter from being disturbed by the background noise at the learning stage and enlarge the search region. Moreover, in order to avoid the drifting problem, the update of the model is stopped once part of the target is occluded. The feature weighting coefficients, which reflect the discriminability of the feature channels, are used at the location stage to improve the discriminability. The experimental results show that our method is capable of achieving state-of-the-art performance on Princeton RGB-D tracking benchmark among all public tracking algorithms.

1 Introduction

Visual tracking has been a popular and challenging topic in recent years, while the challenges come from a number of factors, such as illumination and scale changes, rotation in and out of plane, movement by target or camera, etc. Quite a number of visual tracking methods[1-8] based on DCF[9] have been proposed and have shown impressive performance in all standard benchmarks since Bolme[10] introduced correlation filter into visual tracker. More discriminative features such as Histogram of Oriented Gradient (HOG)[11], Colour Names (CN)[12], Deep Convolutional Neural Networks (CNNs)[4,6,13,14], and tricks such as kernel[1-2], ridge regression[2], spatial and temporal regularization are adopted to improve the performance and robustness of visual tracking[3,5]. However, the visual tracking problem is still far from being solved.

The above mentioned DCF-based trackers benefit from the periodic assumption of training samples, which can be learned efficiently in the frequency domain via Fast Fourier Transform (FFT). However, the DCF-based tracker also introduces unwanted boundary effects, which may severely degrade the quality of the tracking model. This is because that, firstly, inaccurate negative training patches would reduce the discriminability of the learned model. Secondly, the detection scores are only accurate around the center of the search region, while the remaining scores are heavily influenced by the periodic repetitions of the detection samples. This leads to a restricted target search region at the location stage.

For the sake of avoiding the boundary effects, Kiani Galoogahi proposed Limited Boundaries Correlation Filter (LBCF)[15] which addressed the problem that occurs due to learning with circular correlation from small training regions. They proposed a learning framework that artificially increases

（a） (b) (c) (d)

Figure 1.The result of our proposed method

the filter size by implicitly zero padding the filter, which reduces the boundary effects artifacts by increasing the number of training examples in constrained filter learning. Danelljan proposed Spatially Regularized DCF (SRDCF)[3] which introduced spatial regularization by reformulating the loss function to penalize nonzero filter values outside the object boundary. Though SRDCF outperforms LBCF[15], the learned filter of SRDCF is still limited by the rectangular shape assumption. Lukezic proposed Channel and Spatial Reliability DCF (CSR-DCF)[16] which introduced the spatial reliability map to eliminate the limitations of periodic assumption and the rectangular shape assumption, and they also introduced the channel weights based on the discriminability to further avoid the different scales issue of each feature channels's contribution to the final response. The weakness of CSR-DCF is that the quality of the estimated spatial reliability map on colour image using spatial 2D priors and color segmentation is relatively low. The map can be segmented exactly with spatial information.

The RGB-D sensor can provide spatial information without extra computational cost in 3D tracking, which gains robustness in scenarios such as illumination changes and scale changes, and provides strong cues for shape changes and occlusion.

In 3D tracking, one of the first methods was proposed by Song[17], which combined an SVM detector and an optical flow tracker, who released the Princeton benchmark that includes 100 videos with 11 categories for both RGB and RGB-D visual tracking. In the tracker, HOG feature was extracted on both depth and colour image. The speed of the method is only 0.26 fps on average, due to exhaustive search and optical flow. However, in terms of precision it outperformed state-of-the-art RGB-only trackers, which demonstrates the importance of depth information in visual tracking.

Based on Princeton benchmark, quite a number of tracking methods[18-21] have been proposed. Bible[19] presented a part-based sparse tracker in particle filter framework. The target location was

firstly estimate by optical flow, then particles are sampled in rotation and translation space. In addition, they proposed an automated method to abate the noise of synchronization and registration between the color and depth streams, which achieved the state-of-the-art performance on Princeton benchmark. However, this method based on PCL library which greatly limited it's practical applications may lead to high computational complexity.

Depth Scale-Kernelized Correlation Filter (DS-KCF)[18] proposed by Hannuna integrated both depth and colour features in the Kernelized Correlation Filter (KCF)[2] framework, which included target object segmentation to obtain the region of target. The depth distribution of target extracted from the the region of target on depth image is used for scale changes, occlusions and aspect ratio changes of the tracking model. But the method based on KCF is limited by boundary effects and fixed shape assumption.

