ICIC Express Letters
Part B: Applications
Volume **9**, Number **5**, May **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **381–387**

OBJECT TRACKING WITH SUPERPIXEL-BASED MODEL

Shinfeng D. Lin and Dung-Han Yang

Department of Computer Science and Information Engineering
National Dong Hwa University
No. 1, Sec. 2, Da Hsueh Rd., Shoufeng, Hualien 97401, Taiwan
david@gms.ndhu.edu.tw; odinqq@gmail.com

Received November 2017; accepted February 2018

Abstract. *In this paper, we propose a superpixel-based model for object tracking in the Bayesian framework. The mid-level visual cues via superpixel with sufficient structure information are used to represent the object. This tracking algorithm contains three models, including appearance model, motion model, and online learning model. By adding texture feature, our method has improved the previous superpixel tracking (SPT) method on reducing the interference of neighboring similar color and increasing the success rate of grayscale sequences. We conducted experiments using videos from literature. Experimental results show that the proposed method outperforms the existing methods.*
Keywords: Object tracking, Superpixel, Appearance model, Online-learning model, OTSM (object tracking with superpixel-based model)

1. **Introduction.** Object tracking has been one of the major issues in computer vision and related fields such as intelligent surveillance, traffic monitoring, and human-computer interaction. Many object tracking algorithms [1-3] were proposed to face various challenges, including large variation of scale, heavy occlusion and drifts. For example, the multiple instance learning (MIL) [1] uses positive and negative samples to learn a discriminative model for object tracking. The tracking-learning-detection (TLD) [4] tracks objects and learns their appearance at the same time, detecting larger area when heavily occluded.

Many object tracking methods, based on the pixel information, have been proposed for years. However, a single pixel can only provide little information for human vision since human eyes are merely able to recognize a meaningful image composed of a group of associated pixels. Superpixel is an area which consists of a series of adjacent pixels, having similar color, brightness, texture, and other characteristics. These segmental areas provide effective message from small portions of an image, and retain the edge's information of target object. Currently, many researches develop a method of superpixel algorithm to solve numerous vision problems, including object segmentation [5-7], saliency detection [7-9], and recognition [10,11]. However, less attention has been paid to mid-level visual cues for object tracking. Yang *et al.* [12] first used the method of superpixel tracking (SPT) to separate target objects from background according to a superpixel-based appearance model. Some literatures have been published based on superpixel-based method such as human tracking [13], traffic surveillance [14], and superpixel-driven level set tracking [15]. Nevertheless, the method proposed by Yang *et al.* [12] is not effective for grayscale images because of the color-based characterization. Therefore, additional features should be considered to improve the tracking accuracy.

In addition to the color-based feature, in this article, we also consider the texture feature in our proposed tracking algorithm to improve Yang et al.'s method. The proposed tracking algorithm contains three models, including appearance model, motion model, and online learning model. By adding texture feature, our method has improved the previous

superpixel tracking (SPT) method on increasing the success rate of grayscale sequences and reducing the interference of neighboring similar color. Experiments conducted using videos from literature show that the proposed method outperforms the existing state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 presents our proposed tracking method. Experimental results are demonstrated in Section 3. Finally, Section 4 summarizes the conclusions.

2. **The Proposed Tracking Method.** The procedures of our proposed tracking algorithm are addressed as follows. First of all, we will demonstrate the Bayesian theory used in our tracking algorithm which contains three models: (1) appearance model, (2) motion model, and (3) online learning model. The tracking algorithm is shown in Figure 1 and described in the following.

- Step 1: In the appearance model, we preform superpixel segmentation to obtain the color and texture features. Using these features, we then generate the confidence maps to separate the target from background.
- Step 2: We utilize the particle filter to select N candidates in the motion model. To find the best result, we apply the observation and motion models to estimating the posterior of the most likely object movement.
- Step 3: Based on the above result, we create a continuously updating online learning model to take account of the appearance changes during the tracking process, and reduce the occlusion effect.

