Visual Tracking with Spatio-Temporal Dempster-Shafer Information Fusion

Xi Li, Anthony Dick, Chunhua Shen, Zhongfei Zhang, Anton van den Hengel, Hanzi Wang

***Abstract*—*A key problem in visual tracking is how to effectively combine spatio-temporal visual information from throughout a video to accurately estimate the state of an object. We address this problem by incorporating Dempster-Shafer information fusion into the tracking approach. To implement this fusion task, the entire image sequence is partitioned into spatially and temporally adjacent subsequences. A support vector machine (SVM) classifier is trained for object/non-object classification on each of these subsequences, the outputs of which act as separate data sources.*

To combine the discriminative information from these classifiers, we further present a spatio-temporal weighted Dempster-Shafer (STWDS) scheme. Moreover, temporally adjacent sources are likely to share discriminative information on object/non-object classification. In order to use such information, an adaptive SVM learning scheme is designed to transfer discriminative information across sources. Finally, the corresponding Dempster-Shafer belief function of the STWDS scheme is embedded into a Bayesian tracking model. Experimental results on challenging videos demonstrate the effectiveness and robustness of the proposed tracking approach.

CONTENTS

X. Li, A. Dick, C. Shen, and A. van den Hengel are with School of Computer Sciences, University of Adelaide, SA 5005, Australia.

Correspondence should be address to C. Shen (e-mail: chunhua.shen@adelaide.edu.au).

Z. Zhang is with State University of New York, Binghamton, NY 13902, USA.

H. Wang is with Center for Pattern Analysis and Machine Intelligence, Xiamen University, 361005 China.

I. Introduction

Visual tracking has a wide range of potential applications including video surveillance for security and traffic management, health care, human-computer interaction, robotic vision, object detection, and multimedia. A popular approach to visual tracking is to learn a discriminative appearance model for coping with complicated appearance changes. Typically, this assumes that the object/non-object discriminative information from different frames is generated from a temporally homogeneous source. However, this assumption may not hold in practice, as object appearance and environmental conditions vary dynamically over time. In addition, some intrinsic factors also affect object appearance, including shape deformation, pose variation, out-of-plane rotation, etc. In the face of such challenging factors, fitting a static discriminative model is unlikely to optimally distinguish an object from its background.

Our approach is to break incoming video into spatially and temporally adjacent subsequences and to treat each subsequence as a separate, but related, set of data to which a discriminative model is fitted. In this way, we obtain a sequence of discriminative models acting as separate information sources as visual tracking proceeds. Effectively combining these information sources plays a critical role in our approach. Most existing fusion techniques treat the sources equally or independently, and thus ignore the spatio-temporal differences and correlations among them. To address this issue, we propose a spatio-temporal weighted Dempster-Shafer (STWDS) scheme for combining the discriminative information from multiple information sources. The proposed STWDS scheme is capable of capturing both time-related and space-related discriminative information for object/non-object classification, leading to robust tracking results. The main contributions of this paper are three-fold:

- We introduce multi-source discriminative learning into visual tracking. The problem of visual tracking is converted to that of discriminative learning in a sequence of spatially and temporally adjacent video subsequences, each of which is treated as a discriminative information source for object/non-object classification.
- We present a spatio-temporal weighted Dempster-Shafer (STWDS) scheme for combining the evidence from both time-related and space-related discriminative information sources (as described in Sec. IV-B). The corresponding Dempster-Shafer belief function of the STWDS scheme is used as the observation model of a particle filter, resulting in robust tracking results. We believe this is the first time that such a fusion method has been adapted to visual tracking.
- We design an adaptive SVM learning (ASL) scheme for transferring space-related discriminative information across time-adjacent sources. To capture time-related discriminative information, the ASL scheme uses adaptive SVM learning for propagating the discriminative information of the prior SVM classifier into the current source. After acquiring new discriminative information in the current source, it seeks to adaptively adjust the propagation direction, resulting in a new SVM classifier in the current source.

II. Related work

This section gives a brief review of the related tracking approaches using the discriminative learning and information fusion techniques.

A. Discriminative learning based tracking

Discriminative learning based tracking approaches try to build a strong classifier for distinguishing a tracked foreground object from background patterns. An online AdaBoost classifier [5] is employed for discriminative feature selection, which enables the tracking approach to adapt to appearance variations caused by out-of-plane rotations and illumination changes. Following the work of [5], Grabner *et al.* [6] present a semi-supervised online boosting approach for visual tracking. This approach can significantly alleviate the model drifting problem caused during updating the model for the online AdaBoost classifier. Avidan [10] constructs a confidence map by pixel classification using an ensemble of online learned weak classifiers. Collins *et al.* [12] propose an online feature selection approach for visual tracking. This approach tries to find the most discriminative linear combinations of the RGB color space in each frame. Liu and Yu [13] propose an efficient online boosting approach based on gradient-based feature selection. Babenko *et al.* [9] present a tracking system based on online multiple instance boosting, which takes the uncertainty of object localization into account. Moreover, Avidan [7] proposes an off-line SVM-based tracking approach for distinguishing the target vehicle from background. Tian *et al.* [8] utilize the ensemble of linear SVM classifiers for visual tracking. These classifiers can be adaptively weighted according to their discriminative abilities during different periods, resulting in the robustness to large appearance variations during tracking. In order to capture the contextual information on object samples, Li *et al.* [11] construct a contextual kernel based on graph mode seeking, and then embed the contextual kernel into the SVM tracking process.

In addition, Yang *et al.* [31] propose a discriminative tracker that can track non-stationary visual appearances by data-driven subspace adaptation, which encourages the mutual closeness of the positive data to their projections and the mutual separability of the negative data from their projections. Besides, Fan *et al.* [32] present a human tracking approach based on convolutional neural networks, which can effectively learn spatio-temporal features from image pairs of two adjacent frames. Recently, discriminative metric learning has also been successfully applied to visual tracking [34], [35], [36], [33]. It aims to learn a distance metric to capture the correlation information between different feature dimensions for robust tracking. Furthermore, Li *et al.* [39] propose a compact and discriminative 3D-DCT (3D discrete Cosine transform) object representation that poses object tracking as a signal compression and reconstruction problem (solved by the fast Fourier transform).

B. Information fusion based tracking

In general, information fusion based tracking approaches are based on two types of techniques: i) single-modal multi-source information fusion; and ii) multi-modal information fusion, which aims to fuse information from different types of sensors (e.g. [22], [23], [24], [25]). Here, we focus on single modal techniques, aiming to utilize multiple visual cues obtained from a visible light camera at different times and locations to compensate for noisy, partial or missing observations.

For instance, Wu and Huang [16] propose a co-inference approach for integrating and tracking multiple cues (e.g., color and shape), resulting in an accurate and robust state estimation on the image observations. Hua and Wu [17] present a part-based visual tracking framework for detecting and integrating inconsistent measurements (i.e., contradictory to the majority of their neighbors). By eliminating these inconsistent measurements, the presented framework can achieve promising tracking results. Han and David [20] develop a tracking approach based on robust fusion, which is used for estimating the complicated motion parameters. By fusing individual motion components, the developed tracking approach is capable of obtaining robust motion parameters that are immune to outliers. Xu *et al.* [21] propose a tracking approach which fuses partial estimates from different sources. By exploring the connection between the probabilistic data fusion and computational geometry, the proposed tracking approach can obtain the global optimal estimation in a high-dimensional parameter space. Fan *et al.* [18] develop a multiple collaborative kernel tracking approach for increasing the “kernel observability” for articulated objects. Multiple kernels are utilized to explore the diverse information from different perspectives. Tang *et al.* [19] present an online semi-supervised learning based tracker. The method constructs two feature-specific SVM classifiers (i.e., color and HOG features) in a co-training framework, and thus is capable of improving each individual classifier using the information from other features. Adam *et al.* [1] propose a tracking approach based on a patch-division object representation with a histogram-based feature description. The final tracking position is determined by combining the vote maps of all the patches (represented by grayscale histograms). The combination mechanism can eliminate the influence of the outlier vote maps caused by occlusion.

