Opt Rev (2016) 23:244–253
DOI 10.1007/s10043-015-0175-5

REGULAR PAPER

Visible and infrared tracking based on multi-view multi-kernel fusion model

Xiao Yun[1] · Zhongliang Jing[1] · Bo Jin[2]

Received: 29 June 2015 / Accepted: 26 December 2015 / Published online: 12 January 2016

Abstract In the visual tracking problem, fusion of visible and infrared sensors provides complementarily useful features and can consistently help distinguish the target from the background efficiently. Recently, multi-view learning has received growing attention due to its enormous potential in combining diverse view features containing consistent and complementary characteristics. Therefore, in this paper, a visible and infrared fusion tracking algorithm based on multi-view multi-kernel fusion (MVMKF) model is presented. The proposed MVMKF model considers the diversities of visible and infrared views and embeds complementary information from them. Furthermore, the multi-kernel framework is used to learn the importance of view features so that an integrated appearance representation is made with regard to the respective performance. Besides, the tracking task is completed with naive Bayes classifier in sophisticated compressive feature domain, considering the high performances of classifier-level and sophisticated feature-level learning for multiple views. The experimental results demonstrate that the MVMKF tracking algorithm performs well in terms of accuracy and robustness.

Keywords Visual tracking · Visible and infrared fusion tracking · Multi-sensor fusion · Multi-view learning

✉ Xiao Yun
yunxiao@sjtu.edu.cn

[1] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China

[2] Software Engineering Institute, East China Normal University, Shanghai 200062, China

1 Introduction

Multi-view learning has attracted much attention recently in the fields of image classification, word embedding, and food recognition [7, 15, 16]. Multiple views may be different viewpoints or descriptions from different features. In conventional machine learning algorithms, features in different views are concatenated into one single view, which is not physically meaningful because of the specific statistical property of each view. Instead, multi-view learning considers the diversity of multiple views and introduces an integrated model to learn them jointly [30]. Existing algorithms on multi-view learning can be grouped into two major categories: feature space-level and classifier-level learning [15]. A direct way to solve multi-view learning is to combine information from multiple views in the feature level. Meltzer et al. [17] proposed a method to learn feature descriptors using multi-view images. Each feature in the appearance model is learned by kernel principal component analysis that is supposed to yield a high-computational efficiency and a compact representation of the algorithms. Multi-view embedding was solved by [10] using a semi-supervised learning framework, with which feature embedding can be learned from unlabeled data via predicting one view from another. In [6], a multi-view spectral clustering algorithm was reported in which kernel matrix learning and spectral clustering optimization are integrated into one framework. White et al. [27] presented a convex formulation for learning a shared feature subspace of multiple views. In this formulation, an implicit convex regularizer is exploited and the corresponding reconstruction model is recovered jointly and optimally. Classifier-level learning is another strategy for multiple views. Co-training [3] is one of the earliest classifier learning algorithms to solve multi-view learning. Sindhwani et al. [22]

proposed a co-regularization model in which the view classifiers are learned through multi-view regularization. Wang et al. [26] presented a recursive nonparametric discriminant analysis method to construct probabilistic classifiers for multi-view face detection by learning histograms. Timofte et al. [24] proposed a pipeline for traffic sign detection, recognition, and 3D localization. The accomplished 2D detections in multiple views through the Adaboost are combined to generate 3D techniques for improving results. DietCam [7] was presented as a combination approach for ingredient detection and food classification. From the detected ingredients, food categories are classified using a multi-view kernel support vector machine. It has been proved [13, 23] that classifier-level learning outperforms simple feature-level learning, while sophisticated feature-level learning can usually be better than that of classifier-level [12].

Visual tracking is an important topic in applications of computer vision, such as intelligent video surveillance, human-machine interfaces, medical diagnosis, and public transportation systems [14, 21]. Some papers worked on visual tracking using multi-view data. For example, Wang and Ji [25] proposed a graphical model to track multi-view faces in a cluttering environment. The factorial and switching hidden Markov models are combined in this method. An online detection-based two-stage model [29] was presented for multi-view multi-object tracking. The two-stage online tracking framework seeks both the local optimum for each object and the global optimum for all the tracked objects. Mittal and Davis [19] presented a multi-view approach for segmenting, detecting, and tracking multiple people in a cluttered scene. This model includes a region-based stereo algorithm capable of finding 3D points and a segmentation algorithm using the Bayes classification. In [8], tracking is casted as a multi-task multi-view sparse learning problem, and the cues from multiple views are exploited to be integrated in a particle framework. However, the above trackers fail when the target undergoes severe appearance changes due to limited data supplied by single-sensor feature [35].

Multi-sensor cooperation has higher precision, certainty, and reliability compared with single sensor. Fusion of visible and infrared sensors, one of the typical multi-sensor cooperation, provides complementarily useful features and consistently helps recognize the target from the background efficiently in visual tracking. For instance, an infrared camera can dramatically improve the success rate of distinguishing hot people from comparatively colder background. However, while walking past a group of other humans, a human target may be lost because they all show up as similarly shapes in infrared images. With a visible camera, differences in the colors and texture of different peoples' clothing can make it possible to track the target [18]. By fusing information from both visible and infrared images, they can benefit from one another to achieve more accurate and robust tracking. Besides, features of different views in multi-view learning algorithms obey two major principles that ensure their success: consensus and complementary principles [30]. As the fusion of visible and infrared sensors obeys the major principles of multi-view learning, it is feasible to cast it as a multi-view learning problem. Nevertheless, reports are still few on this topic. Therefore, in this paper, a visible and infrared fusion tracking algorithm is proposed based on multi-view multi-kernel fusion model. The main contributions of our work are:

- The consistent and complementary characteristics of visible and infrared fusion are explored, and then multi-view learning is applied to solve the problem of multi-sensor fusion tracking.
- The proposed multi-view multi-kernel fusion (MVMKF) model embeds complementary information from different views, and uses the multi-kernel framework to learn the importance of view features so as to make an integrated appearance representation with regard to respective performance.
- Our tracking task is completed with naive Bayes classifier in sophisticated compressive feature domain due to the outperformances of classifier-level and sophisticated feature-level learning of multiple views.

The rest of this paper is organized as follows. In Sect. 2, we describe the proposed MVMKF tracking algorithm in detail. The experimental results are presented in Sect. 3. Section 4 concludes with a general discussion.