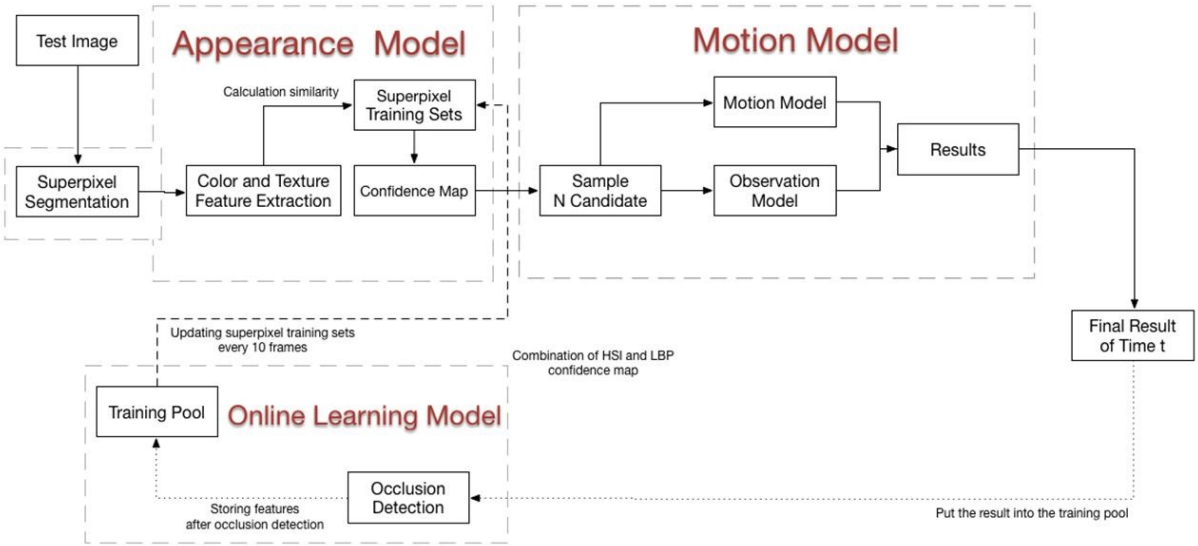


FIGURE 1. The flow chart of our tracking algorithm

Besides, the online learning model in the tracking algorithm can be used to handle large variations of object sizes, appearance change and tracking drift. Experimental results will demonstrate the superiority of the proposed method. In the following, we first introduce the Bayesian theory in Section 2.1, and then illustrate Steps 1, 2, 3 in Sections 2.2, 2.3 and 2.4, respectively.

2.1. **The Bayesian theory.** Our tracking algorithm is based on Bayesian statistical inference used in appearance model and motion model. The maximum posterior probability $p(x|z_{1:t})$ of the object position at time t is described below:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx_{t-1} \tag{1}$$

where x_t is the object state at time t, and $z_{1:t}$ indicates the observation from time 1 to time t. $p(z_t|x_t)$, $p(x_t|x_{t-1})$ indicate the appearance model and motion model, respectively.

In appearance model, x_t is defined as $x_t = (x_t^{cx}, x_t^{cy}, x_t^{sx}, x_t^{sy})$, where x_t^{cx} and x_t^{cy} represent the center location x and y of the object, x_t^{sx} and x_t^{sy} indicate the size of bounding box in the x-axis and y-axis, respectively. The motion model $p(x_t|x_{t1})$ describes the probability of position between two consecutive frames.

The appearance model $p(z_t|x_t)$ indicates the likelihood of the observation z_t at object state x_t, which plays an important role in the object tracking. The appearance model in our method is built by:

$$p(z_t|x_t) \propto \hat{C}(x_t) \tag{2}$$

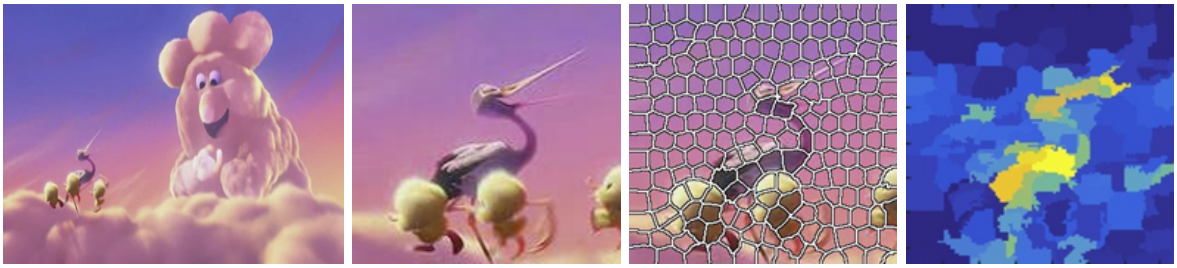
where $\hat{C}(x_t)$ performs the confidence of an observation of the object state at time t, which can be obtained by the maximum a posterior estimation over the N samples in the bounding boxes at each t. We set x_t^l to indicate the l-th sample of the object state x_t.

$$\hat{x}_t = \arg\max_{x_t^l} p\left(x_t^l|z_{1:t}\right), \quad \forall l = 1, \ldots, N \tag{3}$$

where $\hat{x}$ indicates the estimation of the maximum a posterior among N samples.