III. Tracking approach: overview

The goal of a visual tracking method is to estimate the location of a target at each frame, given only the previous frame data. As is common, we assume that the target motion is approximated by a Markov model. Let $\mathbf{Z}_t = (\mathcal{X}_t, \mathcal{Y}_t, \mathcal{S}_t)$ denote the motion parameters at frame t including $\mathcal{X}$ translation, $\mathcal{Y}$ translation, and scaling. Then we estimate:

$$p(\mathbf{Z}_t|\mathcal{O}_t) \propto p(\mathbf{o}_t|\mathbf{Z}_t)\int p(\mathbf{Z}_t|\mathbf{Z}_{t-1})p(\mathbf{Z}_{t-1}|\mathcal{O}_{t-1})d\mathbf{Z}_{t-1}, \quad (1)$$

where $\mathcal{O}_t = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$ are the observed data, $p(\mathbf{o}_t \mid \mathbf{Z}_t)$ denotes the observation model, and $p(\mathbf{Z}_t \mid \mathbf{Z}_{t-1})$ represents the state transition model. The motion model between two consecutive frames is assumed to be a Gaussian distribution:

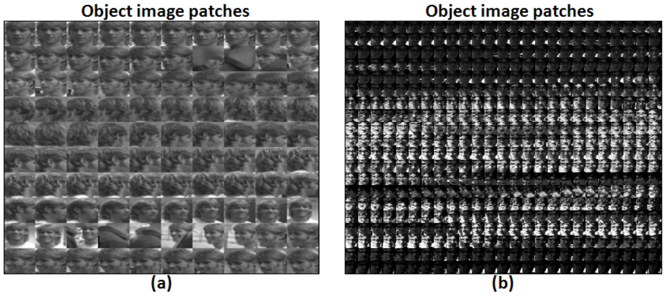

Fig. 1. Illustration of multi-source distribution properties. (a) and (b) show 100 and 500 consecutive object image patches obtained by the proposed tracker on the “seq-jd” and the “trellis70” video sequences, respectively. While the appearance of nearby patches is correlated, it changes significantly over the course of the videos.

$$p(\mathbf{Z}_t|\mathbf{Z}_{t-1}) = \mathcal{N}(\mathbf{Z}_t; \mathbf{Z}_{t-1}, \Sigma), \quad (2)$$

where Σ denotes a diagonal covariance matrix with diagonal elements: $\sigma^2_{\mathcal{X}}$, $\sigma^2_{\mathcal{Y}}$, and $\sigma^2_{\mathcal{S}}$. In this framweork, the optimal object state $\mathbf{Z}^*_t$ at time t is determined by solving the maximum a posterior (MAP) problem:

$$\mathbf{Z}^*_t = \arg\max_{\mathbf{Z}_t} p(\mathbf{Z}_t|\mathcal{O}_t), \quad (3)$$

The workflow of the proposed tracking approach is summarized in Algorithm 1. We approximate the distribution $p(\mathbf{Z}_t \,|\, \mathcal{O}_t)$ using a pool of samples which is updated at each frame. Each sample is a translation and scaling of the original target location and size, and thus represents an image patch in the current frame. Our main contribution in this paper is the use of a Dempster-Shafer fusion strategy for the observation (or likelihood) model $p(\mathbf{o}_t|\mathbf{Z}_t)$ used to evaluate each sample, which is described in the following Section.

IV. Dempster-Shafer theory for visual tracking

A. Motivation

Dempster-Shafer is a data fusion theory that combines evidence from different information sources [27], [28], [29], [30]. Typically, it is applied to the decision-making process using different information sources with uncertainty.

In this paper, we apply the Dempster-Shafer theory to visual tracking in order to combine evidence from multiple frames, and from multiple regions within each frame. This allows us to include information from other frames and locations in the video in our implementation of the likelihood function $p(\mathbf{o}_t \,|\, \mathbf{Z}_t)$. The key advantage of the Dempster-Shafer theory over Bayesian formulations is that it allows us to represent ambiguity and mutual contradiction more explicitly, which in turn leads to more robust tracking results.

Before describing our tracking approach, we introduce the following terminologies:

- *Mass function.* Let $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ denote a set of mutually exhaustive and exclusive hypotheses. The power set of Θ is defined as the set containing all the 2^N possible subsets of Θ, denoted as $P(\Theta)$:

$$P(\Theta)=\{\emptyset,\{\theta_1\},\dots,\{\theta_N\},\{\theta_1,\theta_2\},\{\theta_1,\theta_3\},\dots,\Theta\}, \quad (4)$$

```
Initialization:
    – $N = 1$, $t = 1$.
    – Number of samples $V$.
    – Maximum buffer size $W$.
    – Maximum number of sources $Q$.
    – Manually set initial object state $\mathbf{Z}_1^*$.
    – Collect positive and negative samples to form a training set $\mathcal{F}$ (see Sec. V-A) and extract features for all the sub-regions such that $l \in \{1, 2, 3, 4, 5\}$.
    – Train the SVM classifiers $\{h_{l1}\}_{l=1}^5$ over the features.
At frame t:
Input: Existing SVM classifiers $\left\{\{h_{ln}\}_{l=1}^5\right\}_{n=1}^N$, new frame $t$, previous object state $\mathbf{Z}_{t-1}^*$.
begin
    • Sample $V$ candidate object states $\{\mathbf{Z}_{tj}\}_{j=1}^V$ by Eq. (2).
    • Extract the features $\{\mathbf{x}_{tj}^l\}_{l=1}^5$ for each $\mathbf{Z}_{tj}$
    • for each $\mathbf{Z}_{tj}$ do
        – Time-related Dempster-Shafer information fusion.
            – Compute the normalized SVM confidence score $g_{ln}(\mathbf{x}_{tj}^l)$ according to Eq. (14).
            – Calculate the time-weighted mass function $m_{ln}^*(A_n)$ in Eq. (15).
            – Obtain the combined mass function $\mathbb{M}_l(A)$ in Eq. (16).
        – Space-related Dempster-Shafer information fusion.
            – Compute the spatial-weighted mass function $\mathbb{M}_l^*(A)$ in Eq. (17).
            – Calculate the combined mass function $\mathcal{M}(A)$ in Eq. (18).
            – Obtain the belief function $Bel(A)$ in Eq. (19).
      end
    • Determine the optimal object state $\mathbf{Z}_t^*$ by the MAP estimation in Eq. (3).
    • Collect positive (or negative) samples $\mathcal{Z}_t^+$ (or $\mathcal{Z}_t^-$) (referred to in Sec. V-A).
    • Update the training sample sets $\mathcal{F}$ with $\mathcal{F} \bigcup \mathcal{Z}_t^+ \bigcup \mathcal{Z}_t^-$.
    if $t$ mod $W = 0$ then
        • Extract features from samples in $\mathcal{F}$.
        • Run adaptive SVM learning over the extracted features and $h_{lN}$ to generate a new SVM classifier $h_{l(N+1)}$.
        Create a new source associated with $\{h_{l(N+1)}\}_{l=1}^5$, $N = N + 1$, and $\mathcal{F} = \emptyset$.
    end
    if $N > Q$ then
        $\left\{\{h_{ln}\}_{l=1}^5\right\}_{n=1}^N$ is truncated to keep the last $Q$ sources.
    end
    $t = t + 1$.
end
Output: SVM classifiers $\left\{\{h_{ln}\}_{l=1}^5\right\}_{n=1}^N$, current object state $\mathbf{Z}_t^*$, and updated training sample sets $\mathcal{F}$.
```

Algorithm 1: Overview of the tracking algorithm.

where $\emptyset$ denotes the empty set. A probability mass function is introduced to define a mapping: $m : P(\Theta) \rightarrow [0, 1]$ which has the following properties:

$$\sum_{A \in P(\Theta)} m(A) = 1, \; m(\emptyset) = 0. \tag{5}$$

If $m(A) > 0$, the set A is called a focal element. The set of all the focal elements corresponds to a body of evidence.

- *Belief function.* Given a mass function m, a belief function is defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B), \tag{6}$$

where $Bel(A)$ reflects the total belief degree of A.