2 Problem formulation

In the following, the proposed MVMKF tracking algorithm is described in detail.

2.1 Compressive features

The compressive feature vectors are constructed in this section. Each test sample is convolved with a set of Haar-like feature filters at multiple scales [2]. These filtered samples are represented as a very high-dimensional multi-scale image feature vector $\boldsymbol{x} \in \mathbb{R}^n$ [32]. Then, $\boldsymbol{x}$ can be embedded into an extremely compressive feature vector $\boldsymbol{u} \in \mathbb{R}^m$ by a random projection $R \in \mathbb{R}^{m \times n}$. This linear transformation is expressed as [34]

$$\boldsymbol{u} = R\boldsymbol{x}, \tag{1}$$

where $m \ll n$. R has to satisfy the Johnson-Lindenstrauss lemma [1] to restructure $\boldsymbol{x}$ from $\boldsymbol{u}$ with minimum error. Thus, R is defined as [32]

$$r_{ij} = \sqrt{s} \times \begin{cases} 1 \text{ with prob. } 1/2s \\ 0 \text{ with prob. } 1 - 1/s \\ -1 \text{ with prob. } 1/2s. \end{cases} \quad (2)$$

Symbol s is set to be $m / 4$ which satisfies the Johnson-Lindenstrauss lemma and makes a very sparse random matrix, and prob. stands for probability. Therefore, this matrix is data-independent of any training samples and is very easy to compute, thereby resulting in a very efficient method. Based on the linear transformation, each element in the compressive feature vector $\boldsymbol{u}$ in Eq. (1) is a linear combination of spatially distributed rectangle features at different scales [32]. The compressive sensing theory makes the original image be described by the extracted features.

2.2 Multi-view multi-kernel fusion model

Visual tracking from a single view would be inaccurate due to lack of information diversity. To deal with this problem, we develop a multi-view multi-kernel framework for fusion tracking task, by considering target appearances from multiple sensors.

All elements u_i (where $i = 1, \ldots, m$) in compressive feature vector $\boldsymbol{u} = (u_1, \ldots, u_m)$ are assumed independently distributed [32], and the probability density functions (PDFs) of being a target or background are defined as $f_1(u_i)$ and $f_0(u_i)$, respectively. Symbols 1 and 0 denote the labels of target (corresponding to positive sample) and background (corresponding to negative samples), respectively. Kernel density estimation (KDE) [20] is a non-parametric way to estimate PDFs. In this paper, MVMKF model extends KDE to multiple views for multi-feature integration. Therefore, the PDFs of u_i are estimated as

$$\begin{aligned} f_1(u_i) &= \sum_v w_{i,v} k_v(u_i; \mu_i^1, \sigma_i^1), \\ f_0(u_i) &= \sum_v w_{i,v} k_v(u_i; \mu_i^0, \sigma_i^0), \end{aligned} \quad (3)$$

where $v = \{\text{vis}, \text{ir}\}$ denotes the labels of the visible and infrared views. An appropriate kernel function can either be estimated from data or selected as a priori [17]. In this paper, the view kernel functions are assumed to be Gaussian distributed based on empirical study:

$$\begin{aligned} k_v(u_i; \mu_i^1, \sigma_i^1) &= \frac{1}{\sqrt{2\pi\sigma_{i,v}^1}} \exp\left(-\frac{\|u_{i,v} - \mu_{i,v}^1\|^2}{2(\sigma_{i,v}^1)^2} \right), \\ k_v(u_i; \mu_i^0, \sigma_i^0) &= \frac{1}{\sqrt{2\pi\sigma_{i,v}^0}} \exp\left(-\frac{\|u_{i,v} - \mu_{i,v}^0\|^2}{2(\sigma_{i,v}^0)^2} \right), \end{aligned} \quad (4)$$

where $(\mu_{i,v}^1, \sigma_{i,v}^1, \mu_{i,v}^0, \sigma_{i,v}^0)$ are mean and variance parameters.

To cater for flexible circumstances if multiple views compete intensely with each other, the adaptive view weights in Eq. (3) are defined as

$$w_{i,v} = \exp(-\lambda_w \rho_{i,v}^2), \quad (5)$$

where λ_w is a controlling parameter that controls the importance of each view. $\rho_{i,v} = \frac{1}{N_{i,v}} \|u_{i,v} - u_{i,v}(T)\|$ measures the Euclidean distance [5] between $u_{i,v}$ and the template $u_{i,v}(T)$, where $N_{i,v}$ denotes the normalizing vector. Considering short-time tracking without great appearance changes, the feature template is set as the feature in the previous frame.

2.3 Classifier construction and updating

The tracking problem is completed with the naive Bayes classifier [11] as

$$H(\boldsymbol{u}) = \log\left(\frac{\prod_{i=1}^{m} p(u_i|y=1) p(y=1)}{\prod_{i=1}^{m} p(u_i|y=0) p(y=0)} \right) = \sum_{i=1}^{m} \log\left(\frac{p(u_i|y=1)}{p(u_i|y=0)} \right), \quad (6)$$

where $p(u_i|y=1) = f_1(u_i)$ and $p(u_i|y=0) = f_0(u_i)$. The positive and negative probabilities are assumed to be $p(y=1) = p(y=0)$ [34]. The parameters $\left(\mu_{i,v}^1, \sigma_{i,v}^1, \mu_{i,v}^0, \sigma_{i,v}^0\right)$ are incrementally updated as $\mu_{i,v}^1 \leftarrow \lambda \mu_{i,v}^1 + (1-\lambda)\mu_v^1$ and $\sigma_{i,v}^1 \leftarrow \sqrt{\lambda\left(\sigma_{i,v}^1\right)^2 + (1-\lambda)\left(\sigma_v^1\right)^2 + \lambda(1-\lambda)\left(\mu_{i,v}^1 - \mu_v^1\right)^2}$, where $\lambda > 0$ is a learning parameter, and μ_v^1 and σ_v^1 are the mean and covariance parameters computed from the historical frames [32]. Then, we get the tracking result by finding the test sample with the maximal classification score $H(\boldsymbol{u})$.

