Spatiogram and Fuzzy Logic Based Multi-modality Fusion Tracking with Online Update

Canlong Zhang[1,2(✉)], Zhixin Li[1,2], Zhiwen Wang[3], and Ting Han[1]

[1] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, China
zcltyp@163.com
[2] Guangxi Experiment Center of Information Science, Guilin University of Electronic Technology, Guilin, China
[3] College of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou, China

Abstract. Multi-modality fusion tracking is an interesting but challenging task. Many previous works just consider the fusion of different features from identical spectral image or identical features from different spectral images alone, which makes them be quite distinct from each other and be difficult to be integrated naturally. In this study, we propose an unified tracking framework to naturally integrate multiple different modalities via innovative use of spatiogram and fuzzy logic. Specifically, each modal target and its candidate are first represented by second-order spatiogram and their similarity is measured. Next, a novel objective function is built by integrating all modal similarities, and then a joint target center-shift formula is gained by performing mathematical operation on the objective function. Finally, the optimal target location is gained recursively by applying the mean shift procedure. Besides, a model update scheme via particle filter is developed to capture the appearance variations. Our framework allows the modalities to be original pixels or other extracted features from single image or different spectral images, and provides the flexibility to arbitrarily add or remove modality. Tracking results on the combination of infrared gray-HOG and visible gray-LBP clearly demonstrate the excellence of the proposed tracker.

Keywords: Spatiogram · Fuzzy logic · Particle filter · Mean shift

1 Introduction

Target tracking is a major foundation in visual surveillance, human-machine interface, vehicle navigation, video scene analysis, etc. Numerous methods have been reported [1], and these methods can be classified into two categories: using single-modality [2–5] and using multi-modality [6–10]. By leveraging the combined benefit of using different modalities while compensating for failure in individual modality, multi-modality system offers considerable advantage over single-modality system.

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 113–127, 2017.
https://doi.org/10.1007/978-981-10-7299-4_10

The fusion of different spectral images is significant in multi-modality tracking, because each spectral image provides disparate, yet complementary information about a scene. It offers an improved operational robustness, because distinct physical sensing principles compensate for particular perception shortcomings. Besides the fusion of multi-spectral images, the combination of multiple features from identical spectral image is also widely researched. The multi-feature fusion can better represent the target than single feature under complex dynamic circumstance by taking advantage of the complementarity of different features. However, above two types of multi-modality tracking method act in their own way, and few method can integrate them in a natural mean. In this paper, we propose a unified framework to cope with the problem by employing spatial histogram representation and mean shift search. Being fast and robust, the Mean Shift Tracker (MST) [5] has been used widely and many extensions and variants were proposed [11–14]. Instead of completely ignoring the spatial structure of the object features in original MST, Birchfield et al. [11] introduced a spatial histogram (abbreviated to *spatiogram*) that was formed by weighting each bin of histogram with the mean and covariance of the locations of the pixels that contribute to that bin. The spatiogram-based method not only improves the accuracy of the target appearance model but also preserves the rapidity of mean shift tracking, so we choose it to build our framework. The main contributions of our work are as follows:

- Establish a unified framework of multi-modal target tracking based on joint spatiogram representation, and the framework is flexible enough to handle any number of modalities;
- Design a fast multi-input multi-output fuzzy system to adaptively adjust the weight of each modality for evaluating the target state pre frame;
- Develop a particle filter based model update scheme to keep track of the most representative reference spatiogram throughout the tracking procedure;

The rest of paper is organized as follows. In Sect. 2, we first introduce the related work. Next, the spatiograms and their similarity is sketched in Sect. 3. After that, respectively detail the proposed tracking algorithm and fuzzy logic based weight adjust method in Sect. 4. Our model update scheme based particle filter is described in Sect. 5. Experimental results on the method are reported in Sect. 6. We conclude this paper in Sect. 7.

2 Related Work

The cooperation of visible and infrared imagery is the most active topic in multi-spectral combination. Infrared sensor can detect relative difference in the amount of thermal energy emitted from objects, and is independent of illumination, making it more effective than visible camera under poor lighting condition, but infrared imagery can't provide color and texture features. Visible sensor is oblivious to temperature difference, but is more effective than infrared sensor when objects are at thermal crossover, provided that the scene is well illuminated and

the objects have color signatures different from the background. Therefore, by combining their data, a tracker can perform better than one that uses only single sensor. There are many methods for tracking color-infrared target. Reference [6] proposed a framework by fusing the outputs of multiple spatiogram trackers. Reference [7] presented a multi-cue mean-shift tracking approach based on fuzzified region dynamic image fusion. Reference [8] proposed a tracking approach for visible-infrared target using tracking-before-fusion, in which the visible and infrared targets were tracked individually, and their results were fused. However, the final result may be very bad when any one of their tracking results is poor. Reference [9] proposed a compressive spatial-temporal Kalman fusion tracking algorithm that allowed independent features to be integrated, whose fusion model was formulated with both spatial and temporal fusion coefficients. However, its Bayesian classification requires a large number of samples to gain enough accuracy, which leads to large computational burden. Reference [10] proposed an infrared-visible target tracking method using joint sparse representation, in which only one target template was required for infrared or visible image. The method need solve only one ℓ_1 norm minimization problem, so its time cost is dramatically decreased. However, it is the single target template that leads to the loss of diversity of target templates, thus reducing the robustness of tracker.

In recent years, many multi-feature fusion tracking algorithms have been reported. Reference [15] proposed a fusion tracking method that employed local steering kernel descriptor and color histogram to represent the target object. Reference [16] proposed a new joint sparse representation model for robust feature-level fusion tracking, which dynamically prevented unreliable feature being fused by taking advantage of the sparse representation. Reference [17] developed a particle filter tracking algorithm based on the fusion of color histogram and edge orientation histogram. Reference [18] located the target through independent multi-feature fusion and region-based temporal difference model under particle filter framework. Reference [19] proposed a multi-feature fusion tracking framework by employing hashing method to fuse different features to generate compact binary feature. Reference [20] presented a novel online object tracking algorithm by using multi-feature channels with adaptive weights.