B. Spatio-temporal Dempster-Shafer information fusion

1) Weighted Dempster-Shafer theory: According to the traditional Dempster-Shafer theory, all the information sources are equally weighted. However, due to noise or errors, information sources may have different properties and confidence values, so their outputs should have different weights [29], [38]. Specifically, let h_n be an existing information source, w_n be the confidence weight of h_n, m_n be the original mass function of h_n, m_n^* be the weighted mass function of h_n, and $P(\Theta)$ (referred to in Eq. (4)) be the power set of h_n. Mathematically, m_n^* is defined as:

$$m_n^*(A) = \left\{ \begin{array}{cl} 0, & A = \emptyset; \\ 1 - w_n \left(1 - m_n(\Theta)\right), & A = \Theta; \\ w_n m_n(A), & A \in P(\Theta) \setminus \{\Theta, \emptyset\}. \end{array} \right. \tag{7}$$

The proof that m_n^* is a mass function is given as follows. According to Eq. (5), we have the following relation:

$$m_n(\Theta) + \sum_{A \in P(\Theta) \setminus \Theta} m_n(A) = 1. \tag{8}$$

Based on this relation, the sum of m_n^* over $P(\Theta)$ can be formulated as:

$$\begin{aligned} \sum_{A \in P(\Theta)} m_n^*(A) &= m_n^*(\Theta) + \sum_{A \in P(\Theta) \setminus \Theta} m_n^*(A) \\ &= 1 - w_n \left(1 - m_n(\Theta)\right) + w_n \sum_{A \in P(\Theta) \setminus \Theta} m_n(A) = 1. \end{aligned} \tag{9}$$

2) Application to tracking: In practice, how to effectively obtain the confidence weights (i.e., w_n) of different information sources is another key issue to solve. In the field of visual tracking, the data taken from different frames are likely to have different statistical properties due to the influence of the extrinsic environmental conditions (e.g., illumination) and the intrinsic object-specific factors (e.g., shape deformation, rotation, or pose variation), as shown in Fig. 1. From Fig. 1, we see that the image patches from sequential frames usually have different but correlated appearance properties, as the targets often move in a continuous, dynamic, and periodical manner.

Even if the data are taken from the same object region but

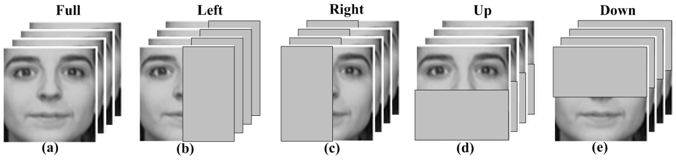

Fig. 2. Illustration of the spatial block-division strategy with five sub-regions (i.e., full, left, right, up, and down) over the ensemble of data samples. The image regions occluded by the masks are not used during tracking.

different locations, they are also likely to possess different statistical properties. Thus, we need to consider both space-related and time-related information about object appearance for robust visual tracking. These two types of information should be associated with appropriate confidence weights according to their spatial or temporal properties. Motivated by this observation, we propose a weighted information fusion mechanism for combining the belief evidences from different spatio-temporal SVMs.

3) Spatio-temporal weighted SVM evidence combination: During visual tracking, discriminative information from different frames and different spatial regions within each frame constitutes different sources for object or non-object classification. As in [3], we divide each sample into five spatially related sources: $\{$*full*, *left*, *right*, *up*, *down*$\}$, as shown in Fig. 2. Each subdivision is indexed by a number $l \in \{1, 2, 3, 4, 5\}$. Clearly, these five sources contain different spatial salience information of the appearance inside each object sub-region. In order to capture this discriminative information, a SVM classifier is learned for each sub-region. Thus, each frame is associated with five SVM classifiers that act as space-related discriminative information sources. For a single sub-region, the corresponding SVM classifiers from different frames can be considered as a sequence of time-related discriminative information sources. Thus, we have a set of spatio-temporal discriminative information sources which we combine using weighted Dempster-Shafer information fusion. Fig. 3 gives an intuitive illustration of the proposed spatio-temporal SVM evidence combination scheme.

At this point we introduce some terminology needed to describe space and time related information fusion. Suppose that there are N time-adjacent sub-sequences, each of which is associated with the five source-specific SVM classifiers denoted as $\{h_{ln}\}$ for $1 \leq l \leq 5$ and $1 \leq n \leq N$. Thus, there are two dimensions, spatial and temporal, for all the SVM classifiers. These SVM classifiers work as spatio-temporal discriminative information sources during visual tracking. As visual tracking is solved as a binary classification problem, we define the corresponding power set of the spatio-temporal discriminative information sources as:

$$P(\Theta) = \{\emptyset, \{-1\}, \{1\}, \Theta\}, \tag{10}$$

where $\Theta = \{-1, 1\}$ denotes both classes, $\{1\}$ represents the object class, and $\{-1\}$ stands for the non-object class. In order to effectively combine these space-related and time-related discriminative information sources, we introduce the

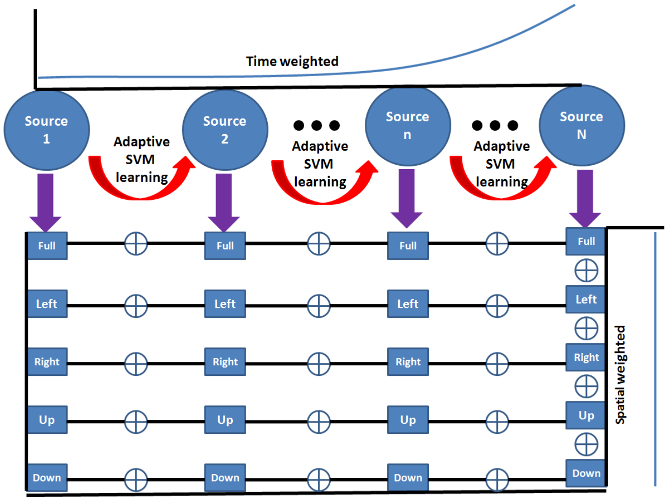

Fig. 3. Illustration of the spatial-temporal weighted Dempster-Shafer information fusion.

Dempster's orthogonal fusion rule:

$$\begin{aligned} m(A) &= (m_1 \bigoplus m_2 \bigoplus \cdots \bigoplus m_N)(A) \\ &= \begin{cases} \frac{\sum\limits_{\bigcap_{n=1}^{N} A_n = A} \left(\prod_{n=1}^{N} m_n(A_n)\right)}{1-K}, & \text{if } A \neq \emptyset; \\ 0, & \text{otherwise;} \end{cases} \end{aligned} \tag{11}$$

where $\bigoplus$ is the combination operator, m_n is the corresponding mass function of $A_n \in P(\Theta)$, K is a probability mass measuring the degree of the conflict among the N mass functions (i.e., $\{m_n\}_{n=1}^{N}$), and the term $1 - K$ is a normalization factor. $K = 0$ indicates that there is no conflict among $\{m_n\}_{n=1}^{N}$, while $K = 1$ implies that $\{m_n\}_{n=1}^{N}$ are completely contradictory to each other. Mathematically, K can be formulated as:

$$K = \sum_{\bigcap_{n=1}^{N} A_n = \emptyset} \left(\prod_{n=1}^{N} m_n(A_n) \right). \tag{12}$$

Time-related combination. By concatenating the SVM classifiers of a particular spatial region along the temporal dimension, we obtain a set of time-related SVM classifiers denoted as $\{h_{ln}\}_{n=1}^{N}$. Let $\{m_{ln}\}_{n=1}^{N}$ be the corresponding mass functions of $\{h_{ln}\}_{n=1}^{N}$. Mathematically, the mass function m_{ln} is defined as:

$$m_{ln}(A_n) = \begin{cases} 0, & A_n = \emptyset; \\ \zeta, & A_n = \Theta; \\ (1-\zeta) g_{ln}(\mathbf{x}_t^l), & A_n = \{1\}; \\ (1-\zeta)(1 - g_{ln}(\mathbf{x}_t^l)), & A_n = \{-1\}; \end{cases} \tag{13}$$

where ζ is the class uncertainty degree ($\zeta = 0.1$ in the experiments), $\mathbf{x}_t^l$ is a candidate sample associated with the l-th spatial block-division sub-region at time t, and $g_{ln}(\mathbf{x}_t^l)$ is the normalized SVM confidence score defined as:

$$g_{ln}(\mathbf{x}_t^l) = \frac{1}{1 + \exp(-\gamma h_{ln}(\mathbf{x}_t^l))}. \tag{14}$$