The proposed MVMKF tracking scheme is summarized in Algorithm 1. In MVMKF, it is assumed that visible and infrared sequences have been registered in time and space spans beforehand. The tracking window in the first frame is located manually or by other detection methods. First, we take some test samples at each frame. Let $l_t(\boldsymbol{x}) \in \mathbb{R}^2$ denote the location of sample $\boldsymbol{x}$ at the tth frame, and $\boldsymbol{x}^*$ represents the sample of the tracking result. At the tth frame in visible image, we select some patches $X^\gamma = \{\boldsymbol{x} | \|l_t(\boldsymbol{x}) - l_{t-1}(\boldsymbol{x}^*)\| < \gamma.\}$ surrounding the target location $l_{t-1}(\boldsymbol{x}^*)$ in the $(t-1)$th frame and set them as the test samples. To extract multi-view features for each sample, we crop patches in infrared image with the same locations. Secondly, we use these test samples to construct the compressive feature vector $\boldsymbol{u}$. At the next step, for each element u_i in $\boldsymbol{u} = (u_1, \ldots, u_m)$, we compute multi-view kernel functions $k_v(u_i; \mu_i^1, \sigma_i^1)$ and $k_v(u_i; \mu_i^0, \sigma_i^0)$ and the view weights $w_{i,v}$ so as to obtain the multi-view multi-kernel fusion PDFs $f_0(u_i)$ and $f_0(u_i)$. Then, the naive Bayes classifier is applied to find

the tracking location $l_t(\boldsymbol{x}^*)$ with the maximal classifier response. After that, we extract a set of positive and negative training samples by randomly cropping patches $X_v^{\alpha} = \{\boldsymbol{x} | \|l_t(\boldsymbol{x}) - l_t(\boldsymbol{x}^*)\| < \alpha\}$ and $X_v^{\varsigma,\beta} = \{\boldsymbol{x} | \varsigma < \|l_t(\boldsymbol{x}) - l_t(\boldsymbol{x}^*)\| < \beta.\}$ surrounding $l_t(\boldsymbol{x}^*)$, respectively, where $\alpha < \varsigma < \beta$. At last, we use these training samples to update the classifier.

designed to observe the contribution of the multi-kernel model to the proposed MVMKF algorithm. In MVSK, the visible and infrared image features are vectorized into a single compressive feature and used to obtain the PDF by single kernel. Figures 1, 2, 3, 4, 5, 6, 7, 8 and

Algorithm 1 General scheme of MVMKF tracking algorithm

Input: tth visible and infrared frames

1. Select some test samples by $X^{\gamma} = \{\boldsymbol{x} | \|l_t(\boldsymbol{x}) - l_{t-1}(\boldsymbol{x}^*)\| < \gamma\}$ in visible and infrared images.
2. Obtain compressive feature vector $\boldsymbol{u}$ using Eq. (1).
3. For each Haar-like feature $i = 1, \cdots, m$, compute kernel functions of each view $k_v(u_i; \mu_i^1, \sigma_i^1)$ and $k_v(u_i; \mu_i^0, \sigma_i^0)$ using Eq. (4).
4. Compute view weights $w_{i,v}$ with Eq. (5).
5. Obtain PDFs of positive and negative samples $f_1(u_i)$ and $f_0(u_i)$ using Eq. (3), and obtain classifier probabilities $p(u_i|y = 1)$ and $p(u_i|y = 0)$.
6. Input classifier probabilities to the Bayes classifier, and find the tracking location $l_t(\boldsymbol{x}^*)$ with the maximal classifier response from $H(\boldsymbol{u})$.
7. Extract positive and negative samples by $X_v^{\alpha} = \{\boldsymbol{x} | \|l_t(\boldsymbol{x}) - l_t(\boldsymbol{x}^*)\| < \alpha\}$ and $X_v^{\varsigma,\beta} = \{\boldsymbol{x} | \varsigma < \|l_t(\boldsymbol{x}) - l_t(\boldsymbol{x}^*)\| < \beta\}$ in visible and infrared images.
8. Update the classifier parameters $(\mu_{i,v}^1, \sigma_{i,v}^1, \mu_{i,v}^0, \sigma_{i,v}^0)$ by $\mu_{i,v}^1 \leftarrow \lambda\mu_{i,v}^1 + (1-\lambda)\mu_v^1$ and $\sigma_{i,v}^1 \leftarrow \sqrt{\lambda(\sigma_{i,v}^1)^2 + (1-\lambda)(\sigma_v^1)^2 + \lambda(1-\lambda)(\mu_{i,v}^1 - \mu_v^1)^2}$.

Output: Tracking location and classifier parameters

3 Experiments

In this section, the MVMKF tracking algorithm is tested on several challenging real-world sequences, and some qualitative and quantitative analyses are performed on the tracking results.

3.1 Experimental setup and evaluation criteria

The sample parameters are set as $\alpha = 4$, $\beta = 30$, $\varsigma = 8$, and $\gamma = 20$, which generate 45 positive samples, 50 negative samples, and 1100 test samples [32]. The controlling and learning parameters are set as $\lambda_w = 1.8$ and $\lambda = 0.85$, respectively. We set the initial values of the view weights as $w_{i,v} = 0.5, v = \{\text{vis}, \text{ir}\}$, meaning that the importance of each view is equal in the tracking beginning, which provides them enough competitive space. The dimension of compressive feature vector is set as $m = 50$ and other parameters are set according to [32].

The performance of the MVMKF tracking algorithm is compared with state-of-the-art CT (compressive tracking) [32], MIL (multiple instance learning) [2], ODFS (online discriminative feature selection) [33], STC (spatio-temporal context) [31], FRDIF (fuzzified region dynamic image fusion) [28], and MVSK (multi-view single-kernel) tracking algorithms. MVSK is

Tables 2, 3 present the experimental results in six challenging sequences named *Labman*, *Cross*, *Shadow*, *Occlusion 1*, *Occlusion 2*, and *Ourlab*. The first sequence is downloaded from the AIC dataset [4]. Sequences *Cross*, *Shadow*, *Occlusion 1*, and *Occlusion 2* come from the OTCBVS dataset which is available at http://vcipl-okstate.org/pbvs/bench/. The last sequence is recorded in our lab. The details of the test sequences (including the visible and infrared cameras, the arrangement of two sensors, frame rate, image resolution, target size, etc) are presented in Table 1.