3 The Spatiograms and Their Similarity

Spatiogram [11] is generalization of histogram that includes potentially higher order moments. Conventional histogram is a zeroth-order spatiogram, while second-order spatiogram contains spatial mean and covariance for each histogram bin.

Denote by $\hat{\boldsymbol{h}} = \{\hat{p}_u, \hat{\boldsymbol{\mu}}_u, \widehat{\boldsymbol{\Sigma}}_u\}_{u=1}^m$ the m-bin reference second-order spatiogram of the target. In each frame of the image sequence, the center point $\boldsymbol{z}$ of the image region in which the spatiogram $\boldsymbol{h}(\boldsymbol{z}) = \{p_u(\boldsymbol{z}), \boldsymbol{\mu}_u(\boldsymbol{z}), \boldsymbol{\Sigma}_u(\boldsymbol{z})\}_{u=1}^m$ is closest to $\hat{\boldsymbol{h}}$ is sought. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ denote the (normalized) coordinates of the pixels in a candidate region centered at $\boldsymbol{z}$, then the feature probability of the uth bin is

$$p_u(\boldsymbol{z}) = \sum_{i=1}^{n} k(\|\boldsymbol{x}_i - \boldsymbol{z}\|^2)\delta[b(\boldsymbol{x}_i) - u], \tag{1}$$

where $b(\boldsymbol{x})$ is the bin number $(1, \cdots, m)$ associated with the feature at location $\boldsymbol{x}$, δ is the Kronecker delta function, and $k(x)$ is a kernel profile that assigns smaller weights to pixels farther from the circle center. Note that the kernel is normalized such that its profile satisfies $\sum_{i=1}^{n} k(\|\boldsymbol{x}_i - \boldsymbol{z}\|^2) = 1$. The coordinate mean $\boldsymbol{\mu}_u(\boldsymbol{z})$ and covariance $\boldsymbol{\Sigma}_u(\boldsymbol{z})$ of pixels belong to the bin u are

$$\boldsymbol{\mu}_u(\boldsymbol{z}) = \frac{1}{\sum_{k=1}^{n} \delta[b(\boldsymbol{x}_k) - u]} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{z})\delta[b(\boldsymbol{x}_i) - u], \tag{2}$$

$$\boldsymbol{\Sigma}_u(\boldsymbol{z}) = \frac{1}{\sum_{k=1}^{n} \delta[b(\boldsymbol{x}_k) - u] - 1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_u(\boldsymbol{z}))(\boldsymbol{x}_i - \boldsymbol{\mu}_u(\boldsymbol{z}))^{\mathrm{T}} \delta[b(\boldsymbol{x}_i) - u], \tag{3}$$

Similarly, the calculation of reference spatiogram $\hat{\boldsymbol{h}}$ can be regarded as a special case of $\boldsymbol{h}(\boldsymbol{z})$ that $\boldsymbol{z} = 0$.

The similarity between the spatiograms is measured by

$$\rho(\boldsymbol{z}) \triangleq \rho[\boldsymbol{h}(\boldsymbol{z}), \hat{\boldsymbol{h}}] = \sum_{u=1}^{m} \psi_u(\boldsymbol{z}) \sqrt{p_u(\boldsymbol{z})\hat{p}_u}, \tag{4}$$

where $\sqrt{p_u(\boldsymbol{z})\hat{p}_u}$ is used to measure the similarly between features of candidate region and target image, while

$$\psi_u(\boldsymbol{z}) = \frac{4|\boldsymbol{\Sigma}_u(\boldsymbol{z})\widehat{\boldsymbol{\Sigma}}_u|^{\frac{1}{4}}}{|\widetilde{\boldsymbol{\Sigma}}_u|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_u(\boldsymbol{z}) - \hat{\boldsymbol{\mu}}_u)^{\mathrm{T}}(\widetilde{\boldsymbol{\Sigma}}_u(\boldsymbol{z}))^{-1}(\boldsymbol{\mu}_u(\boldsymbol{z}) - \hat{\boldsymbol{\mu}}_u)\right\} \tag{5}$$

is used to measure the similarity in spatial arrangement between these features, and $\widetilde{\boldsymbol{\Sigma}}_u(\boldsymbol{z}) = 2(\boldsymbol{\Sigma}_u(\boldsymbol{z}) + \widehat{\boldsymbol{\Sigma}}_u)$.

4 Proposed Multi-modality Tracker

Assume there are N different modalities (or N reference spatiograms), then it is a fact that each candidate state should correspond to N image patches with different modalities. That is to say, at any time instant t, all well registered N image patches with different modalities should share same dynamic state (including location, size and shape), because they merely use different modalities to describe identical real object.

4.1 Joint Spatiogram Representation

Denote by $\hat{\boldsymbol{h}}_j = \{\hat{p}_u^j, \hat{\boldsymbol{\mu}}_u^j, \widehat{\boldsymbol{\Sigma}}_u^j\}_{u=1}^m$ the jth modality reference spatiogram. Given a candidate state centered at $\boldsymbol{z}$, then there are N candidate spatiograms with different modalities, and we denote by $\boldsymbol{h}_j = \{p_u^j(\boldsymbol{z}), \boldsymbol{\mu}_u^j(\boldsymbol{z}), \boldsymbol{\Sigma}_u^j(\boldsymbol{z})\}_{u=1}^m$ the jth

modality candidate spatiogram. The similarity between the jth modality candidate spatiogram and its reference spatiogram is measured by

$$\rho_j(\boldsymbol{z}) = \sum_{u=1}^{m} \psi_u^j(\boldsymbol{z}) \sqrt{p_u^j(\boldsymbol{z}) \hat{p}_u^j} \tag{6}$$

where

$$\psi_u^j(\boldsymbol{z}) = \frac{4|\boldsymbol{\Sigma}_u^j(\boldsymbol{z}) \widehat{\boldsymbol{\Sigma}}_u^j|^{\frac{1}{4}}}{|\widetilde{\boldsymbol{\Sigma}}_u^j|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\mu}_u^j(\boldsymbol{z}) - \hat{\boldsymbol{\mu}}_u^j)^{\mathrm{T}} (\widetilde{\boldsymbol{\Sigma}}_u^j(\boldsymbol{z}))^{-1} (\boldsymbol{\mu}_u^j(\boldsymbol{z}) - \hat{\boldsymbol{\mu}}_u^j) \right\}, \tag{7}$$

and $\widetilde{\boldsymbol{\Sigma}}_u^j(\boldsymbol{z}) = 2(\boldsymbol{\Sigma}_u^j(\boldsymbol{z}) + \widehat{\boldsymbol{\Sigma}}_u^j)$.