2.2. **Appearance model and confidence map.** Based on Equation (3), when a new frame arrives, the system estimates the probabilities of random positions of N bounding boxes (i.e., x_t), and defines the best position by the highest probability. This procedure is so called appearance model which is based on confidence map. Therefore, the first important task is to generate a confidence map. The procedure to construct a confidence map and an appearance model for both target and background is described below.

- Step 1: A new frame arrives at time t (Figure 2(a)). Then extract the surrounding region of the target in the frame t (Figure 2(b)).
- Step 2: Perform a superpixel segmentation as shown in Figure 2(c).
- Step 3: Build a confidence map [12] of superpixels as shown in Figure 2(d).
- Step 4: Create an appearance model based on the confidence map.



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

FIGURE 2. The procedure to construct a confidence map and an appearance model

2.3. **Motion model.** After finishing the appearance model, we obtain N different likelihood probabilities of x_t. To obtain more reliable tracking result, we further weigh each probability of x_t by the motion model. The motion model is constructed by the distribution of the positional difference between two nearby frames, x_t and x_{t-1}. This distribution is assumed to be Gaussian distributed:

$$p(x_t|x_{t-1}) = N(x_t; x_{t-1}, \Psi) \tag{4}$$

where Ψ is a diagonal covariance matrix, which is the function of the standard deviations for location and scale, i.e., σ_c and σ_s. The values of σ_c and σ_s define the motion and scale change in the proposed algorithm.

2.4. **Online learning model.** The motivation to do the online learning model is to keep our feature pool F continuously updated during the tracking process. In this way, we can always have the most current model of appearance features instead of using unchangeable appearance model (i.e., off-line appearance model). This can help to increase the tracking accuracy and reduce the effects of occlusion and drifts.

During the tracking process, we build a training pool to store the information of object features, including colors and texture features. A sequence of 10 frames is stored in the pool. The first four frames are fixed in the entire training process to reduce the feature contamination by occlusion, scale change and non-rigid deformation during long term tracking. The rest of six frames are used to update the new appearance features following the flow chart in Figure 3. The flow chart shows the strategy of updating method depending on the condition of occlusion.

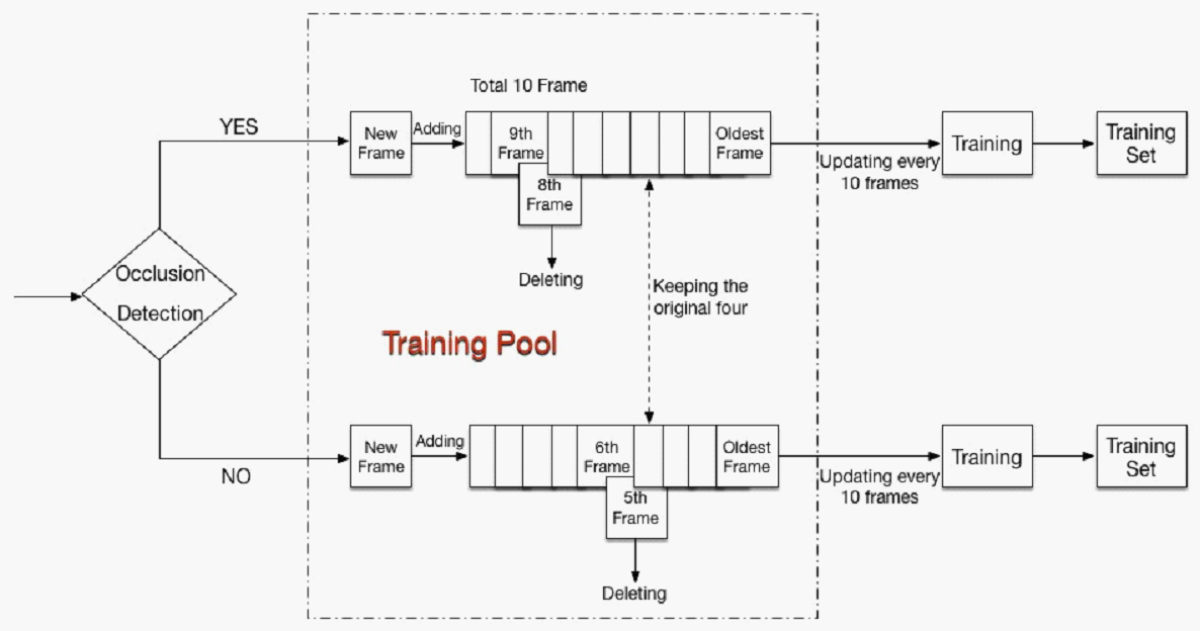


FIGURE 3. The strategy of updating method in the online learning model

If no occlusion happened: when a new frame arrives, we add the new one to be the latest 10^{th} frame in the pool and delete the oldest 5^{th} one. The 1^{st} to 4^{th} frames are fixed.