Two metrics, i.e., location error (pixel) [9] and overlapping rate [33], are used to evaluate the tracking results of MVMKF quantitatively. The location error is computed as $\text{error} = \sqrt{(x_G - x_T)^2 + (y_G - y_T)^2}$, where (x_G, y_G) and (x_T, y_T) are the ground truth and tracking bounding box centers, respectively. The tracking overlapping rate is defined as $\text{overlapping} = \text{area}(\text{ROI}_G \cap \text{ROI}_T)/\text{area}(\text{ROI}_G \cup \text{ROI}_T)$, where ROI_T and ROI_G denote the tracking bounding box and ground truth which is manually located, respectively. Symbol $\text{area}(\cdot)$ denotes the rectangular area function. A smaller location error and a larger overlapping rate indicate higher accuracy and robustness. Next, the performance of each sequence is described in detail.

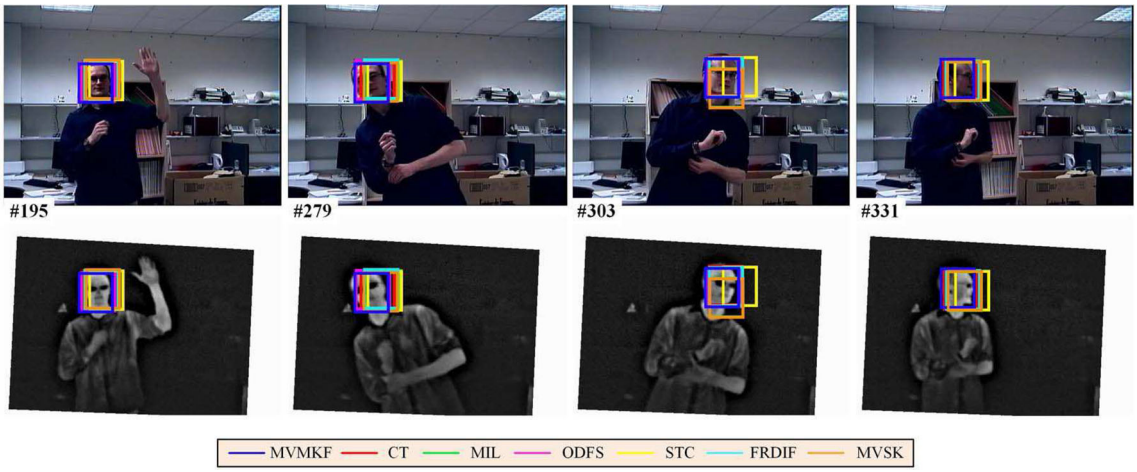

Fig. 1 Tracking results of Sequence *Labman* in visible and infrared images

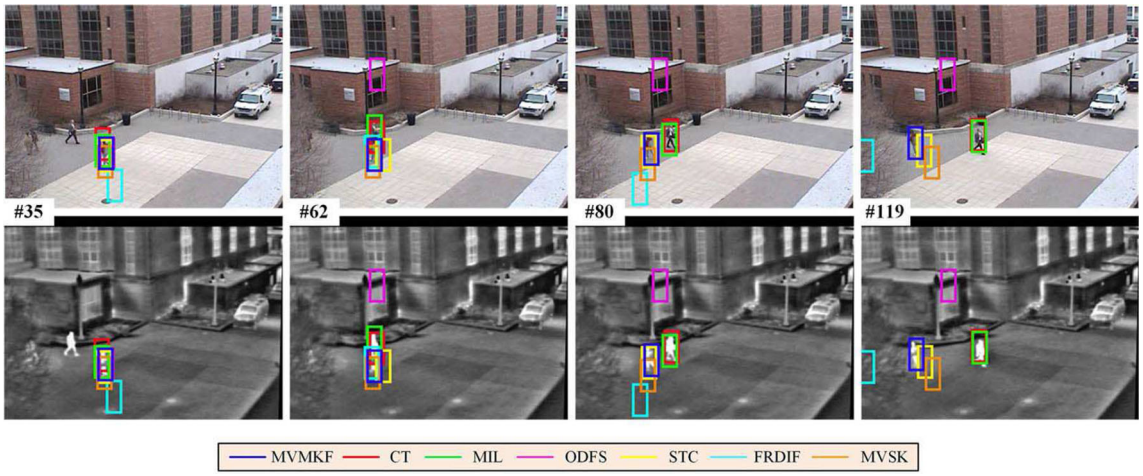

Fig. 2 Tracking results of Sequence *Cross* in visible and infrared images