For the multi-modality fusion tracking, whether a candidate state should be accepted or not is decided by joint similarity of all modality candidates and their corresponding targets, so we define the joint similarity (or object function) as

$$\rho(\boldsymbol{z}) \triangleq \sum_{j=1}^{N} \alpha_j \rho_j(\boldsymbol{z}), \tag{8}$$

where $0 \leq \alpha_j \leq 1$ is the weight that reflects the reliability of the jth modality to evaluating the target state in current frame, and $\sum_{j=1}^{N} \alpha_j = 1$.

4.2 Target Localization

The goal of target localization is to estimate the target translation $\hat{\boldsymbol{z}}$ that maximizes the joint similarity in (8). Denote by $\hat{\boldsymbol{z}}_0$ the estimated target location in the previous frame. Approximating the joint similarity (8) in the current frame by its first-order Taylor expansion around the values $p_u^j(\hat{\boldsymbol{z}}_0)$ and $\boldsymbol{\mu}_u^j(\hat{\boldsymbol{z}}_0)$ results in

$$\begin{aligned} \rho(\boldsymbol{z}) \approx & \sum_{j=1}^{N} \sum_{u=1}^{m} \alpha_j \psi_u^j(\hat{\boldsymbol{z}}_0) \sqrt{p_u^j(\hat{\boldsymbol{z}}_0) \hat{p}_u^j} ((\widetilde{\boldsymbol{\Sigma}}_u^j(\hat{\boldsymbol{z}}_0))^{-1} (\hat{\boldsymbol{\mu}}_u^j - \boldsymbol{\mu}_u^j(\hat{\boldsymbol{z}}_0)) \boldsymbol{\mu}_u^j(\boldsymbol{z}) + \cdots \\ & \sum_{j=1}^{N} \sum_{u=1}^{m} \frac{\alpha_j}{2} \psi_u^j(\hat{\boldsymbol{z}}_0) \sqrt{\hat{p}_u^j / p_u^j(\hat{\boldsymbol{z}}_0)} p_u^j(\boldsymbol{z}) + C, \end{aligned} \tag{9}$$

where

$$p_u^j(\boldsymbol{z}) = \sum_{i=1}^{n} k(\|\boldsymbol{x}_i^j - \boldsymbol{z}\|^2) \delta[b(\boldsymbol{x}_i^j) - u], \tag{10}$$

$$\boldsymbol{\mu}_u^j(\boldsymbol{z}) = \frac{1}{\sum_{k=1}^{n} \delta[b(\boldsymbol{x}_k^j) - u]} \sum_{i=1}^{n} (\boldsymbol{x}_i^j - \boldsymbol{z}) \delta[b(\boldsymbol{x}_i^j) - u], \tag{11}$$

and C is independent of $\boldsymbol{z}$. Taking the derivative of (9) with respect to $\boldsymbol{z}$ yields

$$\frac{\partial \rho(\boldsymbol{z})}{\partial \boldsymbol{z}} = \sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j k'(\|\boldsymbol{z} - \boldsymbol{x}_i^j\|^2)(\boldsymbol{z} - \boldsymbol{x}_i^j) - \sum_{j=1}^{N} \sum_{u=1}^{m} \boldsymbol{w}_u^j, \tag{12}$$

where

$$\left.\begin{aligned} w_i^j &= \sum_{u=1}^{m} \frac{\alpha_j}{2} \psi_u^j(\hat{\boldsymbol{z}}_0) \sqrt{\hat{p}_u^j / p_u^j(\hat{\boldsymbol{z}}_0)}\, \delta[b(\boldsymbol{x}_i^j) - u] \\ \boldsymbol{w}_u^j &= \alpha_j \psi_u^j(\hat{\boldsymbol{z}}_0) \sqrt{p_u^j(\hat{\boldsymbol{z}}_0) \hat{p}_u^j}\, ((\widetilde{\boldsymbol{\Sigma}}_u^j(\hat{\boldsymbol{z}}_0))^{-1} (\hat{\boldsymbol{\mu}}_u^j - \boldsymbol{\mu}_u^j(\hat{\boldsymbol{z}}_0)) \end{aligned}\right\} \tag{13}$$

Set $\frac{\partial \rho(\boldsymbol{z})}{\partial \boldsymbol{z}} = 0$ and solve for $\boldsymbol{z}$:

$$\hat{\boldsymbol{z}} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j g(\|\boldsymbol{z}_0 - \boldsymbol{x}_i^j\|^2) \boldsymbol{x}_i^j - \sum_{j=1}^{N} \sum_{u=1}^{m} \boldsymbol{w}_u^j}{\sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j g(\|\boldsymbol{z}_0 - \boldsymbol{x}_i^j\|^2)}, \tag{14}$$

where $g(x) = -k'(x)$. If we use the Epanechnikov profile [5] then the derivative of the kernel is constant, thus the iteration (14) is reduced to

$$\hat{\boldsymbol{z}} = \left(\sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j \boldsymbol{x}_i^j - \sum_{j=1}^{N} \sum_{u=1}^{m} \boldsymbol{w}_u^j \right) \Big/ \sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j, \tag{15}$$

To enable the physical meaning of (15) to evident, we rearrange the numerator to obtain

$$\hat{\boldsymbol{z}} = \sum_{j=1}^{N} \sum_{i=1}^{n} (w_i^j \boldsymbol{x}_i^j - \boldsymbol{v}_i^j) \Big/ \sum_{j=1}^{N} \sum_{i=1}^{n} w_i^j, \tag{16}$$

where $\boldsymbol{v}_i^j = \sum_{u=1}^{m} (\boldsymbol{w}_u^j \delta[b(\boldsymbol{x}_i^j) - u] / \sum_{k=1}^{n} \delta[b(\boldsymbol{x}_k^j) - u])$. It is easy to see from (16) that each pixel in each modality candidate patch casts a vote proportional to $\|w_i^j \boldsymbol{x}_i^j - \boldsymbol{v}_i^j\|$ in the direction of $\boldsymbol{x}_i^j - \boldsymbol{v}_i^j / w_i^j$ for the Mean-Shift offset toward or away from $\hat{\boldsymbol{z}}$, which also shows that the optimal location of the target is determined by all modalities.