Here, γ is a scaling factor ($\gamma = 0.6$ in the experiments), and $h_{ln}(\mathbf{x}_t^l)$ is the associated prediction function of the SVM clas-

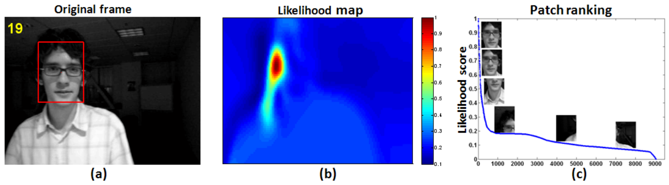

Fig. 4. Demonstration of the likelihood evaluation based on spatio-temporal Dempster-Shafer information fusion using $Bel(\{1\})$. (a) shows the original frame; (b) displays a likelihood map, each element of which corresponds to an image patch in the entire image search space; and (c) exhibits the curve of the likelihood ranking for all the image patches.

sifier h_{ln}. Note that there is still some uncertainty associated with a candidate sample even if the normalized confidence score g_{ln} takes the value of 1.0. Since visual tracking is a time-varying process, the SVM classifier h_{ln} is associated with a time-varying weight defined as: $\exp(-\frac{N-n}{N})$. Hence, the previous discriminative information can be forgotten gradually. According to Eq. (7), we define the time-weighted mass function m^*_{ln} of m_{ln} as:

$$m^*_{ln}(A_n) = \begin{cases} 0, & A_n = \emptyset; \\ 1 - \exp(-\frac{N-n}{N})(1-\zeta), & A_n = \Theta; \\ \exp(-\frac{N-n}{N})(1-\zeta)g_{ln}(\mathbf{x}^l_t), & A_n = \{1\}; \\ \exp(-\frac{N-n}{N})(1-\zeta)(1-g_{ln}(\mathbf{x}^l_t)), & A_n = \{-1\}. \end{cases} \tag{15}$$

According to Eq. (11), the combined mass function $\mathbb{M}_l$ of $\{m^*_{ln}\}^N_{n=1}$ is formulated as:

$$\begin{aligned} \mathbb{M}_l(A) &= (m^*_{l1} \bigoplus m^*_{l2} \bigoplus \cdots \bigoplus m^*_{lN})(A) \\ &= \begin{cases} \dfrac{\sum_{\cap^N_{n=1} A_n = A} \left(\prod^N_{n=1} m^*_{ln}(A_n)\right)}{1 - \sum_{\cap^N_{n=1} A_n = \emptyset} \left(\prod^N_{n=1} m^*_{ln}(A_n)\right)}, & \text{if } A \neq \emptyset; \\ 0, & \text{otherwise}; \end{cases} \end{aligned} \tag{16}$$

where $A, A_1, A_2, \ldots, A_N \in P(\Theta)$. By extension, we obtain all the combined mass functions corresponding to the five spatial sub-regions: $\{\mathbb{M}_l(A)\}^5_{l=1}$.

Space-related combination. Following discriminative information fusion from multiple timesteps, we now aim to fuse the spatial-related discriminative information from the combined mass functions $\{\mathbb{M}_l(A)\}^5_{l=1}$, which can be viewed as five spatial-related discriminative information sources. We assign equal weights to these sources because of their similar spatial configurations. Consequently, the corresponding confidence weights of $\{\mathbb{M}_l(A)\}^5_{l=1}$ are uniform. In this case, we have the following spatial-weighted mass function $\mathbb{M}^*_l(A)$ of $\mathbb{M}_l(A)$:

$$\mathbb{M}^*_l(A) = \begin{cases} 0, & A = \emptyset; \\ \zeta, & A = \Theta; \\ (1-\zeta)\frac{\mathbb{M}_l(A)}{1-\mathbb{M}_l(\Theta)}, & A = \{1\} \text{ or } \{-1\}. \end{cases} \tag{17}$$

According to Eq. (11), the final combined mass function $\mathcal{M}(A)$ of $\{\mathbb{M}_l(A)\}^5_{l=1}$ is formulated as:

$$\begin{aligned} \mathcal{M}(A) &= (\mathbb{M}^*_1 \bigoplus \mathbb{M}^*_2 \bigoplus \cdots \bigoplus \mathbb{M}^*_5)(A) \\ &= \begin{cases} \dfrac{\sum_{\cap^5_{l=1} A_l = A} \left(\prod^5_{l=1} \mathbb{M}^*_l(A_l)\right)}{1 - \sum_{\cap^5_{l=1} A_l = \emptyset} \left(\prod^5_{l=1} \mathbb{M}^*_l(A_l)\right)}, & \text{if } A \neq \emptyset; \\ 0, & \text{otherwise}; \end{cases} \end{aligned} \tag{18}$$

where $A, A_1, A_2, \ldots, A_5 \in P(\Theta)$. According to Eq. (6), the belief function $Bel(A)$ associated with $\mathcal{M}(A)$ is defined as:

$$\begin{aligned} Bel(A) &= \textstyle\sum_{B \subseteq A} \mathcal{M}(B) \\ &= \begin{cases} 0, & A = \emptyset; \\ \mathcal{M}(\{-1\}) + \mathcal{M}(\{1\}) + \mathcal{M}(\Theta), & A = \Theta; \\ \mathcal{M}(\{1\}), & A = \{1\}; \\ \mathcal{M}(\{-1\}), & A = \{-1\}. \end{cases} \end{aligned} \tag{19}$$

According to Dempster-Shafer, $Bel(A)$ reflects the degree of belief when the decision A is made, in the interval $[0, 1]$. $Bel(\{1\})$ is the belief that the test sample belongs to the object class. Therefore, $Bel(\{1\})$ is used for evaluation of the likelihood $p(\mathbf{o}_t|\mathbf{Z}_t)$ in Eq. (1), taking into account weighted discriminative information from N frames and five spatial subdivisions.

Fig 4 illustrates the effectiveness of $Bel(\{1\})$ as a likelihood function. As shown in Fig. 4(a), a bounding box highlighted in red is shifted pixel by pixel from left to right and from top to bottom. After calculating the normalized likelihood score using Eq. (19) at each location, we have a likelihood map which is shown in Fig. 4(b). From Fig. 4(b), we see that the likelihood map has an obvious peak, which indicates that the proposed observation model is able to discriminative the object of interest from the rest of the image. For an intuitive illustration, we compute the corresponding likelihood scores of all the rectangular image patches, then sort them in a descending order, and finally show them in Fig. 4(c). Clearly, the larger the likelihood score is, the more likely the rectangular image patch belongs to the foreground object class.

C. Adaptive SVM learning across sources

In general, the time-adjacent sources for visual tracking are temporally correlated with each other. We design an adaptive SVM learning scheme in order to make use of this correlation by transferring discriminative information between the SVM classifiers that occur in adjacent sources.