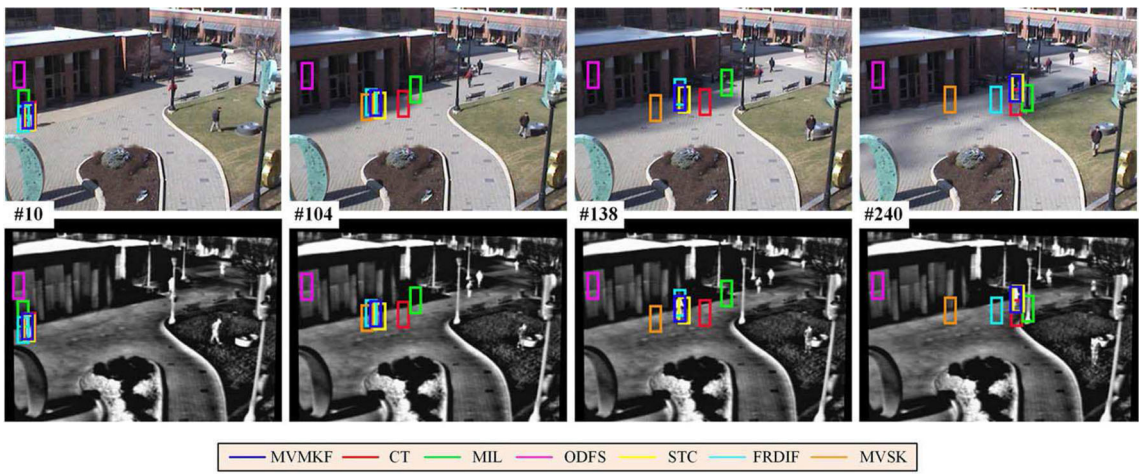

Fig. 3 Tracking results of Sequence *Shadow* in visible and infrared images

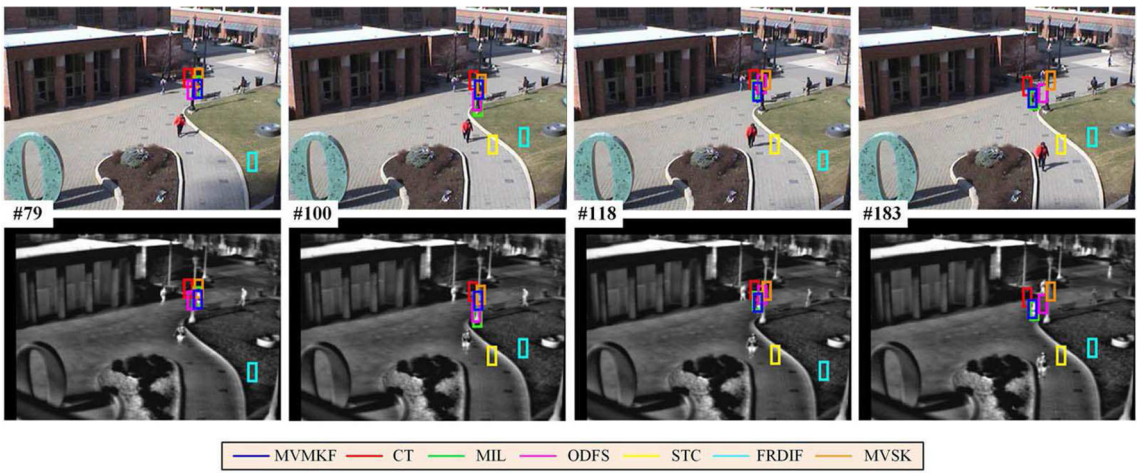

Fig. 4 Tracking results of Sequence *Occlusion 1* in visible and infrared images

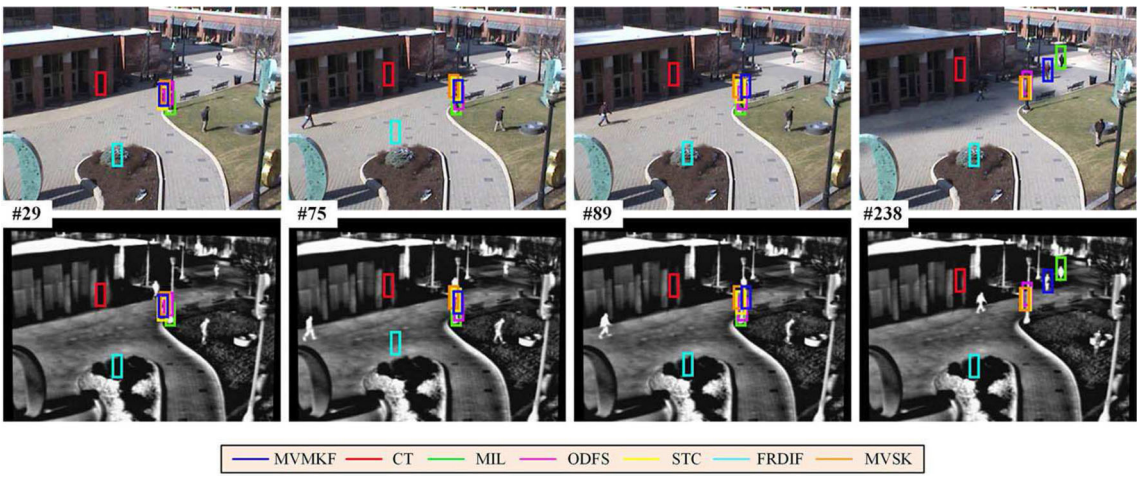

Fig. 5 Tracking results of Sequence *Occlusion 2* in visible and infrared images

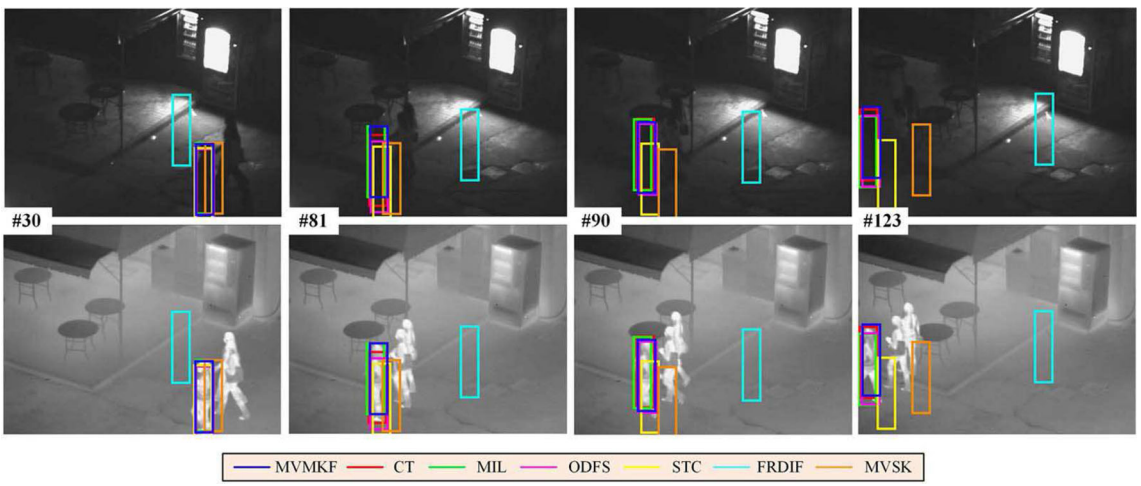

Fig. 6 Tracking results of Sequence *Ourlab* in visible and infrared images

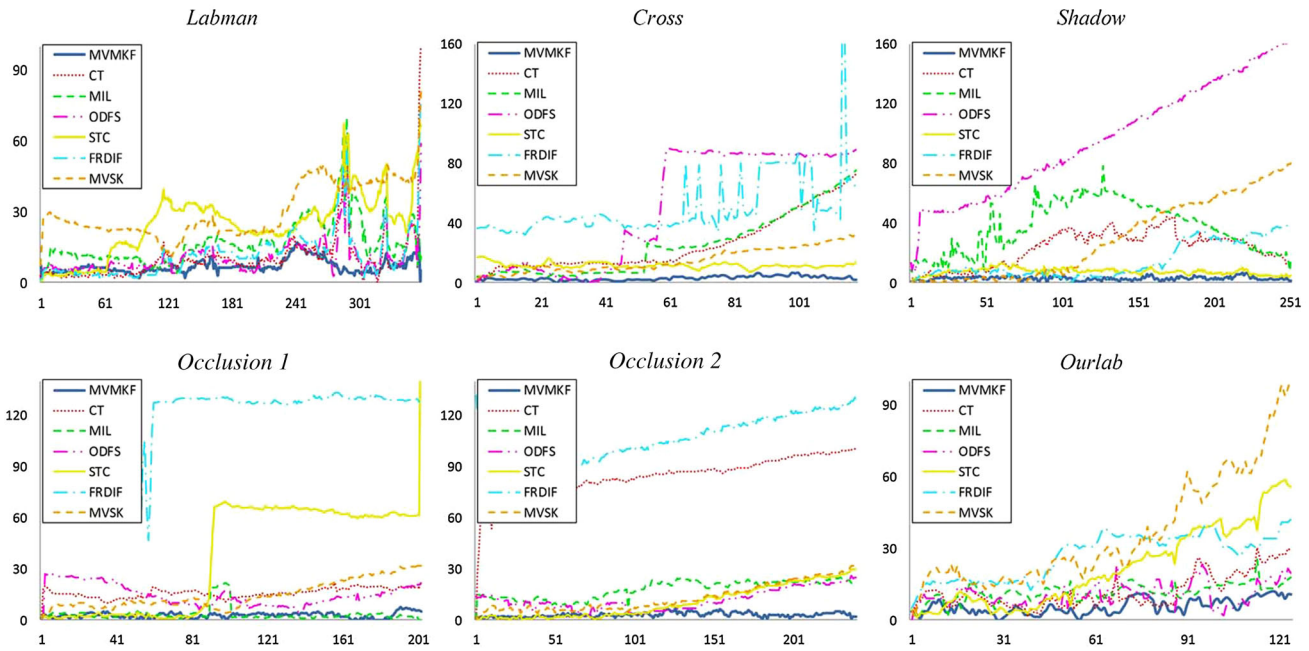

Fig. 7 Comparisons on location error (pixel) of the test sequences

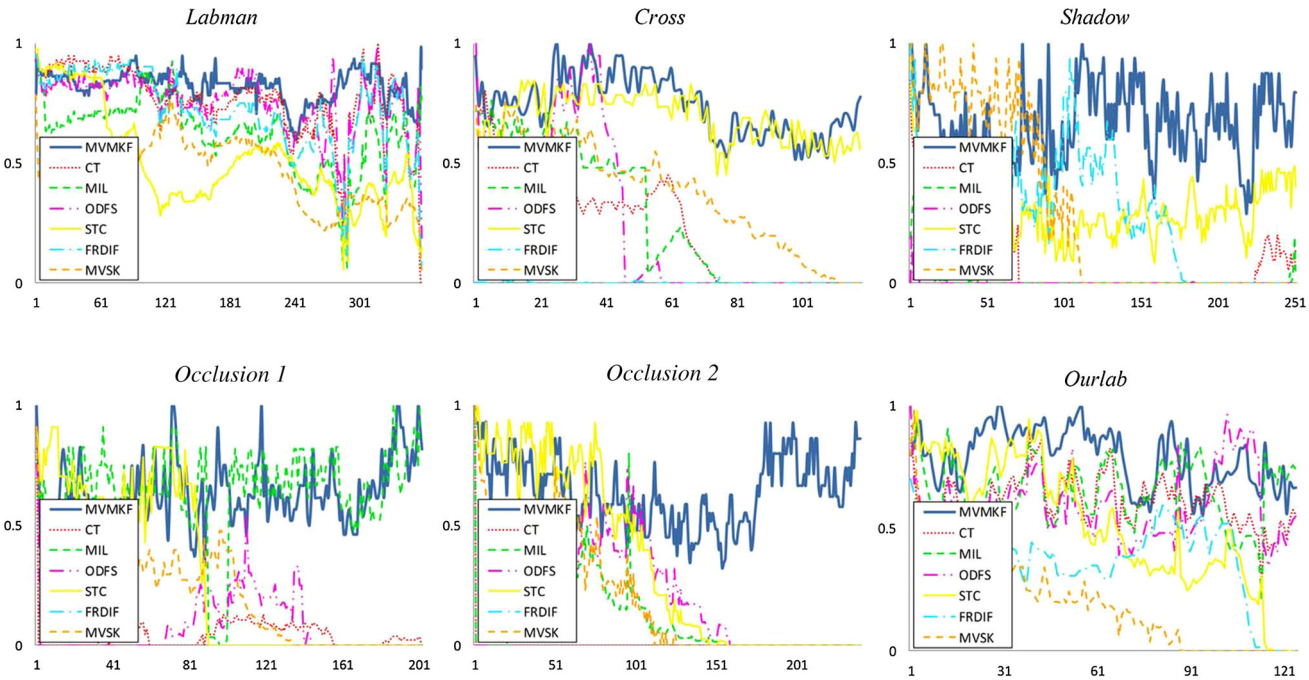

Fig. 8 Comparisons on overlapping rate (%) of the test sequences

3.2 Tracking results

3.2.1 Abrupt rotation and movement

The efficiency of MVMKF is demonstrated by using Sequence *Labman* (360 frames in total), in which abrupt rotation and movement is displayed. Due to the limitation of space, only four frames of each sequence are displayed. At the tracking beginning, the CT, MIL, ODFS, STC, FRDIF, and MVSK trackers can all track the target successfully. But when the man starts to shake his head from right to left abruptly from Frame #279 to #303 in Fig. 1,

Table 1 Details of the test sequences

Sequences	*Labman*	*Cross*	*Shadow*	*Occlusion 1*	*Occlusion 2*	*Ourlab*
Visible camera	Panasonic WV-CP470	Sony TRV87 Handycam	Sony TRV87 Handycam	Sony TRV87 Handycam	Sony TRV87 Handycam	UNIQ UM-301
Infrared camera	Raytheon Thermal IR-2000B	Raytheon PalmIR 250D	Raytheon PalmIR 250D	Raytheon PalmIR 250D	Raytheon PalmIR 250D	CEDIP IR camera
Arrangement	Aligned	Aligned	Aligned	Aligned	Aligned	Aligned
Frame rate (FPS)	25	25	25	25	25	25
Image resolution (pixel)	640*480	320*240	320*240	320*240	320*240	480*360
Target size (pixel)	76*91	16*37	12*30	10*21	10*27	31*123

CT, MIL, STC, FRDIF, and MVSK fail to locate the target accurately. And we can see that MVMKF performs the best when the man turns around this head to the left at Frame #331. The comparisons about location error and overlapping rate for the seven trackers are presented in Figs. 7, 8. MVMKF is able to overcome the abrupt appearance change and performs well on this sequence.