Kart proposed Depth Mask-DCF (DM-DCF)[20] which introduced the 2D spatial reliability map[16] into 3D tracking, where they get the spatial reliability map on depth image by producing a foreground probability map. In our work, a more accurate and robust segmentation method is adopted to get the reliability map. At the same time, an occlusion handling mechanism is introduced to improve robustness. The result of our proposed method is shown as figure 1 in the case of scale variance (a)(c)(d), deformations (a)(b)(c)(d) and out-of-plan rotations (c). Yellow rectangle represents the search window when occlusions occur and the red rectangle represents the target. Against this background, our novel contributions are:

1. We get more accurate shape of target using clustering on depth histogram than the methods using foreground probability map or colour segmentation.
2. We construct the CCF with the shape of target to overcome the boundary effects and prevent the filter from being disturbed by the background.
3. We closely integrate depth and the colour image, and gain the best performance on the Princeton dataset among all public tracking algorithms.

The remainder of the paper is organized as follows: Section 2 introduces the DCF framework. Our proposed scheme is described in Section 3, while experiments and conclusion are presented in Section 4 and Section 5 respectively.

2. DCF Framework

The aim of standard DCF formulation is to learn a multi-channel convolution filter h by minimizing the L^2-error between the response $g(h)$ on the training feature f and the desired output labels g. DCF is formulated as a ridge regression problem as

$$\xi(h) = \left\| g(h) - g \right\|^2 + \lambda \left\| h \right\|^2 , \tag{1}$$

where, λ is the weight of the regularization term. The label g is a Gaussian function, which decays smoothly from one at the target center to zero for other shifts and we have

$$g(h) = \sum_{d=1}^{N_c} f_d * h_d \quad , \tag{2}$$

where * represents circular correlation between $f_d \in \mathbb{R}^{M \times N}$ and $h_d \in \mathbb{R}^{M \times N}$, M and N are the width and height of the target region, respectively. N_C is the number of the feature channels of the image region to be detected or to be trained. Let us further convert the loss function equation (1) to the Fourier domain as

$$\xi(h) = \left\| \sum_{d=1}^{N_c} diag(f_d) \overline{h_d} - g \right\|^2 + \lambda \sum_{d=1}^{N_c} \left\| h \right\|^2 . \tag{3}$$

Here, $h_d \in \mathbb{R}^{D \times 1}$ denotes the column vector of the Discrete Fourier Transformation (DFT) of h_d, with $D = M \cdot N$. $diag(f_d)$ is a $D \times D$ diagonal matrix formed from f_d and $\overline{(\cdot)}$ is the complex conjugate operator. The closed form solution of equation (3) is

$$h_d = diag(f_d \overline{g_d}) \left(\sum_{d=1}^{Nc} diag(f_d \overline{f_d}) + \lambda \right)^{-1}, \quad (4)$$

where, the fraction indicates element-wise division. The efficiency of the DCF is apparent in equation (4) as it reduces the computational cost of the general ridge regression problem from $O(N_c \cdot D^2)$ to $O(N_c \cdot D \cdot \log(D))$ of the element-wise division operations and DFT/IDFT included in equation (4).