To describe the ASL scheme, we introduce the following notations. Let $\mathcal{D}^l_{n-1}$ denote the previous source associated with the SVM classifier $h_{l(n-1)}(\mathbf{x})$. Corresponding to $\mathcal{D}^l_{n-1}$, the training data generated from the current source $\mathcal{D}^l_n$ are denoted as $\{(\mathbf{x}^l_{ni}, y^l_{ni})\}^{J^l_n}_{i=1}$, where $\mathbf{x}^l_{ni}$ is the i-th data vector, $y^l_{ni} \in \{-1, 1\}$ is its binary label, and J^l_n is the training data size. The ASL scheme tries to effectively learn a corresponding SVM classifier $h_{ln}(\mathbf{x})$ of $\mathcal{D}^l_n$ using $h_{l(n-1)}(\mathbf{x})$ and $\{(\mathbf{x}^l_{ni}, y^l_{ni})\}^{J^l_n}_{i=1}$. In other words, the learned SVM classifier $h_{ln}(\mathbf{x})$ needs to not only account for the historical discriminative information from $h_{l(n-1)}(\mathbf{x})$ but also adapt to the current training data $\{(\mathbf{x}^l_{ni}, y^l_{ni})\}^{J^l_n}_{i=1}$. Mathematically, the ASL scheme

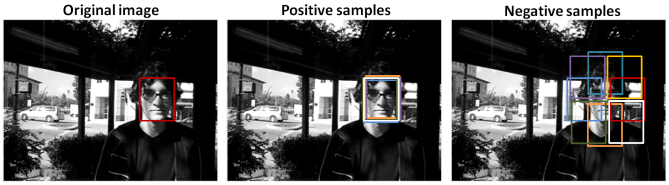

Fig. 5. Illustration of training sample selection. The left subfigure plots the bounding box corresponding to the current tracker location; the middle subfigure shows the selected positive samples; and the right subfigure displays the selected negative samples. Different colors are assoicated with different samples.

can be formulated as the following optimization problem:

$$\begin{array}{ll}\max\limits_{\boldsymbol{\alpha}} & \sum_{i=1}^{J_n^l}(1-\lambda_i)\alpha_i - \frac{1}{2}\sum_{i=1}^{J_n^l}\sum_{j=1}^{J_n^l}\alpha_i\alpha_j y_{ni}^l y_{nj}^l K(\mathbf{x}_{ni}^l, \mathbf{x}_{nj}^l), \\ \text{s.t.} & 0 \leq \alpha_i \leq C, \, \forall i \end{array} \tag{20}$$

where $\lambda_i = y_{ni}^l h_{l(n-1)}(\mathbf{x}_{ni}^l)$, C is a regularization factor, and $K(\mathbf{x}, \mathbf{x}')$ is a kernel function. The optimization problem (20) can be efficiently solved by using the iterative parameter learning approach [14], which first chooses working variables and then optimizes them until convergence. As a result, we have the SVM classifier $h_{ln}(\mathbf{x}) = \sum_{i=1}^{J_n^l}\alpha_i y_{ni}^l K(\mathbf{x}, \mathbf{x}_{ni}^l)$.

D. Feature description

The kernel function $K(\mathbf{x}_a, \mathbf{x}_b)$ (referred to in Eq. (20)) is defined as:

$$K(\mathbf{x}_a, \mathbf{x}_b) = \exp(-\beta\|\mathbf{F}(O_a) - \mathbf{F}(O_b)\|^2), \tag{21}$$

i.e. a Gaussian RBF kernel, where $\|\cdot\|$ is the ℓ_2 norm, β is a scaling factor, and **F** is a feature descriptor.

In our implementation, the feature descriptor **F** is associated with a Radon matrix [15], but other features may equally be used. The Radon matrix $T_{\mathfrak{R}}$ can be derived by directly applying the Radon transform [15] to an image patch O after affine warping and histogram equalization:

$$T_{\mathfrak{R}}(\rho, \theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} O_{xy}\delta(x\cos\theta + y\sin\theta - \rho)\mathrm{d}x\mathrm{d}y, \tag{22}$$

where $\delta(\cdot)$ is the Dirac delta-function, $\theta \in [0, \pi]$, and $\rho \in [-\infty, +\infty]$. $T_{\mathfrak{R}}(\rho, \theta)$ is the integral of O over the line $\rho = x\cos\theta + y\sin\theta$. Consequently, the Radon matrix $T_{\mathfrak{R}}$ encodes the spatial integral information of O in different directions. In practice, the angle θ takes several discrete values which are uniformly sampled from the interval $[0, \pi]$ ($\theta \in \{\frac{k\pi}{6}\}_{k=1}^{6}$ in the experiments), and the radial coordinate ρ lies in a particular interval $[\rho_{\min}, \rho_{\max}]$ determined by the image size. In this case, we obtain a Radon feature vector for the image region O, i.e., $\mathbf{F}_{\mathfrak{R}}(O) = \mathbb{U}(T_{\mathfrak{R}})$ with $\mathbb{U}(\cdot)$ being a row flattening operator. A separate feature is calculated for each spatial subdivision of the image patch.

V. Experiments

A. Data description and implementation details

In the experiments, we evaluate the proposed tracker (referred to as STWDS) on twenty video sequences, which are taken in different scenes and composed of 8-bit grayscale or 24-bit color images. In these videos, several complicated factors cause drastic appearance changes of a tracked object, including illumination variation, occlusion, out-of-plane rotation, background distraction, small target size, motion blurring, pose variation and so on. In order to demonstrate the effectiveness of the proposed tracker on these videos, a number of experiments are conducted. Such experiments have two main purposes: to verify the robustness of the proposed STWDS in various challenging situations, and to evaluate the ability of STWDS to adapt to complex appearance changes.

As mentioned in Section III, tracking is based on a pool of samples that is updated at each frame. Like [9], we take a spatial distance-based strategy for training sample selection. Namely, the image regions from a small neighborhood around the object location are selected as positive samples, and the negative samples are generated by selecting the image regions which are relatively far from the object location. Specifically, we draw a number of samples $\mathcal{Z}_t$ from Eq. (2), and then an ascending sort for the samples from $\mathcal{Z}_t$ is made according to their spatial distances to the current object location, resulting in a sorted sample set $\mathcal{Z}_t^s$. By selecting the first few samples from $\mathcal{Z}_t^s$, we have a subset $\mathcal{Z}_t^+$ that is the final positive sample set, as shown in the middle part of Fig. 5. The negative sample set $\mathcal{Z}_t^-$ is generated in the area around the current tracker location, as shown in the right part of Fig. 5. In the first frame, the object location is manually labeled.

The proposed STWDS is implemented in Matlab on a workstation with an Intel Core 2 Duo 2.66GHz processor and 3.24G RAM. The average running time of the proposed STWDS is about 0.3 second per frame. The number of samples in each frame (i.e., V) is set to 200. The parameters W and Q in Algorithm 1 are set to 6 and 10, respectively. The scaling factor β defined in Eq. (21) is set to 0.01. These parameters remain the same throughout all the experiments.

B. Competing trackers

We compare STWDS with several state-of-the-art trackers both qualitatively and quantitatively. These trackers are all recently proposed, and have had significant impact on the visual tracking community. For descriptive convenience, they are respectively referred to as FragT (Fragment-based tracker [1]), MILT (multiple instance boosting-based tracker [9]), VTD (visual tracking decomposition [26]), OAB (online AdaBoost [5]), IPCA (incremental PCA [2]), L1T (ℓ_1 minimization tracker [4]), and CSVM (conventional SVM tracker without multi-source discriminative learning). In implementation, we directly use the publicly available source codes of FragT, MILT, VTD, OAB, IPCA, and L1T. In the experiments, there are two different versions for OAB, i.e., OAB1 and OAB5. They use two different search radiuses (i.e., r=1 and r=5 selected in the same way as [9]) to generate the training samples for learning AdaBoost classifiers, respectively. In implementation, the proposed STWDS adopts the same initialization strategy (i.e., manual annotation with a bounding box) as the competing trackers.

The reasons for selecting these competing trackers are as follows. CSVM is close to STWDS but does not use Dempster

Shafer information fusion to fuse multiple SVM outputs, instead using a self-learning strategy to train and update the SVM classifiers in a single source. MILT is a recently proposed discriminant learning-based tracker, which uses multiple instance boosting for object/non-object classification. In comparison, OAB utilizes online boosting for object/non-object classification. Thus, comparing STWDS with MILT and OAB can demonstrate the discriminative capabilities of STWDS in handling large appearance variations. FragT is a fragment-based visual tracker using the integral histogram. By combining the vote maps of the multiple patches, FragT captures the spatial layout information of the object region, resulting in the robustness to partial appearance changes. IPCA uses incremental principal component analysis to construct the eigenspace-based observation model for visual tracking. L1T treats visual tracking as a sparse approximation problem using ℓ_1-regularized minimization. VTD uses sparse principal component analysis to decomposes the observation (or motion) model into a set of basic observation (or motion) models, each of which covers a specific type of object appearance (or motion). Thus, comparing STWDS with FragT, IPCA, L1T, and VTD can show their capabilities of tolerating complicated appearance changes.