4.3 Adjusting Weights Using Fuzzy Logic

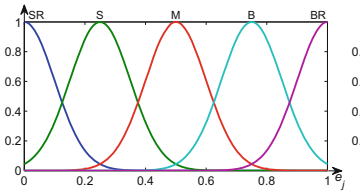
The weights in most algorithms are assumed to be unchanged during the tracking, but the fact is that the importance (reliability) of each modality changes over time. Usually, the target doesn't change drastically between consecutive frames, but has always a few difference, so we can only fuzzily rather than determinately estimate the change. In this paper, we use the similarity between the tracking result and the reference model to measure the change, and employ the fuzzy logic method to calculate the weights. Specifically, the inputs of fuzzy system are the reliabilities

$$e_j(\hat{\boldsymbol{z}}) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{1 - \rho_j(\hat{\boldsymbol{z}})}{2\delta^2} \right) \tag{17}$$

index $j = 1, \ldots, N$, where $\rho_j(\hat{\boldsymbol{z}})$ is the similarity between the tracking result and the reference model in current frame. Notice that the smaller the similarity,

the lower the reliability. The outputs of fuzzy system are the weights α_j index $j = 1,\ldots,N$ in the next frame.

In this study, we apply the singleton fuzzification, product inference, and centroid defuzzification to build the fuzzy system [21]. Each input variable e_j is fuzzified with five linguistic variables, labeled SR, S, M, B, BR, partitioned on the interval [0, 1], and each output variable α_j is also fuzzified with nine linguistic variables, labeled ST, VS, SR, S, M, B, BR, VB, BT, partitioned on the interval [0, 1], where ST stands for smallest, VS for very smaller, SR for smaller, S for small, M for middle, B for big, BR for bigger, VB for very bigger, BT for biggest. The membership functions of e_j and α_j are all Gaussian function, as shown in Fig. 1(a) and (b).



(a) (b)

e_i \ α_j	SR	S	M	B	BR
SR	M	B	BR	VB	BT
S	S	M	B	BR	VB
M	SR	S	M	B	BR
B	VS	SR	S	M	B
BR	ST	VS	SR	S	M

(c)

Fig. 1. (a) Membership function for e_j (b) membership function for α_j (c) the control rule bases for a dual-input and single-output fuzzy system.

Our fuzzy system is a typical multi-input multi-output system with N inputs and N outputs, and if directly using the IF-THEN rule of N inputs inferring one output then resulting in $N \times 5^N$ rules, which is time-consuming. Therefore, we convert the fuzzy system into several dual-input and single-output (DISO) fuzzy systems, and define the fuzzy rule of each DISO system as

$$\text{IF } e_j \text{ is } A_1^k \text{ and } e_i \text{ is } A_2^k \text{ THEN } \alpha_j^i \text{ is } B^k, k = 1,\ldots,K \tag{18}$$

where A_1^k and A_2^k belong to {SR, S, M, B, BR}, B^k belongs to {ST, VS, SR, S, M, B, BR, VB, BT}, α_j^i $(i \neq j)$ is the weight of modality j relative to modality i, and K is the total number of rules. The control rule bases originating from (18) is shown in Fig. 1(c), apparently, $K = 25$ and $\alpha_i^j = 1 - \alpha_j^i$. After performing the fuzzy control procedure to acquire all α_j^i, the α_j is calculated by

$$\alpha_j = (\alpha_j^1 + \cdots + \alpha_j^{(j-1)} + \alpha_j^{(j+1)} + \cdots + \alpha_j^N)/(N-1) \tag{19}$$

Finally, all α_j are normalized such that they satisfy $\sum_j \alpha_j = 1$. Since the derivation of each α_j needs $N-1$ noninteractive DISO fuzzy systems, the total number of DISO systems is $N(N-1)/2$, thus the total fuzzy rules can be reduced to $25 \times N(N-1)/2$.

Algorithm 1 summarizes the proposed Multi-modality Joint Spatiogram Tracking (MJST) procedure, which is implemented in MATLAB+MEX. The model update will be detailed in the next section.

Algorithm 1. Multi-modality fusion tracking via spatiogram and fuzzy logic

Input: each modal target image $\{I(\boldsymbol{x}_i^j, 1)\}_{i=1}^n$ centered at $\boldsymbol{z}_1$ in the first frame, the corresponding reference spatiogram $\{\hat{p}_u^j(1), \hat{\boldsymbol{\mu}}_u^j(1), \widehat{\boldsymbol{\Sigma}}_u^j(1)\}_{u=1}^m$, its weight α_j, and γ;