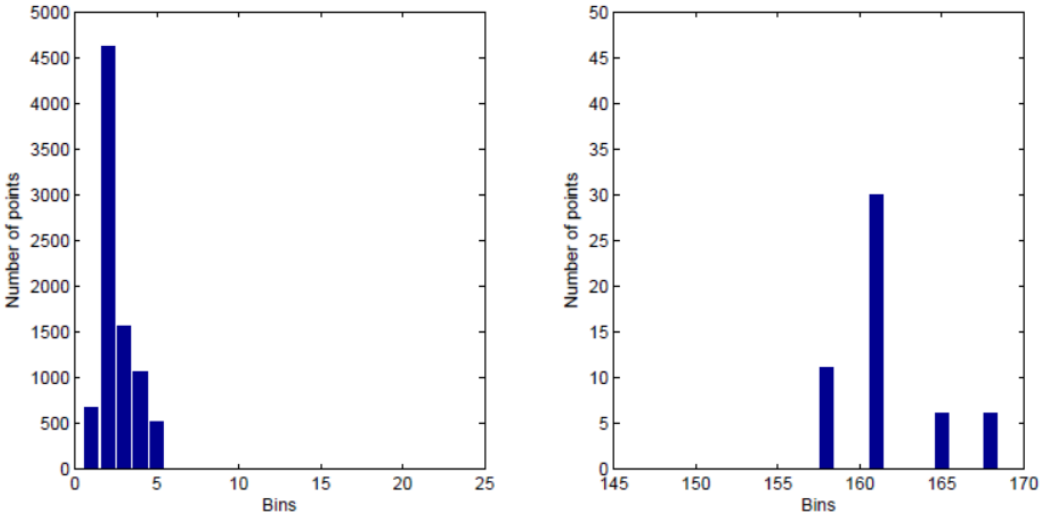
3. Proposed Scheme

In this section, we detail our proposed scheme. Firstly, the shape of target is segmented from the rectangular target area on the depth image as the spatial reliability map. Secondly, CCF is constructed to eliminate the boundary effects with the spatial reliability map. Thirdly, for the next frame, the target location is estimated by the final response which is the sum of the product of each feature channel's response and the corresponding channel reliability weights. Finally, the occlusion is detected and handled by the final response and the result of segmentation.

3.1 Depth Image Segmentation

In DCF formulation, the size of filter and the shape of target shall be fixed. Since the target shape is arbitrary, the filter will inevitably be corrupted by the background at the learning stage. If target shape is available, the limitation of fixed shape can be overcame. Depth information gains robustness in many scenarios such as scale changes, shape changes and occlusion. Our proposed method can get the mask of the target by depth information which is either one for the target or zero otherwise.

The K-Means clustering algorithm is capable of extracting target from colour image[22.23]. The main drawback of K-Means is that its result is very sensitive to the initial cluster seeds and outlier, and that number of cluster K needs to be known in advance. As the computational burden of K-Means may lead to a disaster when the amount of data is large, we implement the K-Means cluster on the depth histogram to reduce the number of points to be clustered. The local maxima of the depth histogram $h(d_j)$ are good seeds d^{t_0} for K-Means to reduce the convergence time which is determined by Non-Maxima Suppression (NMS). The number of cluster K equals to the number of local maxima.



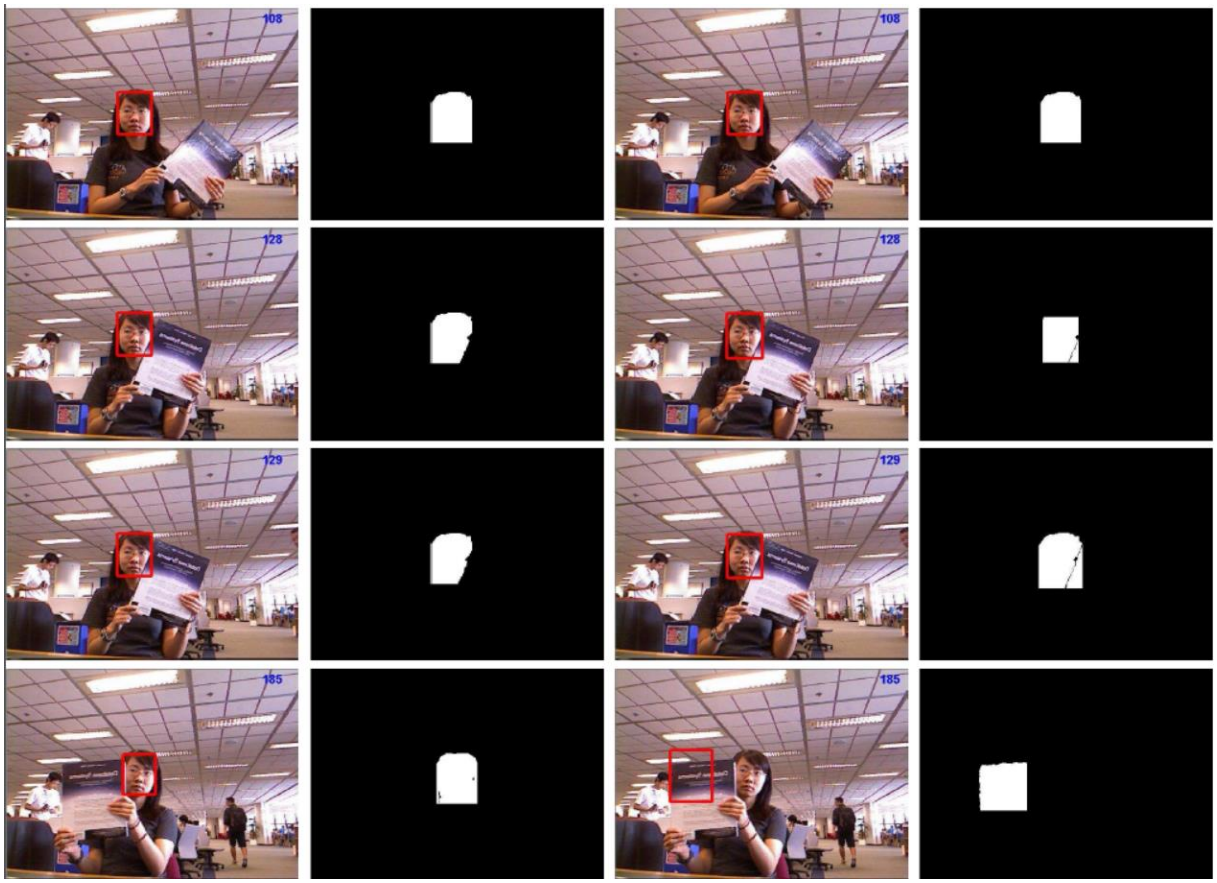
(a) (b)