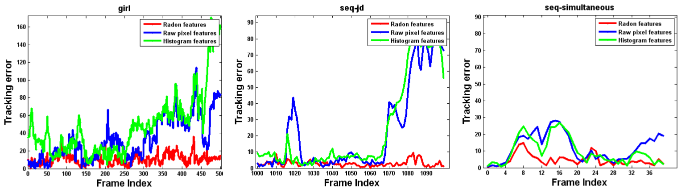

Fig. 6. Quantitative tracking performances of the tracking approaches using different visual features on the three video sequences. Clearly, the tracking approach using the Radon feature achieves the best tracking results.

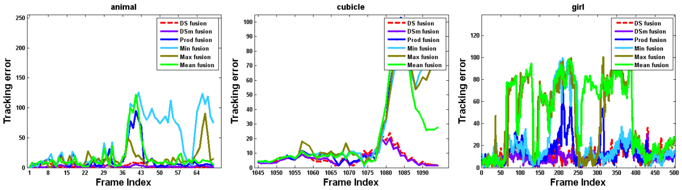

Fig. 7. Quantitative tracking performances of the tracking approaches using different information fusion methods on the three video sequences, including Dempster-Shafer (DS) fusion, Dezert-Smarandache (DSm) fusion, Prod fusion, Min fusion, Max fusion, and Mean fusion. Clearly, the tracking approaches using DS fusion and DSm fusion achieve better tracking performances. Moreover, the tracking results obtained by DS fusion is on par with those obtained by DSm fusion.

C. Quantitative evaluation criteria

The object center locations of the eighteen video sequences are manually labeled and used as the ground truth. Hence, we can quantitatively evaluate the tracking performances of the nine trackers by computing their corresponding pixel-based tracking location errors (TLEs) with respect to the ground truth of the twenty video sequences.

To measure the tracking accuracy of each tracker, we define a quantitative evaluation criterion: $\mathrm{TLEWH} = \mathrm{TLE}/\max(\mathrm{W}, \mathrm{H})$. Here, TLE is the above-mentioned pixel-based tracking location error with respect to the ground truth, W is the width of the ground truth bounding box for object localization, and H is the height of the ground truth bounding box. For quantitative comparison, we compute the average TLEWH of each tracker on each video sequence. The smaller the average TLEWH, the more accurate tracking results the tracker achieves.

In order to better evaluate the quantitative tracking performance of each tracker, we introduce a per-frame criterion to judge whether a tracker succeeds in object tracking. Namely, if TLEWH is smaller than 0.25, the tracking result for each frame is considered to be successful. We compute the tracking success rate (defined as $\frac{\#\text{Success frames}}{\#\text{Total frames}}$) for each video sequence to quantitatively evaluate the performance of each tracker.

D. Evaluation of features and information fusion methods

In order to evaluate the effect of feature description, we embed 3 different visual features into our tracking approach: the raw pixel feature, the intensity histogram feature, and the Radon feature. Fig. 6 shows the tracking error curves of the tracking approaches using different visual features on the three video sequences (i.e., “girl”, “seq-jd”, and “seq-simultaneous”). From Fig. 6, we see that the tracking approach using the Radon feature achieves better tracking performance, particularly in the presence of large rotation.

We also conduct an experiment comparing six different information fusion methods including Dempster-Shafer (DS) information fusion, Dezert-Smarandache (DSm) information fusion [37], and simple information fusion methods (i.e., the maximum, minimum, average, and product of the confidence scores from different information sources). By embedding these information fusion methods into our tracking approach, we can evaluate their quantitative tracking performances on the three video sequences (i.e., “animal”, “cubicle”, “girl”), as shown in Fig. 7. Clearly, it is seen from Fig. 7 that the tracking approaches using Dempster-Shafer and Dezert-Smarandache information fusion obtain better tracking results than those using the simple information fusion methods. In addition, the tracking performance achieved by Dempster-Shafer information fusion is on par with that by Dezert-Smarandache information fusion. As an extension of the Dempster-Shafer theory, the Dezert-Smarandache theory mainly focuses on the fusion of uncertain, highly conflicting, and imprecise quantitative or qualitative sources of evidence. In our case, the spatio-temporal sources of evidence for visual tracking are mutually correlated and lowly conflicting. As a result, the Dezert-Smarandache theory for information fusion performs comparably to the Dempster-Shafer theory during tracking.

E. Comparison with and without adaptive SVM learning

We compare the quantitative tracking performance of our approach with and without using adaptive SVM learning. These two variants make use of the same multi-source discriminative learning scheme. The only difference between them is

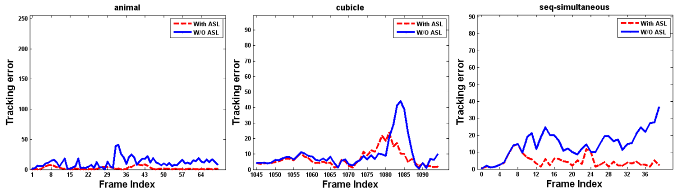

Fig. 8. Quantitative tracking performances of the tracking approaches with and without adaptive SVM learning (ASL) on the three video sequences. Clearly, the tracking approach with ASL achieves better tracking results.

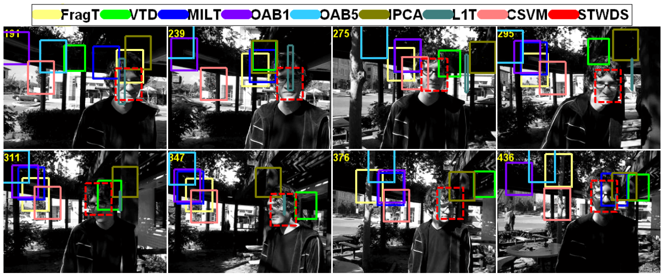

Fig. 9. The tracking results of the nine trackers over the representative frames (191, 239, 275, 295, 311, 347, 376, 436) of the "*trellis70*" video sequence in the scenarios with drastic illumination changes and head pose variations.

that the former utilizes adaptive SVM learning to generate the multi-source discriminative information while the latter takes advantage of standard SVM learning. Fig. 8 shows the frame-by-frame center location error curves of the two tracking approaches (i.e., with and without adaptive SVM learning) on the three video sequences. From Fig. 8, we see that the tracking approach using adaptive SVM learning obtains more accurate tracking results, especially for later frames in each video.

F. Qualitative comparison results with competing trackers

Due to the space limit, we provide the corresponding tracking results of the nine trackers (highlighted by the bounding boxes in different colors) over the representative frames of the four video sequences, as shown in Figs. 9–12. The complete tracking results for all the video sequences can be found in the supplementary file.

1) Illumination changes: Video sequences "*trellis70*" (Fig. 9), "*car11*", "*car4*", and "*shaking*" demonstrate tracker performance in the presence of drastic illumination changes.

As a representative example, in "*trellis70*", VTD and OAB5 begin to lose the face after shadowing at frame 170 while OAB1, IPCA, MILT, and FragT fail to track the face after further lighting changes at frames 182, 201, 202, and 205 respectively. L1T begins to lose the face from frame 252. After frame 123, CSVM greatly drifts away from the true location of the face. Only the proposed STWDS successfully tracks the face throughout each of these sequences.

2) Pose and viewpoint changes: Video sequences "*seq-simultaneous*" (Fig. 10), "*animal*", "*football2*", and "*girl*" test tracker performance for rapid object rotation and translation, sometimes in combination with other factors such as occlusion or motion blur.

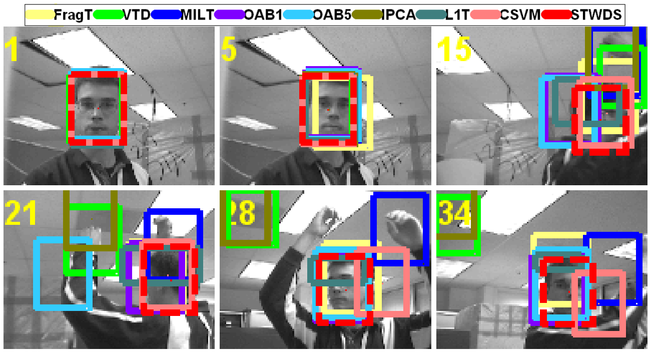

Fig. 10. The tracking results of the nine trackers over the representative frames (1, 5, 15, 21, 28, 34) of the "*seq-simultaneous*" video sequence in the scenarios with partial occlusion, out-of-plane rotation, and head pose variation.