1: **for** $t = 2, \ldots, T$ **do**
2: Compute the current candidate spatiogram $\{p_u^j(\boldsymbol{z}_{t-1}), \boldsymbol{\mu}_u^j(\boldsymbol{z}_{t-1}), \boldsymbol{\Sigma}_u^j(\boldsymbol{z}_{t-1})\}_{u=1}^m$ by (1) to (3) and its similarity $\rho_j(\boldsymbol{z}_{t-1})$ with its reference spatiogram by (6) to (7), as well as their $\rho(\boldsymbol{z}_{t-1})$ by (8);
3: Compute each weights w_i^j and $\boldsymbol{w}_u^j$ by (13), and find new candidate location $\boldsymbol{z}_t$ by (15);
4: Compute the new candidate spatiogram $\{p_u^j(\boldsymbol{z}_t), \boldsymbol{\mu}_u^j(\boldsymbol{z}_t), \boldsymbol{\Sigma}_u^j(\boldsymbol{z}_t)\}_{u=1}^m$ by (1) to (3) and its similarity $\rho_j(\boldsymbol{z}_t)$ with its reference spatiogram by (6) to (7), and their $\rho(\boldsymbol{z}_t)$ by (8);
5: If $\rho(\boldsymbol{z}_t) < \rho(\boldsymbol{z}_{t-1})$, then set $\boldsymbol{z}_t \leftarrow \frac{1}{2}(\boldsymbol{z}_{t-1} + \boldsymbol{z}_t)$ and go to Step 4;
6: If $\boldsymbol{z}_t = \boldsymbol{z}_{t-1}$ or the number of iterations reached M_{max}, then go to Step 7. Otherwise, set $\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{z}_t$ and $\rho(\boldsymbol{z}_{t-1}) \leftarrow \rho(\boldsymbol{z}_t)$, and go to Step 3;
7: Compute all reliability $e_j(\boldsymbol{z}_t)$ by (17) and adjust all weight α_j using fuzzy logic method;
8: For each modality, establish the particle filter with the state variable $\{I(\boldsymbol{x}_i^j, t-1)\}_{i=1}^n$ and its current observation $\boldsymbol{h}_j(\boldsymbol{z}_t)$ by (20) to (22), run the filter procedure to gain optimal state $\{\hat{I}(\boldsymbol{x}_i^j, t)\}_{i=1}^n$, and obtain the new reference spatiogram $\{\hat{p}_u^j(t), \hat{\boldsymbol{\mu}}_u^j(t), \widehat{\boldsymbol{\Sigma}}_u^j(t)\}_{u=1}^m$ according to (23);
9: **end for**

5 Model Update via Particle Filter

In practical tracking scenarios, the target model should be updated so that it can capture the appearance variations due to illumination or pose changes. Some self-learning methods use the weighted sum of the current tracking result and current model to get new model, which is easy to realize but also prone to drift from the target because of the accumulation of errors. Other semi-supervised learning methods are suggested to avoid the drift problem, but the classifier-based tracking cannot be applicable to our case. Motivated by the Ref. [12], we propose a model update mechanism based on particle filter, and its details are introduced as follows.

5.1 Spatiogram Filtering

In most of tracking methods, the particle filter keeps tracking of object changes in position and velocity, not the changes of object appearance. We use particle filter for filtering object spatiogram so as to obtain the optimal estimate of the target model.

In frame t, let $I(\boldsymbol{x}_i, t)$ be the feature value (e.g., intensity, gradient, texture, etc.) at the normalized coordinate $\boldsymbol{x}_i$ in object image, and $\boldsymbol{h}(\{I(\boldsymbol{x}_i, t)\}_{i=1}^n)$ be the corresponding spatiogram that is obtained by performing (1) to (3) on $\{I(\boldsymbol{x}_i, t)\}_{i=1}^n$. The object appearance variations in essence are the value changes of all $I(\boldsymbol{x}_i, t)$. Since the change of each $I(\boldsymbol{x}_i, t)$ is independent, we can filter each of them independently. Thus, the state prediction equation of each $I(\boldsymbol{x}_i, t)$ is given by

$$I(\boldsymbol{x}_i, t) = I(\boldsymbol{x}_i, t-1) + \omega_i(t-1), \qquad i = 1, 2, \cdots, n, \tag{20}$$

where $I(\boldsymbol{x}_i, t)$ is called the state of pixel $\boldsymbol{x}_i$ in the frame t, and $I(\boldsymbol{x}_i, t-1)$ is its counterpart in the frame $t-1$. $\omega_i(t-1)$ specifies the state noise owing to the object appearance variation, which is assumed to be Gaussian, and furthermore, to have the same variance σ_ω^2 for all $\boldsymbol{x}_i$.

To obtain measurements of the filters, the previous reference spatiogram $\boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n)$ is matched with the current frame, and yielding a new spatiogram $\boldsymbol{h}(\boldsymbol{z}_t)$ at the convergence position $\boldsymbol{z}_t$ provided by the mean-shift tracking algorithm. Then, $\boldsymbol{h}(\boldsymbol{z}_t)$ is used as measurement for $\boldsymbol{h}(\{I(\boldsymbol{x}_i, t)\}_{i=1}^n)$, and the observation equation is

$$\boldsymbol{h}(\boldsymbol{z}_t) = \boldsymbol{h}(\{I(\boldsymbol{x}_i, t)\}_{i=1}^n) + \boldsymbol{\upsilon}(t), \tag{21}$$

where $\boldsymbol{\upsilon}(t)$ models the noise in the image signal. Since each bin in the spatiogram is independent, the (21) can be rewritten as

$$h_u(\boldsymbol{z}_t) = h_u(t) + \boldsymbol{\upsilon}_u(t), \qquad u = 1, 2, \cdots, m, \tag{22}$$

where $h_u(\boldsymbol{z}_t) = [p_u(\boldsymbol{z}_t), \boldsymbol{\mu}_u(\boldsymbol{z}_t), \boldsymbol{\Sigma}_u(\boldsymbol{z}_t)]^{\mathrm{T}}$, $\boldsymbol{h}(\boldsymbol{z}_t) = \{h_u(\boldsymbol{z}_t)\}_{u=1}^m$, $h_u(t) = [\hat{p}_u(t), \hat{\boldsymbol{\mu}}_u(t), \widehat{\boldsymbol{\Sigma}}_u(t)]^{\mathrm{T}}$, $\boldsymbol{h}(\{I(\boldsymbol{x}_i, t)\}_{i=1}^n) = \{h_u(t)\}_{u=1}^m$, and $\boldsymbol{\upsilon}_u(t)$ is the observation noise, which is assumed to be Gaussian, and furthermore, to have the same variance σ_υ^2 for all u. Because the $\boldsymbol{\Sigma}_u$ is 2-D diagonal matrix, we only update its two diagonal entries, thus the h_u is a 5-D vector.