Figure 2.Depth histogram at the frame 108 of sequence face_occ5 in Princeton datasets.

Let $h(d_j)$ be the depth histogram consisted of j bins with depth value d_j, where each bin is assigned to the closest cluster k. Then updated method of the cluster's centroid becomes

$$d_k^{t+1} = \left(\sum_{d_j \in k} h(d_j)\right)^{-1} \sum_{d_j \in k} d_j \cdot h(d_j). \quad (5)$$

After the algorithm convergences, the cluster with minimum mean depth will be selected as the target cluster. All the points of which the depth value is in the range of target cluster will be selected as the candidate points for target masking. Connected components are formed from the candidate



(a) (b) (c) (d)

Figure 3. The tracking result and mask of "face_occ5" video.

points in the image plane to distinguish objects located within the same depth plane, and to remove clusters corresponding small regions. The spatial distribution of the target is described by the mean μ_{obj} and the standard deviation δ_{obj} of the depth value corresponds the target region. To ensure a correct segmentation, the bin width b_w of depth histogram is adjusted by the δ_{obj} and the noise model N_{depth} of the depth sensor[24], which may be expressed as

$$b_w = \max(\delta_{obj}, N_{depth}). \quad (6)$$

The presence of outliers has great impacts on the mean μ_{obj} and the standard deviation δ_{obj} of a target. More specifically, if the depth distribution of the target is highly dense, and the region of background is relatively small, there may be only one peak. In this case, all the depth data will be within a same cluster. The corresponding mean μ_{obj} and standard deviation δ_{obj} may increase, and the discrimination between target and occlusion will decline. If outliers are detected, we add a K-Means seed at the end of the depth histogram as background to ensure a correct estimation of the target. An example of depth segmentation is shown in figure 2 and figure 3. The depth histogram at the frame 108 of video face_occ5 in Princeton datasets is shown as figure 2 in which (a) and (b) are the depth histogram in the

depth bin [0, 25] and [145, 170], respectively. All bins not display are zeros. If there is only one cluster, the standard deviation of the target's depth will rise from 29mm to 138mm, leading to a tracking result and mask of the video is shown in figure 3. The left two columns with the method adding a cluster seed at the end of depth histogram and the right two without. The left two columns discriminate between target and occlusion with accurate standard deviation, but the right's filter is corrupted by the occlusion object, where it may be seen the performance of the method with outliers handling outperforms than without.