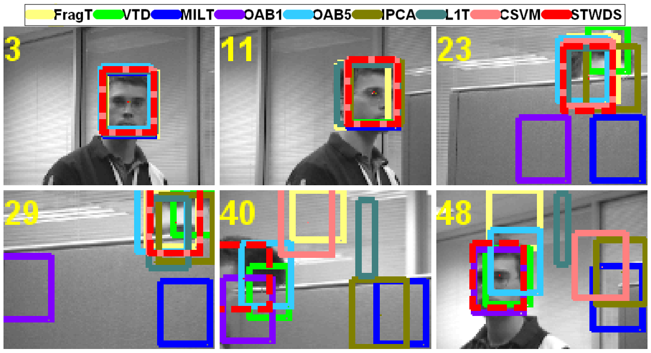

Fig. 11. The tracking results of the nine trackers over the representative frames (3, 11, 23, 29, 40, 48) of the "*cubicle*" video sequence in the scenarios with partial occlusion, out-of-plane rotation, and head pose variation.

For example, in "*seq-simultaneous*", IPCA, VTD, and MILT lose the head of the person thoroughly after frames 19, 20, and 27, respectively. In comparison, L1T, MILT, and OAB5 achieve inaccurate tracking results from early in the video. Before frame 26, CSVM tracks the head inaccurately, and then loses the head entirely after frame 26. Suffering from out-of-plane rotation and occlusion, OAB1 and FragT drift away multiple times. However, the proposed STWDS succeeds in tracking the head throughout this video sequence and others involving significant pose change.

3) Occlusion: Video sequences "*cubicle*" (Fig. 11), "*woman*", "*seq-jd*", and "*girl*" demonstrate tracker performance in the face of severe occlusion, again in conjunction with other factors such as blur and pose variation.

For example, in "*cubicle*", the target head is partially occluded for most frames. IPCA, FragT, L1T, MILT, and OAB1 fail to track the head of the man after frame 38, 38, 39, 15, and 16, respectively. OAB5 and VTD achieve inaccurate tracking results throughout the video sequence. CSVM begins to lose the head after frame 41. However, the proposed STWDS successfully tracks the head in all frames over this and the other sequences.

4) Blur, clutter, and low video quality: Video sequences "*video-car*" (Fig. 12), "*animal*", and "*shaking*" test performance in cases where the depiction of the target is of low quality, either due to motion blur or small size.

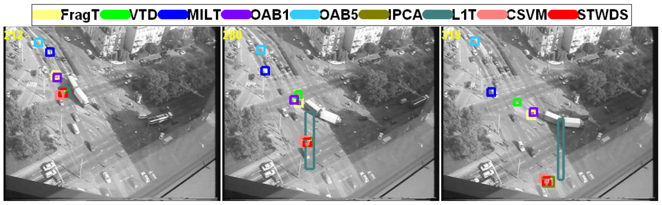

Fig. 12. The tracking results of the nine trackers over the representative frames (212, 280, 319) of the "*video-car*" video sequence in the scenarios with small target size and background clutter.

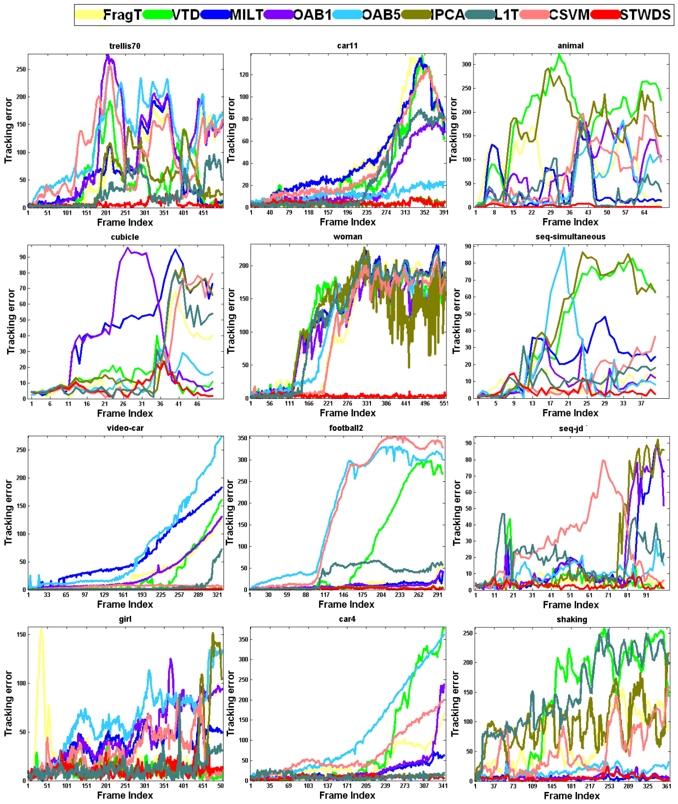

Fig. 13. The tracking location error plots obtained by the nine trackers over the first twelve videos. In each sub-figure, the x-axis corresponds to the frame index number, and the y-axis is associated with the tracking location error.

For example, the sequence "*video-car*" is a cross-road traffic scene from a top-down viewpoint. In this sequence, we track a car that occupies a small number of pixels, is densely surrounded by other cars, and is quite blurred. Due to the influence of the background distraction and the small target size, MILT, OAB5, FragT, OAB1, VTD, and L1T start to lose the car from frames 69, 160, 190, 196, 246, and 313, respectively. On the contrary, IPCA, CSVM, and STWDS can track the car persistently. Among these three trackers, STWDS is able to locate the car most accurately in all frames.

G. Quantitative comparison results with competing trackers

Fig. 13 shows the tracking location error plots obtained by the nine trackers (highlighted in different colors) for the first twelve video sequences. We also compute the mean and standard deviation of the tracking location errors in the experiments, and report the results in Fig. 14. From Fig. 13 and Fig. 14, we see that the proposed STWDS achieves the most robust and accurate tracking performance on most video sequences.

Table I shows the average TLEWHs of the nine trackers on the total twenty video sequences. It is clear that the proposed STWDS achieves the best tracking performances on most video sequences. Table II reports the quantitative tracking results of the nine trackers in the tracking success rate on the total twenty video sequences. From Table II, we can see that the proposed STWDS achieves the best tracking performances on most video sequences.

VI. Conclusion

In this paper, we have proposed a spatio-temporal weighted Dempster-Shafer (STWDS) scheme for combining discriminative information from different sources. In the STWDS scheme, we introduce multi-source discriminative learning to capture the spatio-temporal correlations among different discriminative information sources. Thus, the problem of visual tracking is converted to that of discriminative learning in a sequence of space-related and time-adjacent sources, each of which is associated with a discriminative information source for object/non-object classification. Furthermore, an adaptive SVM learning scheme is designed to transfer discriminative information across time-adjacent sources, which are assumed to have some correlation. Based on the associated Dempster-Shafer belief function of the STWDS scheme, an observation model is constructed and embedded into a visual tracking method. Compared with several state-of-the-art trackers, the proposed tracker is shown to be more robust to illumination changes, pose variations, partial occlusions, background distractions, motion blurring, as well as other complicated appearance changes.

In the future, we plan to extend the method to deal with multiple targets simultaneously by using a multi-class SVM to discriminate between each object and the background. Dempster-Shafer fusion extends naturally to accommodate this by adding extra labels into the set Θ.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp.798-805, 2006.

[2] D. A. Ross, J. Lim, R. Lin, and M. Yang. "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, Vol. 77, Iss. 1-3, pp.125-141 2008.