We call $\boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n)$ *current model*, and $\boldsymbol{h}(\boldsymbol{z}_t)$ *observation model*. Particle filter then yields a tradeoff between the two models and provides us an optimal estimate of the object model that is called *candidate model*, denoted by $\boldsymbol{h}(\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n)$, where $\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n$ is the optimal particle. In implement, the $\{I(\boldsymbol{x}_i, t)\}_{i=1}^n$ is treated as a particle whose likelihood weight is calculated by $\frac{1}{\sqrt{2\pi}\sigma_\upsilon} \exp\{-\frac{1-\rho[\boldsymbol{h}(\boldsymbol{z}_t), \boldsymbol{h}(\{I(\boldsymbol{x}_i, t)\}_{i=1}^n)]}{\sigma_\upsilon^2}\}$ incorporating with (4), and the $\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n$ is determined by taking the weighted mean of all particles.

5.2 Update Criterion

It is not suitable for us to accept the candidate model all the time. We should try to find a robust criterion to decide whether the candidate mode $\boldsymbol{h}(\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n)$ should be accepted because over-update could make tracker sensitive to outliers like occlusions or dramatic appearance changes.

We take the similarity $\rho[\boldsymbol{h}(\boldsymbol{z}_t), \boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n)\})]$ between the observation model and the current model as the criterion. If the similarity is smaller than the threshold γ, which implies that the object encounters dramatic appearance changes, then we reject $\boldsymbol{h}(\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n)$ and remain using current model, otherwise accept it. The final model update formula is as follows

$$\hat{\boldsymbol{h}}(t) = \begin{cases} \boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n\}) & \rho[\boldsymbol{h}(\boldsymbol{z}_t), \boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n\})] < \gamma \\ \boldsymbol{h}(\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n) & \rho[\boldsymbol{h}(\boldsymbol{z}_t), \boldsymbol{h}(\{I(\boldsymbol{x}_i, t-1)\}_{i=1}^n\})] \geq \gamma \end{cases} \tag{23}$$

Once $\boldsymbol{h}(\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n)$ is accepted, the $\{\hat{I}(\boldsymbol{x}_i, t)\}_{i=1}^n$ is also saved.

6 Experiments

To verify the flexibility, accuracy and efficiency of the proposed tracker, we tested six registered infrared-visible sequences which involve general difficulties like night, shade, cluster, crossover and occlusion. We also compared the proposed tracker to the state-of-the-art methods such as ℓ_1 tracker (L1T) [4], joint sparse representation tracker (JSRT) [10], and fuzzified region dynamic fusion tracker (FRD) [7]. All tracking results are obtained by running these trackers on an Intel Dual-Core 2.6 GHz CPU with 8 GB RAM, using the same initial positions for fair comparison. The L1T can tackle only one modality at a time.

Set $M_{max} = 20$, $\gamma = 0.4$, and the number of particles 50 for model update. We use the combination of infrared gray, infrared HOG (histogram of oriented gradient), visible gray, and visible LBP (local binary pattern), so $N = 4$.

6.1 Combination of Infrared Gray-HOG and Visible Gray-LBP

The LBP [22] is very effective to describe the image texture features, and has advantages such as fast computation and rotation invariance, which facilitates the wide usage in the fields of texture analysis, image retrieval, object tracking, etc. In LBP, each pixel is assigned a texture value, which can be naturally combined with the gray of the pixel to represent targets. The LBP operator labels the pixel in an image by thresholding its neighborhood with the center value and considering the result as a binary number.

The HOG technique [23] counts occurrences of gradient orientation in localized portions of an image, and is very effective to describe the object shape feature. HOG is invariant to geometric and photometric transformations, so is used widely in object detection. See [23] for the extraction steps of HOG. To enable HOG to naturally combine with the gray of image, we employ the visualization method suggested in [24].

Generally, the visible image contains more texture features, whereas shape feature is more obvious in the infrared image, so we extract LBP feature from visible image and HOG feature from infrared image. Figure 2 shows the extraction of LBP and HOG features from visible and infrared images respectively.

Figure 3 shows the screenshots of some sampled tracking results on video1-video6.

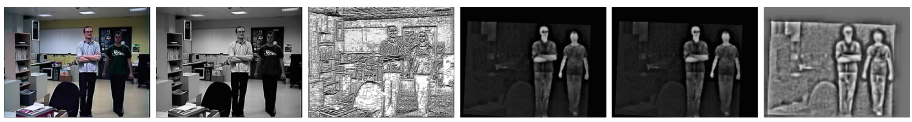


Fig. 2. Extracting LBP and HOG features from visual and infrared images respectively, (a) visible image, (b) visible gray, (c) visible LBP, (d) infrared image, (e) infrared gray, (f) infrared HOG.