If occlusion happened: when occlusion occurs, the incoming frame becomes unreliable; thus we only update the last three frames. When a new frame arrives, we add the new one to be the latest 10^{th} frame in the pool and delete the oldest 8^{th} one. The 1^{st} to 7^{th} frames are fixed.

The detection of occlusion, O_t, is defined by the method [12] as below:

$$O_t = \mu_c - \frac{\max\left(\{C_l\}_{l=1}^{N}\right)}{S(x_{t-1})} \tag{5}$$

where μ_c is the average of normalized confidence of the target measures in the sequence of 10 frames. The ratio of maximum confidence value ($\max(\{C_l\}_{l=1}^{N})$) to the size of target region at time $t-1$, ($S(x_{t-1})$), indicates that how much fraction of area is occluded.

The occlusion detection is repeated every frame, and k-means clustering is applied every ten frames for updating the color and texture training sets. Taking advantage of these continuously updated training sets, the system is adaptive to the changes of environment, and the target object can be effectively separated from the background.

3. **Experimental Results.** We evaluated the effectiveness of the proposed algorithm on 12 video sequences. These sequences were tested in [12] (i.e., SPT sequences). Both

SPT and LBP methods using color and texture features were utilized to characterize the objects. The system is operated using the Matlab (version 2014a) and is performed on a 2.2 GHz 4 cores processor with 8 GB memory. The tracking results for 12 sequences will be compared with previous works in the following discussion. The experimental environment and results are presented below.

3.1. **Experimental environment.** As mentioned in Section 2, our system includes three models: appearance model, motion model and online learning model. In the appearance model, the number of superpixels is set to ~ 300. The parameter of rescaling surrounding region, i.e., λ, is set to 1.5. In the motion model, the standard deviations for location and scale, σ_c and σ_s are set to 7.6 and 7. In the online learning model, the number of frames we store in the training pool is set to 10. The average of normalized confidence, μ_c, is set to 0.5. We set the threshold of occlusion detection $\theta_o = 0.525$. The parameters described above are fixed for all sequences.

Finally, instead of using mean-shift algorithm, we use k-means to cluster the superpixels because of its shorter time for calculation (around 3 to 5 times faster than the mean-shift method in our experiments). The experimental datasets used to examine our algorithm are from [12].

3.2. **Experimental results.** To quantify the quality of our tracking results, we evaluate the success rate and center error for each sequence. For a tracked bounding box and a ground truth box, we regard the percentage of their overlapping area as the success rate.

Following the settings in [12], the tracking result of SPT sequences is considered to be successful when the success rate is greater than 50%. The overall success rates of our results are displayed in the last column of Table 1 (i.e., SPT-LBP). The higher value means the better result. Furthermore, the center error is defined as the center distance between the ground truth and the tracked bounding box [12]. The overall center errors of our results are displayed in the last column of Table 2.

Table 1. The success rates of SPT sequences in percentage

Sequences	IVT	Frag	MIL	PROST	VTD	L1	TLD	Struck	HT	SPT	SPT-LBP
lemming	79	51	83	82	35	17	27	49	21	96	96
liquor	23	79	20	83	27	57	80	23	0	97	99
singer1	93	25	25	–	100	100	100	25	23	99	92
basketball	11	28	28	–	85	4	6	12	60	96	99
woman	8	7	7	–	5	9	6	56	8	50	60
transformer	23	24	24	–	38	27	35	27	59	100	100
bolt	1	3	3	–	57	4	14	3	1	66	91
bird1	1	29	29	–	2	2	6	4	32	21	36
bird2	9	83	83	–	9	74	12	14	81	87	93
girl	7	38	38	–	55	26	11	16	0	96	98
surfing1	8	3	3	–	8	8	40	8	9	34	73
racecar	2	4	4	–	6	8	3	7	8	46	91
Average	22	31	29	83	36	28	28	20	25	74	85

In Table 1, our success rate in average is 85, which is better than that of the famous article [12], 74. In addition, the center error in average is 12 compared with 15 [12]. Based on the comparison results in Tables 1 and 2, it is noted that our proposed method outperforms Yang et al.'s method [12] and other existing methods.