3.2.2 Background similarity and change

Sequence *Cross* (119 frames in total) contains examples of background similarity and change. In this sequence, the human target walks through a cross-background (color change of the earth), and another similar man is passing by him. As seen in Frame #62 in Fig. 2, ODFS and STC lose the target, and CT and MIL fail and track a wrong object. The reason is that the change of the earth color affects much on single-sensor trackers (CT, MIL, ODFS, and STC). The multi-sensor fusion tracker FRDIF and MVSK also fail at around Frame #80 because their fusion method is not equipped with multi-kernel learning models. For the convenience of presentation, the FRDIF tracking curve is not shown entirely in Fig. 7. MVMKF is able to track accurately. As seen in Figs. 7, 8, MVMKF has the smallest location error and largest overlapping rate during most of the tracking process. In contrast, the location errors and overlapping rates of the other six trackers increased and decreased frame by frame, respectively.

3.2.3 Shadow and illumination change

The target in Sequence *Shadow* (252 frames in total) cannot be recognized clearly as a result of the shadow of moving clouds and illumination change. Since the appearance models of CT, MIL, and ODFS are not learned well in this complex circumstance, their location errors in Fig. 7 keep large most of time and overlapping rates in Fig. 8 are almost zero when the target is covered by the cloud shadow. MVSK and FRDIF begin to fail when the target walks into the shadow of the building at around Frame #100 and #240, respectively. Only MVMKF and STC can handle these problems whereas the result of MVMKF is more accurate due to the well learned appearance from multiple views.

3.2.4 Occlusion

In Sequence *Occlusion 1* (203 frames in total), the human target is occluded by a lamppost. Figure 4 indicates that only MIL and MVMKF perform well on this sequence. However, MIL mistakes the lamppost for the target when the occlusion occurs (around Frame #100) because it doesn't have multi-view features, which affects the performance when the occlusion object has similar features with the target in a single view. In comparison with the other six trackers, MVMKF presents the best performance.

The target in Sequence *Occlusion 2* (241 frames in total) is also heavily occluded and encounters complex background disturbance. CT and FRDIF mistake other things for the target for almost the whole tracking process. This is also reflected in overlapping rates which are almost zero in Fig. 8. MIL, ODFS, STC, and MVSK do wrong to track the lamppost or another nearby person. Only MVMKF does well on tracking accuracy frame by frame.

3.2.5 Night tracking

Sequence *Ourlab* (125 frames in total), provided by our lab, presents night tracking. The target in the visible sequence can barely be seen with naked eye. What makes the tracking more challenging is that the human target is surrounded by several people that have similar appearance with it. Once again, MVMKF delivers the best performance for this video due to the effectiveness and robustness of the proposed fusion method, as shown in Figs. 7, 8.

Table 2 Comparisons on average location error (pixel) for the test sequences and average frame per second (FPS)

Sequences	*Labman*	*Cross*	*Shadow*	*Occlusion 1*	*Occlusion 2*	*Ourlab*	*FPS*
MVMKF	**7**	**3**	**3**	**3**	**3**	**6**	39
CT	10	26	23	17	84	12	42
MIL	17	25	37	4	18	12	12
ODFS	11	50	96	16	12	11	36
STC	24	**3**	7	37	10	22	**163**
FRDIF	12	50	14	134	105	27	23
MVSK	30	17	29	15	13	36	34

Bold fonts indicate the best performances

Table 3 Comparisons on success rate (%) for the test sequences

Sequences	*Labman*	*Cross*	*Shadow*	*Occlusion 1*	*Occlusion 2*	*Ourlab*
MVMKF	**100**	**100**	**85**	83	**82**	**100**
CT	94	6	2	0	0	83
MIL	78	30	0	**90**	1	86
ODFS	94	29	0	1	7	71
STC	44	98	4	45	78	55
FRDIF	92	0	17	0	0	15
MVSK	48	39	84	36	60	33

Bold fonts indicate the best performances

Tables 2, 3 are included here to demonstrate the performance on the average location error (pixel) and success rate (%) of the six test sequences. The success rate is defined as the number of times success is achieved in the whole tracking process by considering one frame as a success if the overlapping rate exceeds 0.5 [33]. A smaller average location error and a larger success rate indicate higher accuracy and robustness. In Sequences *Shadow*, *Occlusion 1*, and *Occlusion 2*, most of the trackers do not achieve a large success rate because the target sizes are relatively small (see Table 1) such that a slight drift away from the target may cause a great reduction in the success rates. Although the success rate of MIL is larger than MVMKF in Sequence *Occlusion 1* in Table 3, MIL is unstable with the failure shown in Fig. 4. Besides, the average location error of MIL is smaller than that of MVMKF in Table 2. Trackers MVMKF, CT, FRDIF, and MVSK are implemented using Visual Studio 2010, and MIL, ODFS, and STC are implemented in MATLAB R2010a. MVMKF runs 39 frames per second (FPS) on average on an Intel Dual-Core 1.70 GHz CPU with 4 GB RAM, and the average FPS of the other six trackers are presented in Table 2. Tables 2, 3 shows that although MVMKF is not the fastest in terms of computational time, it has greater stability and better tracking accuracy.

3.3 Computational complexity

Efficiency is one prime characteristic of the proposed MVMKF tracking algorithm. The construction process of compressive feature vector has a low complexity of $o(n)$, because the sparse projection matrix R is independent of training samples, which needs to be computed only once offline and remains fixed throughout the tracking process. Symbol n denotes the dimension of multi-scale image feature vector $\boldsymbol{x}$, which is determined by the inputted images. In computing the multi-view multi-kernel classifier, the computational complexity is $o(\log(m))$ where m denotes the dimension of compressive feature vector. Therefore, the theoretically global computational complexity is really small in practice. Experimentally, the proposed algorithm has a high computational efficiency which runs 39 FPS on average.