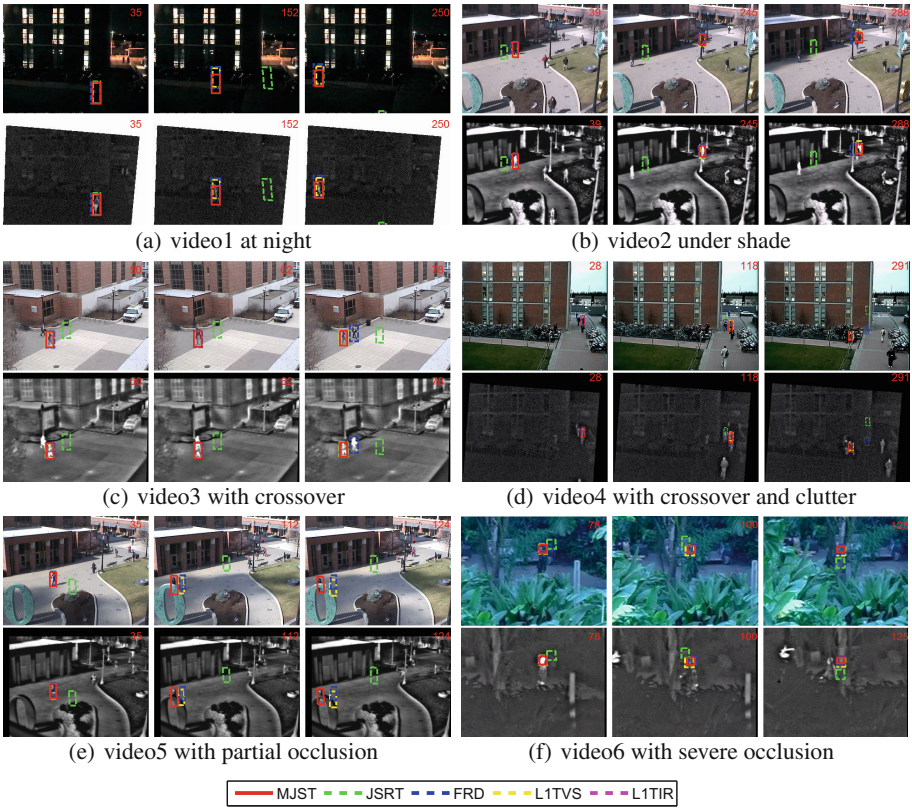


Fig. 3. Screenshots of some sampled tracking results, where L1TIR represents using L1T to track infrared target, and L1TVS represents using L1T to track visible target.

Night and Shade: At night without light, the target in visible image is usually obscure (see Fig. 3(a)), so the tracker that only using visible camera often fails to hold the target. However, by using or cooperating with infrared camera, our tracker can hold well the target, because the infrared sensor is independent of illumination. In Fig. 3(b), the man in black is walking into a shade area, and the contrast between him and his background is very low in visible image but no for infrared image, which is similar to walking at night, so our tracker can hold the target well. It is notable that our tracker successfully overcome the occlusion of lamppost (see frame 245 and 288 in Fig. 3(b)), which is mainly attribute to the use of online model update.

Crossover and Clutter: The tracker is easily attracted by the distracter when target is at thermal crossover or in cluttered background. For example, the FRD is attracted by the right man when two men are at thermal crossover (see frame 78 in Fig. 3(c)), and FRD and L1TVS are attracted by the cluttered background

(see frame 118 and 291 in Fig. 3(d)). The distracter presents similar appearance as the target, so can give good match to the target, which makes it be difficult to discriminate such a distracter only using single source or matching score. However, our tracker can stick with the target all the time because of using joint compressive representation of infrared and visual modalities, thus can always capture the target well.

Partial and Full Occlusion: Occlusion is the most general yet crucial problem in object tracking, and it is classified into partial and full. For partial occlusion, our tracker can successfully overcome it (see frame 245 in Fig. 3(b), frame 112 in Fig. 3(e), and frame 100 in Fig. 3(f)), because both it use model update to handle occlusion. For full occlusion, as showed in Figs. 3(f) and 4(f), almost all trackers lost the targets, which is because that the targets have completely disappeared from our sight.

6.2 Quantitative Comparison

We use five criteria, i.e., center offset error $\varrho = \sqrt{(x_G - x_T)^2 + (y_G - y_T)^2}$, average center offset error $\bar{\varrho} = \frac{1}{q}\sum_i \varrho_i$, overlap ratio $\epsilon = \frac{area(R_G \bigcap R_T)}{area(R_G \bigcup R_T)}$, average overlap ratio $\bar{\epsilon} = \frac{1}{q}\sum_i \epsilon_i$, and success rate $sr = \frac{1}{q}\sum_i f(\epsilon_i - 0.5)$, where (x_G, y_G, R_G) is the center and region of target given by manual, (x_T, y_T, R_T) is the result given by the tacker, and q is the number of frames. The $f(x)$ is step function, if $x \geq 0$, then $f(x) = 1$, else $f(x) = 0$. The overlap ratio is used to evaluate the accuracy of the tracking result pre frame, and whose value is 1 when the estimated region overlaps fully with the ground truth, thereby obtaining best tracking result. An ideal tracker should have less center offset error, and high overlap ratio and success rate.

The L1T can track only one of visible and infrared at a time, for easy to compare with other fusion trackers, we integrate the tracking results of L1TVS and L1TIR as following: (1) the time consumer of L1T is equal to the time sum of L1TVS and L1TIR, and (2) the target state of L1T is equal to the weight average of the target states of L1TVS and L1TIR, where the weight is induced by

Table 1. Quantitative comparison of MJST, JSRT, FRD and L1T.

	MJST			JSRT			FRD			L1T		
	$\bar{\varrho}$	$\bar{\epsilon}$	sr	$\bar{\varrho}$	$\bar{\epsilon}$	sr	$\bar{\varrho}$	$\bar{\epsilon}$	sr	$\bar{\varrho}$	$\bar{\epsilon}$	sr
video1	4.28	0.81	1.00	83.3	0.19	0.21	6.97	0.75	0.93	36.4	0.32	0.36
video2	2.31	0.77	0.98	69.5	0.07	0.07	5.33	0.64	0.77	35.3	0.07	0.06
video3	1.47	0.84	1.00	17.6	0.46	0.51	2.75	0.84	0.95	2.64	0.78	0.90
video4	5.44	0.75	0.93	53.5	0.30	0.38	20.9	0.27	0.36	7.46	0.70	0.87
video5	19.1	0.65	0.78	79.6	0.12	0.15	24.0	0.42	0.59	20.6	0.56	0.68
video6	9.79	0.62	0.73	16.7	0.43	0.51	3.63	0.78	0.85	23.5	0.53	0.61
averag	7.07	0.74	0.90	53.4	0.26	0.31	10.6	0.62	0.74	21.0	0.49	0.58

their observation likelihood in a mean square error way. The quantitative results are summarized in Table 1, where red font indicates the best performance while the blue font indicates the second best ones. The detail of overlap ratio pre frame is shown in Fig. 4. It can be observed that, compared with other trackers, our tracker obtains the highest tracking accuracy and success rate.