3.2 Constructing CCF

We get the spatial reliability map using depth image segmentation to construct the CCF, which allows the system to enlarge the search region and to improve the performance of tracking non-rectangular objects. The constraint can be formalized as $h \equiv h \odot m$, where $\odot$ denotes the element-wise product. Such a constraint does not lead to a closed-form solution, but an iterative approach can be employed for efficiently solving the optimization problem. We first introduce a dual variable h_c and construct the constraint as follow:

$$h_c - m \odot h \equiv 0, \tag{7}$$

which leads to the following augmented Lagrangian

$$L\ (h_c, h, \hat{I} \mid m) = \left\| diag(f)\bar{h}_c - g \right\|^2 + \frac{\lambda}{2}\left\| h_m \right\|^2 + \left[\hat{I}^H (h_c - h_m) + \overline{\hat{I}^H (\bar{h}_c - \bar{h}_m)} \right] + \mu \left\| h_c - h_m \right\|, \tag{8}$$

Where $\hat{I}$ is a complex Lagrange multiplier, $\mu > 0$, and $h_m = h \odot m$ is defined for compact notation. The augmented Lagrangian can be iteratively minimized by the alternating direction method of multipliers (ADMM)[25], which sequentially solves the following sub-problems at each iteration:

$$h_c^{i+1} = \arg\min_{h_c} L\ (h_c, h^i, I^i \mid m), \tag{9}$$

$$h^{i+1} = \arg\ \min_{h} L\ (h_c^{i+1}, h, I^i \mid m), \tag{10}$$

and the Lagrange multiplier is updated as:

$$I^{i+1} = I^i + \mu(h_c^{i+1} - h^{i+1}). \tag{11}$$

Minimizations in equation (9) and equation (10) at each iteration have a close-form solution of

$$h_c^{i+1} = (f \odot \bar{g} + (\mu h_m^i - h^{i+1})), \tag{12}$$

$$h^{i+1} = m \cdot (\lambda \cdot (2D)^{-1} + \mu^i)^{-1} F^{-1}[I^i + \mu^i h_c^{i+1}], \tag{13}$$

where constraint penalty μ is updated by $\mu^{i+1} = \beta\mu^i$, and β is empirically set to 3.

3.3 Channel Reliability

We introduce the channel reliability weights to DCF tracking. In the framework of DCF, the final response is the sum of all the feature channels, irrespective of its discriminability. Each feature's (e.g., HoG[11], CN[12] and grayscale feature) response has an order of magnitude difference in scale. To avoid the issue with different scales, each channel is considered independently. The cost function may be reformulated as

$$\xi(h) = \sum_{d=1}^{N_c} \left\| f_d \bar{h}_d - g \right\|^2 + \lambda \left\| h_d \right\|^2. \tag{14}$$

Now, let us introduce the channel reliability weights $w = (w_d)_{d=1:N_c}$, which can be considered as discriminability. Then, the final response becomes the sum of the product of each feature channel's response and the corresponding channel weights w_d may be expressed as

$$g(h) = \sum_{d=1}^{Nc} f_d * h_d \cdot w_d, \tag{15}$$

where channel reliability weights w_d consists of two reliability measurements, namely the channel learning reliability w_d^{lrn} reflecting the discriminability of the feature channel[16], which is calculated at the filter learning stage, and the channel detection reliability $w_d^{\det}$ reflecting the uniqueness[26], which is calculated at the target localization stage. The channel learning and detection reliability is estimated as

$$w_d^{lrn} = \max(f_d * h_d), \tag{16}$$

and

$$w_d^{\det} = \max(1 - \rho_d^{\max 2} \cdot (\rho_d^{\max 1})^{-1}, 0.5), \tag{17}$$

respectively. We have $\rho_d^{\max 1}$ and $\rho_d^{\max 2}$ denoting the two largest peaks in the response map after NMS. The joint channel reliability w_d at target localization stage is computed as below

$$w_d = w_d^{lrn} \cdot w_d^{\det}, \tag{18}$$

and w_d is normalized as $\sum_d w_d = 1$.

3.4 Detection and Handling Occlusion

The occlusion is detected by both the response of filter and the cluster output. The response between frames has no uniform scale after being weighted by the reliability of channel. As a result, it can be used for finding the maximum value in a same frame but not for comparing between frames. Therefore, at the learning stage of frame $t-1$, the max response without channel weighting $r_{lrn\_\max\_without}^{t-1}$ is calculated and saved for the next frame. At the location stage of frame t, both the reponses with the channel weight $r_{\det\_\max\_with}^{t}$ and without $r_{\det\_\max\_without}^{t}$ are calculated. $r_{\det\_\max\_with}^{t}$ is used for location and the similarity ratio r^t between the frame t and $t-1$ is used for deciding whether occlusions occur

$$r^t = r_{\det\_\max\_without}^{t} \cdot (r_{\det\_\max\_without}^{t-1})^{-1}. \tag{19}$$