4. **Conclusions.** In this paper, we have proposed object tracking with superpixel-based model. The tracking algorithm contains three models, including appearance model, motion model, and online learning model. This tracking method has improved the previous

TABLE 2. The center errors of SPT sequences

Sequences	IVT	Frag	MIL	PROST	VTD	L1	TLD	Struck	HT	SPT	SPT-LBP
lemming	14	84	14	23	98	182	104	134	118	7	7
liquor	238	31	165	22	155	80	28	124	202	9	8
singer1	5	20	20	–	3	3	5	16	52	5	12
basketball	120	14	104	–	11	100	170	153	19	6	7
woman	133	112	120	–	109	113	95	5	122	11	12
transformer	130	47	33	–	43	108	23	54	31	14	11
bolt	382	100	380	–	14	369	90	387	373	6	7
bird1	230	223	270	–	250	226	77	148	203	47	39
bird2	119	28	18	–	50	19	86	88	10	17	17
girl	184	106	55	–	57	177	151	119	232	10	9
surfing1	141	199	319	–	84	228	27	265	287	48	9
racecar	340	94	104	–	196	224	134	202	228	4	4
Average	170	88	134	23	89	152	83	141	156	15	12

superpixel tracking (SPT) method by adding texture feature to reduce the interference of neighboring similar color and increase the success rate of grayscale sequences.

We evaluate the effectiveness of the proposed algorithm on video sequences from [12]. The system demonstrates an improvement on the tracking drift caused by background clutter, occlusion, grayscale tracking, distortion, scale and pose variation. Most of our tracking results have better success rates than those of previous works. These results prove that the SPT-LBP is a robust tracking method for both color and grayscale images. In the future, more techniques can be compared to further demonstrate the performance of the proposed algorithm.

REFERENCES

[1] B. Babenko, M.-H. Yang and S. Belongie, Visual tracking with online multiple instance learning, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.983-990, 2009.

[2] J. Kwon and K. M. Lee, Visual tracking decomposition, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1269-1276, 2010.

[3] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. Hicks and P. Torr, Struck: Structured output tracking with kernels, *IEEE Trans. Pattern Analysis and Machine Intelligence*, no.99, p.1, 2015.

[4] Z. Kalal, K. Mikolajczyk and J. Matas, Tracking-learning-detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.34, no.7, pp.1409-1422, 2012.

[5] B. Fulkerson, A. Vedaldi and S. Soatto, Class segmentation and object localization with superpixel neighborhoods, *IEEE the 12th International Conference on Computer Vision*, pp.670-677, 2009.

[6] X. Li and H. Sahbi, Superpixel-based object class segmentation using conditional random fields, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1101-1104, 2011.

[7] R. Achanta, *Finding Objects of Interest in Images Using Saliency and Superpixels*, Ph.D. Thesis, 2011.

[8] Z. Liu, L. Meur and S. Luo, Superpixel-based saliency detection, *The 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pp.1-4, 2013.

[9] Z. Liu, X. Zhang, S. Luo and O. Le Meur, Superpixel-based spatiotemporal saliency detection, *IEEE Trans. Circuits and Systems for Video Technology*, vol.24, no.9, pp.1522-1540, 2014.

[10] H.-M. Zhu and C.-M. Pun, An adaptive superpixel based hand gesture tracking and recognition system, *The Scientific World Journal*, 2014.

[11] F. Liu, Y. Yin, G. Yang, L. Dong and X. Xi, Finger vein recognition with superpixel-based features, *IEEE International Joint Conference on Biometrics (IJCB)*, pp.1-8, 2014.

[12] F. Yang, H. Lu and M.-H. Yang, Robust superpixel tracking, *IEEE Trans. Image Processing*, vol.23, no.4, pp.1639-1651, 2014.

[13] H. Zhang, J. Zhan, Z. Su, Q. Chen and X. Luo, Online human tracking via superpixel-based collaborative appearance model, *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp.1-6, 2014.

[14] D.-T. Lin and C.-H. Hsu, Crossroad traffic surveillance using superpixel tracking and vehicle trajectory analysis, *The 22nd International Conference on Pattern Recognition (ICPR)*, pp.2251-2256, 2014.

[15] X. Zhou, X. Li, T.-J. Chin and D. Suter, Superpixel-driven level set tracking, *The 19th IEEE International Conference on in Image Processing (ICIP)*, pp.409-412, 2012.