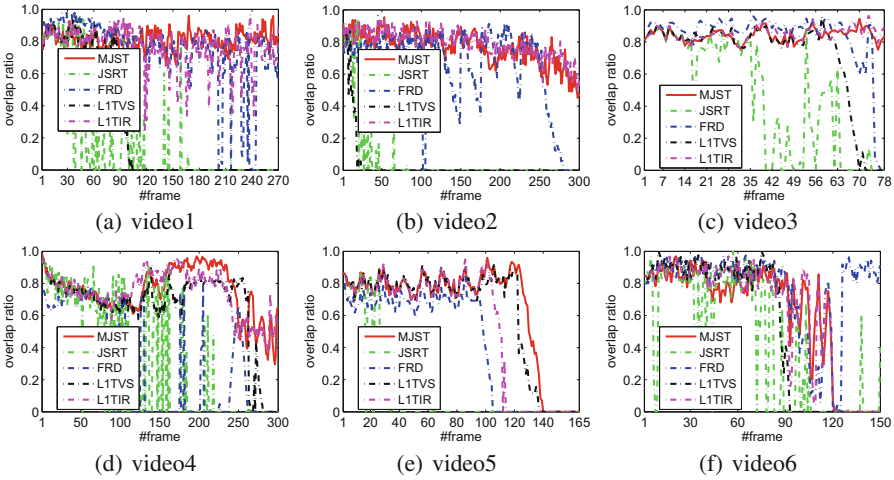


Fig. 4. The overlap ratio between tracked region and manually marked ground truth

7 Conclusions

In this paper we had proposed a unified multi-modality tracking framework by cooperating fuzzy logic. In the framework, different types of modalities can be arbitrary added, removed and naturally integrated through joint spatiogram representation. In addition, the target model update strategy enables the tracker promptly respond to the variations of appearance. Experiment results on six datasets demonstrate that our approach performs best. Our current tracker is not quite suitable for tracking full occlusion targets.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61365009, 61663004, 61462008), the Natural Science Foundation of Guangxi Province of China (Grant Nos. 2014GXNSFAA118368, 2016GXNSFAA380146, 2013GXNSFAA019336), the Doctoral Research Foundation of Guangxi Normal University, and Guangxi Collaborative Innovation Center of Multisource Information Integration and Intelligent Processing. The authors would like to thank the anonymous reviewers for their constructive comments.

References

1. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 2411–2418 (2013)
2. Zhang, C.L., Tang, Y.P., Li, Z.X., et al.: Dual-kernel tracking approach based on second-order spatiogram. J. Electron. Inf. Technol. **37**(7), 1660–1666 (2015)
3. Henriques, J.F., Caseiro, R., Martins, P., et al.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **37**(3), 583–596 (2014)
4. Mei, X., Ling, H.: Robust visual tracking using l1 minimization. In: The 12th International Conference on Computer Vision, pp. 1436–1443 (2009)
5. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **25**(5), 564–577 (2003)
6. Ciarno, C., Noel, E.O., Alan, S.: Thermo-visual feature fusion for object tracking using multiple spatiogram. Mach. Vis. Appl. **19**(5), 483–494 (2008)
7. Xiao, G., Yun, X., Wu, J.M.: A multi-cue mean-shift target tracking approach based on fuzzified region dynamic image fusion. Sci. China Inf. Sci. **55**(3), 577–589 (2012)
8. Xiao, G., Yun, X., Wu, J.M.: A new tracking approach for visible and infrared sequences based on tracking-before-fusion. Int. J. Dyn. Control **4**(1), 40–51 (2016)
9. Yun, X., Jing, Z.L., Xiao, G., et al.: A compressive tracking based on time-space Kalman fusion model. Sci. China: Inf. Sci. **1**, 1–15 (2016)
10. Liu, H.P., Sun, F.C.: Fusion tracking in color and infrared images using joint sparse representation. Sci. China: Inf. Sci. **55**(3), 590–599 (2012)
11. Birchfield S.T.: Spatiograms versus histograms for region-based tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, SanDiego, California, USA, pp. 1158–1163 (2005)
12. Peng, N.S., Yang, J., et al.: Model update mechanism for mean-shift tracking. J. Syst. Eng. Electron. **16**(1), 52–57 (2005)
13. Leichter, I.: Mean shift trackers with cross-bin metrics. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 695–706 (2012)
14. Tomas, V., Jana, N., Jiri, M.: Robust scale-adaptive mean-shift for tracking. Pattern Recogn. Lett. **49**(1), 250–258 (2014)
15. Zoidi, O., Tefas, A., Pitas, I.: Visual object tracking based on local steering kernels and color histograms. IEEE Trans. Circ. Syst. Video Technol. **23**(5), 870–882 (2013)
16. Lan, X., Ma, A.J., Yuen P.C.: Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1194–1201 (2014)
17. Xu, F., Zhao, L.: A particle filter tracking algorithm based on adaptive feature fusion strategy. In: Proceedings of IEEE World Congress on Intelligent Control and Automation (WCICA), pp. 4612–4616 (2012)
18. Lu, X., Song, L., Yu, S., et al.: Object contour tracking using multi-feature fusion based particle filter. In: Proceedings of IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 237–242 (2012)
19. Ma, C., Liu, C.C., Peng, F.R., et al.: Multi-feature hashing tracking. Pattern Recogn. Lett. **69**(1), 62–71 (2016)
20. Jiang, H.L., Li, J.H., Wang, D., et al.: Multi-feature tracking via adaptive weights. Neurocomputing **207**, 189–201 (2016)

21. Ross, T.J.: Fuzzy Logic with Engineering Applications, 3rd edn. McGraw-Hill Inc., New York (1995)
22. Timo, O., Matti, P., Topi, M.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
23. Navneet, D., Bill, T.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition, vol. 1, no. 12, pp. 886–893 (2005)
24. Vondrick, C., Khosla, A., Pirsiavash, H., et al.: Visualizing object detection features. Int. J. Comput. Vis. **119**(2), 145–158 (2016)