Furthermore, the occlusion is detected by both r^t and cluster output as

$$(r^t < \lambda_{r1}) \wedge (p > \lambda_{occ}), \tag{20}$$

where λ_{r1} is the similarity threshold and λ_{occ} is the occlusion threshold. Here, p from cluster output represents the fraction of pixels which does not belong to the target cluster but locates in the rectangle area of target. The value of λ_{occ} and λ_{r1} are determined empirically as $\lambda_{r1} = 0.4$ and $\lambda_{occ} = 0.35$. In addition, when $p > 10\%$, part of target will be occluded, and the mask of target may be occupied by the occluding object. In this case, the model is stopped from updating to prevent it from drifting.

In a state of occlusion, the occluding object is segmented from the depth image and tracked by a new DS-KCF[18] tracker with fixed scale. Searching region is centered on the location of the occluding object. The similarity ratio r_c^t of each cluster output result in the searching region is calculated instead of calculating every location in the searching region to reduce computation cost. Target tracking will be resumed when

$$(r_{\max}^t > \lambda_{r2}) \wedge (p < \lambda_{occ}), \tag{21}$$

where, $r_{\max}^t$ is the maximum of r_c^t, and λ_{r2} is empirically set as $\lambda_{r2} = 0.2$.

4. Experiments

In this section, our result is reported on the original Princeton dataset since it is closer to the practical applications than the rectified version, with synchronization errors in 14% of the sequence[19]. In fact, the RGB-D sensor captures depth and colour image independently, and synchronization errors are inevitable.

HOG feature and CN feature are the most common features in visual tracking methods, CSR-DCF exploit the gray feature. For fair comparison, we use the same features and parameter values as

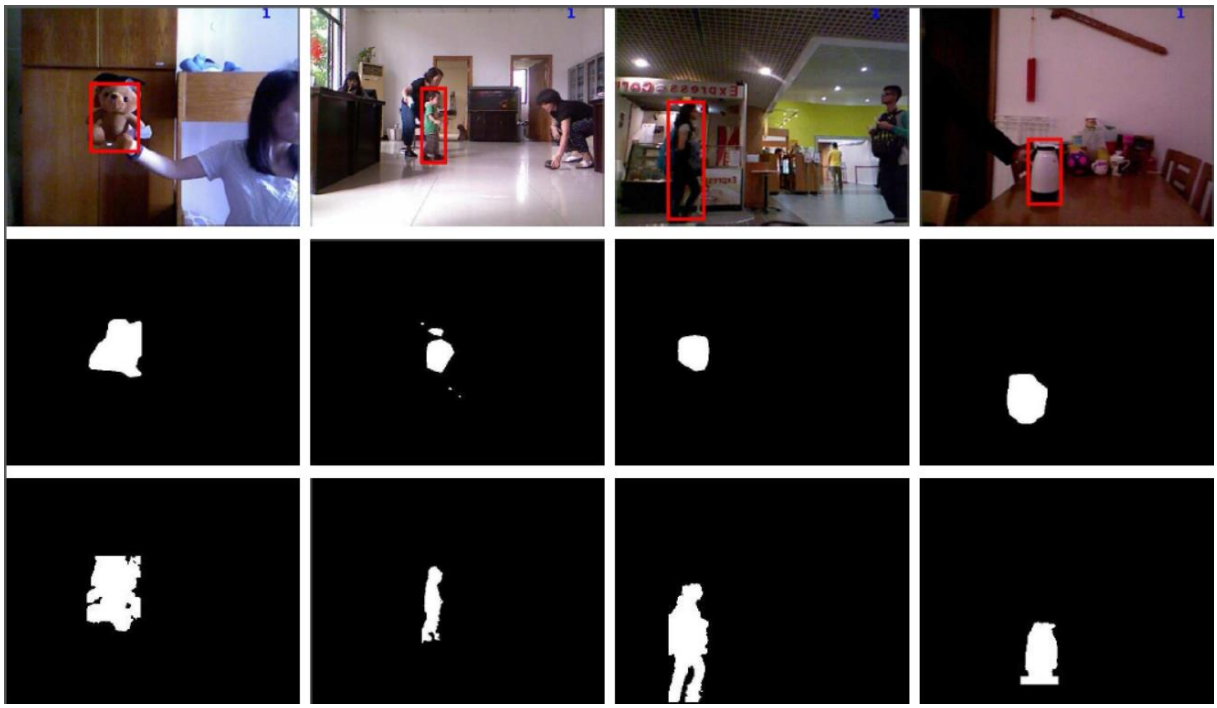


Figure 4. Performance comparison of segmentation method.