4 Conclusion

In this paper, a visible and infrared tracking algorithm based on multi-view multi-kernel fusion (MVMKF) model is developed. In comparison with single sensor, cooperation of multi-sensor features is able to make up the weak points of each other and provides higher precision, certainty, and reliability. Fusion of visible and infrared sensors, one of the typical multi-sensor cooperation, has consistent and complementary properties. Due to these two properties, we apply multi-view learning to solve the problem of visible and infrared fusion tracking. In this paper, the MVMKF model is capable of embedding complementary information from different views to discover an integrated appearance representation. The multi-kernel framework is applied to learn the relative importance of view features and make a combination with regard to respective performance. Besides, the tracking task is

completed with naive Bayes classifier in sophisticated compressive feature domain, considering the better performances of classifier-level and sophisticated feature-level learning for multi-view learning. Numerous real-world video sequences were used to test MVMKF and other state-of-the-art trackers, and here we only selected representative ones for presentation. Experimental results were used to demonstrate that MVMKF is highly accurate and robust.

Acknowledgments This work is supported by the National Natural Science Foundation of China (Grant Nos. 61175028, 61365009) and the Ph.D. Programs Foundation of the Ministry of Education of China (Grant Nos. 20090073110045).

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
2. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1619–1632 (2011)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100. ACM (1998)
4. Conaire, C.Ó., OConnor, N.E., Smeaton, A.: Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. Mach. Vis. Appl. **19**(5–6), 483–494 (2008)
5. Deza, M.M., Deza, E.: Encyclopedia of distances. Springer (2009)
6. Guo, D., Zhang, J., Liu, X., Cui, Y., Zhao, C.: Multiple kernel learning based multi-view spectral clustering. In: Pattern recognition (ICPR), 2014 22nd international conference on, pp. 3774–3779. IEEE (2014)
7. He, H., Kong, F., Tan, J.: Dietcam: Multi-view food recognition using a multi-kernel svm. IEEE J. Biomed. Health Inf. **1**(99), 1–8 (2015)
8. Hong, Z., Mei, X., Prokhorov, D., Tao, D.: Tracking via robust multi-task multi-view joint sparse representation. In: Computer vision (ICCV), 2013 IEEE international conference on, pp. 649–656. IEEE (2013)
9. Jiang, N., Liu, W., Wu, Y.: Learning adaptive metric for robust visual tracking. IEEE Trans. Image Process. **20**(8), 2288–2300 (2011)
10. Johnson, R., Zhang, T.: Semi-supervised learning with multi-view embedding: theory and application with convolutional neural networks. (2015). arXiv:1504.01255
11. Jordan, A.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. Adv. Neural Inf. Process. Syst. **14**, 841 (2002)
12. Klausner, A., Tengg, A., Rinner, B.: Vehicle classification on multi-sensor smart cameras using feature-and decision-fusion. In: Distributed smart cameras, 2007. ICDSC'07. First ACM/IEEE international conference on, pp. 67–74. IEEE (2007)
13. Kludas, J., Bruno, E., Marchand-Maillet, S.: Information fusion in multimedia information retrieval. In: Adaptive multimedia retrieval: retrieval, user, and semantics, pp. 147–159. Springer (2008)
14. Koishi, T., Sasaki, M., Nakaguchi, T., Tsumura, N., Miyake, Y.: Endoscopy system for length measurement by manual pointing with an electromagnetic tracking sensor. Opt. Rev. **17**(2), 54–60 (2010)
15. Luo, Y., Liu, T., Tao, D., Xu, C.: Multiview matrix completion for multilabel image classification. IEEE Trans. Image Process. **24**(8), 2355–2368 (2015)
16. Luo, Y., Tang, J., Yan, J., Xu, C., Chen, Z.: Pre-trained multi-view word embedding using two-side neural network. In: Twenty-eighth AAAI conference on artificial intelligence, pp. 1982–1988 (2014)
17. Meltzer, J., Yang, M.H., Gupta, R., Soatto, S.: Multiple view feature descriptors from image sequences via kernel principal component analysis. In: Computer vision–ECCV 2004, pp. 215–227. Springer (2004)
18. Mihaylova, L., Loza, A., Nikolov, S.G., Lewis, J.J., Canga, E.F., Li, J., Dixon, T.D., Canagarajah, C.N., Bull, D.R.: The influence of multi-sensor video fusion on object tracking using a particle filter. GI Jahrestagung **1**, 354–358 (2006)
19. Mittal, A., Davis, L.S.: M2tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. Int. J. Comput. Vis. **51**(3), 189–203 (2003)
20. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat., pp. 1065–1076 (1962)
21. Peng, C., Chen, Q., Qian, W.X.: Eigenspace-based tracking for feature points. Opt. Rev. **21**(3), 304–312 (2014)
22. Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semi-supervised learning with multiple views. In: Proceedings of ICML workshop on learning with multiple views, pp. 74–79. Citeseer (2005)
23. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on multimedia, pp. 399–402. ACM (2005)
24. Timofte, R., Zimmermann, K., Van Gool, L.: Multi-view traffic sign detection, recognition, and 3d localisation. Mach. Vis. Appl. **25**(3), 633–647 (2014)
25. Wang, P., Ji, Q.: Multi-view face tracking with factorial and switching hmm. In: Application of computer vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on, vol. 1, pp. 401–406. IEEE (2005)
26. Wang, P., Ji, Q.: Multi-view face and eye detection using discriminant features. Comput. Vis. Image Underst. **105**(2), 99–111 (2007)
27. White, M., Zhang, X., Schuurmans, D., Yu, Y.l.: Convex multi-view subspace learning. In: Advances in neural information processing systems, pp. 1673–1681 (2012)
28. Xiao, G., Yun, X., Wu, J.: A multi-cue mean-shift target tracking approach based on fuzzified region dynamic image fusion. Sci. China Inf. Sci. **55**(3), 577–589 (2012)
29. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on, pp. 1200–1207. IEEE (2009)
30. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning, pp. 1–59 (2013). arXiv:1304.5634
31. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.H.: Fast visual tracking via dense spatio-temporal context learning. In: Computer vision–ECCV 2014, pp. 127–141. Springer (2014)
32. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: Computer vision–ECCV 2012, pp. 864–877. Springer (2012)
33. Zhang, K., Zhang, L., Yang, M.H.: Real-time object tracking via online discriminative feature selection. IEEE Trans. Image Process. **22**(12), 4664–4677 (2013)
34. Zhang, K., Zhang, L., Yang, M.H.: Fast compressive tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 2002–2015 (2014)
35. Zhou, P., Yao, J., Pei, J.: Implementation of an energy-efficient scheduling scheme based on pipeline flux leak monitoring networks. Sci. China Ser. F Inf. Sci. **52**(9), 1632–1639 (2009)