CSR-DCF except that the padding is set to 5. The weight of the regularization term λ and constraint penalty term μ are set to 0.05 and 5, respectively. Same parameter value included learning rate and size of HOG, is adopted as DS-KCF when tracking the occluding object. Our proposed method is

Table 1. Results of IoU and rankings evaluated on the Princeton dataset.

Alg	Avg Rank	target type			target size		movement		Occlusion		motion type	
		Human	Animal	Rigid	Large	Small	Slow	Fast	Yes	No	Passive	Active
CSRr-gbd++	3.82	0.77	0.65	0.76	0.75	0.73	0.80	0.72	0.70	0.79	0.79	0.72
our	4.09	0.70	0.65	0.79	0.71	0.73	0.78	0.70	0.64	0.84	0.84	0.67
OAPE	4.45	0.64	0.85	0.77	0.73	0.73	0.85	0.68	0.64	0.85	0.78	0.71
3D-T	4.64	0.81	0.64	0.73	0.80	0.71	0.75	0.75	0.73	0.78	0.79	0.73
RGBDOcc+OF	4.82	0.74	0.63	0.78	0.78	0.70	0.76	0.72	0.72	0.75	0.82	0.70
DSKCF_shape	6.00	0.71	0.71	0.74	0.74	0.70	0.76	0.70	0.65	0.81	0.77	0.70
DM-DCF	6.09	0.76	0.58	0.77	0.72	0.73	0.75	0.72	0.69	0.78	0.82	0.69
DS-KCF	8.00	0.67	0.61	0.76	0.69	0.70	0.75	0.67	0.63	0.78	0.79	0.66
DSKCF-CPP	8.09	0.65	0.64	0.74	0.66	0.69	0.76	0.65	0.60	0.79	0.80	0.64
hiob_lc2	8.55	0.53	0.72	0.78	0.61	0.70	0.72	0.64	0.53	0.85	0.77	0.62
RGBD+OF	9.00	0.64	0.65	0.75	0.72	0.65	0.73	0.66	0.60	0.79	0.74	0.66

implemented with Matlab 2014a and run on a desktop computer with Intel i7 6700 CPU, 16GB RAM and Ubuntu 14.04 OS. The average speed of our proposed method is 6.38 frames per second.

The evaluation metric of the Princeton dataset is Intersection over Union (IoU). If the ratio of overlap area r_i between our results and true bounding boxes is greater than the threshold r_t, the tracking result is success. There are 100 sequences with 11 categories in the dataset, with 5 videos' ground truth being published. The average ranks and IoU can be obtained by uploading our result of the 95 videos to the website of the Princeton dataset online. Where it may be seen that in Table 1, the average rank of our method is just below the csr-rgbd++. The corresponding results show that our method significantly outperforms the DS-KCF, since the spatial and channel reliability map can overcome the limitation of boundary effects and the rectangular fixed shape assumption. Compared to

DM-DCF, our method performs more robustly at most categories, especially at the categories of Animal, Passive and No occlusion. The reason is that more accurate segmentation method and more reasonable occlusion handling mechanism are adopted in our method. Our segmentation method is compared with the segmentation method using spatial 2D priors and colour segmentation in figure 4. The second row and the third row are the results of the colour segmentation method and proposed depth image segmentation method, respectively. Proposed depth image segmentation method has the significantly advantages. Finally, our method wins three categories: Rigid, Small and Passsive in all of the methods.

5. Conclusion

This paper proposed a depth information aided constrained correlation filter for visual tracking. The mask of target is available using the depth segmentation method. Then, the CCF is constructed with the mask, which can eliminate the boundary effects and the limitation of the rectangular fixed shape assumption. Our proposed method integrates the depth and the colour image, and gains the better performance on the Princeton dataset among all public tracking algorithms. The evaluation of Princeton dataset verified that our proposed method is more robust than others.

6. Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grants (No.s 61601304, 61801302 and 61702335), in part by Foundation of ShenZhen under Grants (JCYJ20170302142545838 and JCYJ20170302142545838154149766), in part by Foundation of ShenZhen university under Grants (No.2016057).

7. References

[1] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[M] *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, 2012.

[2] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 2014, **37(3)**, pp 583-596.

[3] Danelljan M, Hager G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking[J] in *IEEE International Conference on Computer Vision(ICCV)*, 2016, pp 4310–4318.

[4] Danelljan M , Robinson A , Khan F S , et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking[M] *Computer Vision – ECCV 2016*. Springer International Publishing, 2016.

[5] Li F, Tian C, Zuo W, et al. "Learning spatial-temporal regularized correlation filters for visual tracking[J]" Preprint cs.cv/1803.08679 (2018).

[6] Danelljan M, Bhat G, Khan F S, et al. Eco: efficient convolution operators for tracking [C] *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2017.

[7] Li Y and Zhu J, A scale adaptive kernel correlation filter tracker with feature integration[C] *European Conference on Computer Vision(ECCV)*. *Springer*, Cham, 2014.

[8] Danelljan M, Hager G, Khan F S, et al. Discriminative scale Sspace tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TP*AMI), 2017, 39(8), pp 1561-1575.

[9] Hester C F, Casasent D. Multivariant technique for multiclass pattern recognition[J]. *Applied Optics*, 1980, **19(11)**, pp 1758-1761.

[10] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C] in Computer Vision and Pattern Recognition(CVPR), 2010, pp 2544–2550.

[11] He N, Cao J and Song L. Scale space histogram of oriented gradients for human detection[C], in *International Symposium on Information Science and Engieering*, 2008, pp 167–170.

[12] Danelljan M, Khan F, Felsberg M, et al. Adaptive color attributes for real-time visual tracking[C] in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2014, pp 1090–1097(2014).
[13] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C] in *IEEE International Conference on Computer Vision(ICCV)*, 2016 pp 3074–3082.
[14] Qi Y, Zhang S, Qin L, et al. Hedged deep tracking in *Computer Vision and Pattern Recognition(CVPR)*, 2016 pp 4303–4311.
[15] Galoogahi H K, Sim T, Lucey S. Correlation filters with limited boundaries[J] in *Computer Vision and Pattern Recognition(CVPR)*, 2015 pp 4630–4638.
[16] Lukezic A, Vojir T, Zajc L C, et al. Discriminative correlation filter with channel and spatial reliability[C] in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* 2017 pp 4847–4856.
[17] Song S, Xiao J. Tracking revisited using RGBD camera: Unified benchmark and baselines[C] in *IEEE International Conference on Computer Vision(ICCV)*, 2014 pp 233–240.
[18] Hannuna S, Camplani M, Hall J, et al. DS-KCF: a real-time tracker for RGB-D data[J]. *Journal of Real-Time Image Processing*, 2016, pp 1-20.
[19] Bibi A, Zhang T, Ghanem B, 3d part-based sparse tracker with automatic synchronization and registration[C] in *Computer Vision and Pattern Recognition(CVPR)*, 2016 pp 1439–1448.
[20] Kart U, Kämäräinen J K, Matas J, et al. Depth masked discriminative correlation filter[J]. Prepint cs.cv/ 1802.09227 2018.
[21] Chrapek D, Beran V, Zemcik P. Depth-Based filtration for tracking boost[J]. *Springer International Publishing* cham 2015 pp 217–228.
[22] Chen T W, Fast image segmentation based on k-means clustering with histograms in hsv color space[J] in *Multimedia Signal Processing*, 2008 IEEE Workshop on, pp 322–325 (2008).
[23] Jin R, Kou C, Liu R, et al. A color image segmentation method based on improved K-Means clustering algorithm proceedings[M] in the *International Conference on Information Engineering and Applications (IEA)* 2012. Springer London, 2013.
[24] Khoshelham K and Elberink S O, Accuracy and resolution of kinect depth data for indoor mapping applications[J] *Se*nsors 12(2), 2012 p 1437.
[25] Boyd S, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J] *Foundations and Trends in Machine* Learning 2010 3(1), pp 1–122 (2010).
[26] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C] in *Computer Vision and Pattern Recognition(CVPR)*, 2010 pp 2544–2550.