[3] X. Li, W. Hu, Z. Zhang, X. Zhang, M. Zhu, and J. Cheng, "Visual Tracking Via Incremental Log-Euclidean Riemannian Subspace Learning," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[4] X. Mei and H. Ling, "Robust Visual Tracking using ℓ_1 Minimization," in *Proc. IEEE International Conference on Computer Vision*, pp. 1436-1443, 2009.

[5] H. Grabner, M. Grabner, and H. Bischof, "Real-time Tracking via On-line Boosting," in *Proc. British Machine Vision Conference*, pp. 47-56, 2006.

[6] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised On-line Boosting for Robust Tracking," in *Proc. European Conference on Computer Vision*, pp.234-247, 2008.

[7] S. Avidan, "Support Vector Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 26, Iss. 8, pp. 1064-1072, 2004.

[8] M. Tian, W. Zhang, and F. Liu, "On-Line Ensemble SVM for Robust Object Tracking," in *Proc. Asian Conference on Computer Vision*, pp. 355-364, 2007.

[9] B. Babenko, M. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 983-990, 2009.

[10] S. Avidan, "Ensemble Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 29, Iss. 2, pp. 261-271, 2007.

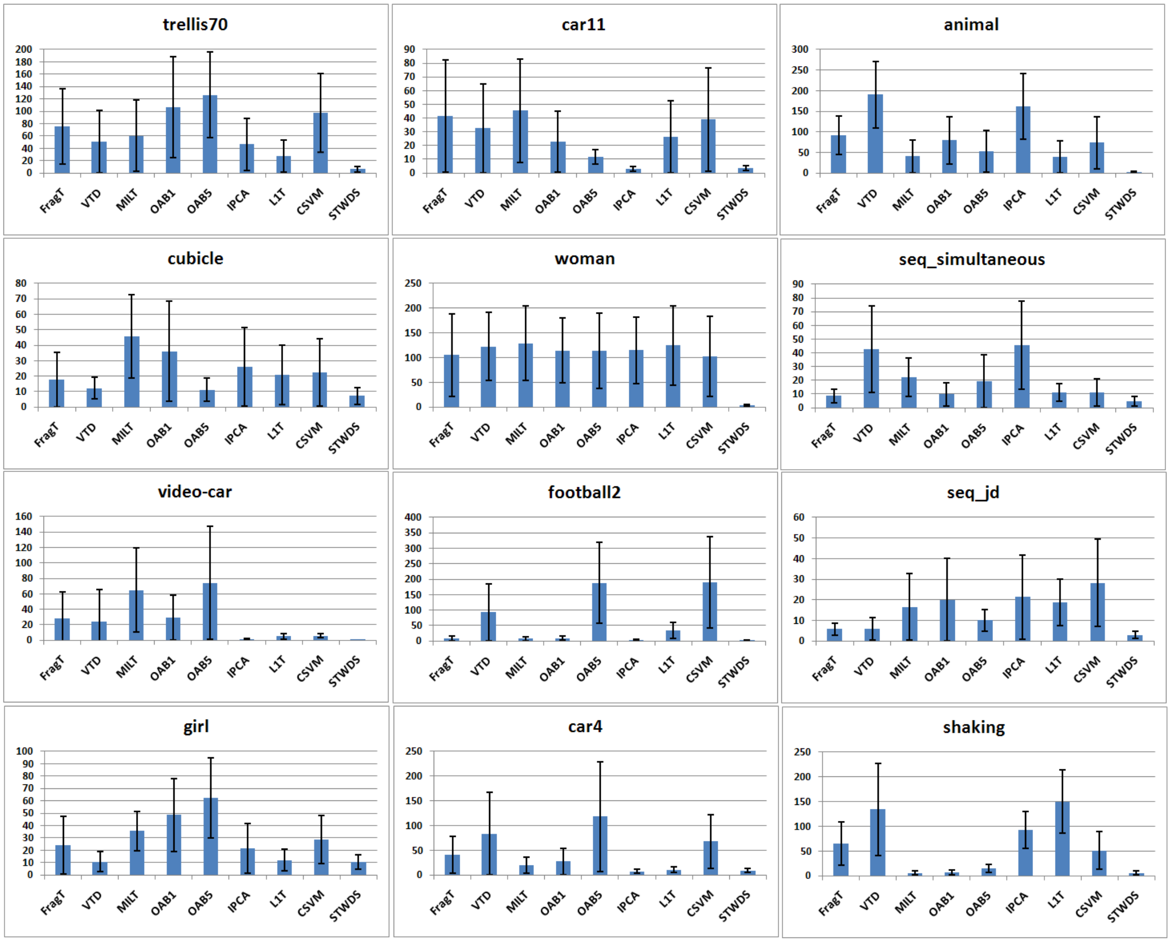

Fig. 14. The quantitative comparison results of the nine trackers over the first twelve videos. The figure reports the mean and standard deviation of their tracking location errors over the first twelve videos. In each sub-figure, the x-axis shows the competing trackers, the y-axis is associated with the mean of their tracking location errors, and the error bars correspond to the standard deviation of their tracking location errors.

[11] X. Li, A. Dick, H. Wang, C. Shen, A. van den Hengel, "Graph mode-based contextual kernels for robust SVM tracking," in *Proc. IEEE International Conference on Computer Vision,*, pp.1156-1163, 2011.

[12] R. T. Collins, Y. Liu, and M. Leordeanu, "Online Selection of Discriminative Tracking Features," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 27, Iss.10, pp.1631-1643, 2005.

[13] X. Liu and T. Yu, "Gradient Feature Selection for Online Boosting," in *Proc. IEEE International Conference on Computer Vision,* pp. 1-8, 2007.

[14] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. ACM Multimedia*, pp. 188-197, 2007.

[15] Stanley R. Deans, "The Radon Transform and Some of Its Applications," *John Wiley & Sons*, New York, 1983.

[16] Y. Wu and T.S. Huang, "Robust Visual Tracking by Integrating Multiple Cues Based on Co-Inference Learning," *International Journal of Computer Vision*, Vol. 58, Iss. 1, pp. 55-71, 2004.

[17] G. Hua and Y. Wu, "Measurement Integration Under Inconsistency For Robust Tracking, in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[18] Z. Fan, M. Yang, and Y. Wu, "Multiple Collaborative Kernel Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 29, pp. 1268-1273, 2007.

[19] F. Tang, S. Brennan, Q. Zhao, and H. Tao, "Co-Tracking Using Semi-Supervised Support Vector Machines," in *Proc. IEEE International Conference on Computer Vision*, 2007.

[20] B. Han and L. S. Davis, "Probabilistic Fusion-based Parameter Estimation for Visual Tracking," *Computer Vision and Image Understanding*, Vol. 113, Iss. 4, pp. 435-445, 2009.

[21] J. Xu, J. Yuan, and Y. Wu, "Multimodal Partial Estimates Fusion," in *Proc. IEEE International Conference on Computer Vision*, 2009.

[22] Y. Wu, E. Blaschz, G. Chen, L. Bai, H. Ling "Multiple Source Data Fusion via Sparse Representation for Robust Visual Tracking," in *Proc. International Conference on Information Fusion*, 2011.

[23] H. Liu and F. Sun, "Fusion Tracking in Color and Infrared Images Using Sequential Belief Propagation," in *Proc. IEEE International Conference on Robotics and Automation*, 2008.

[24] Y. Chen and Y. Rui, "Real-time Speaker Tracking Using Particle Filter Sensor Fusion," *IEEE Trans. Image Processing*, Vol. 92, pp. 485-494, 2004.

[25] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target Tracking Using A Joint Acoustic Video System," *IEEE Trans. on Multimedia*, Vol. 9, Iss. 4, pp. 715-727, 2007.

[26] J. Kwon and K. M. Lee, "Visual Tracking Decomposition," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[27] G. Shafer, "A Mathematical Theory of Evidence," *Princeton University Press*, 1976.

[28] J. Kohlas and P. Monney, "Theory of Evidence – A Survey of Its Mathematical Foundations, Applications and Computational Analysis," *Mathematical Methods of Operations Research*, Vol. 39, Iss.1, pp. 35-68, 1994.

[29] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor Fusion Using Dempster-Shafer Theory," in *Proc. IEEE Instrumentation and Measurement Technology Conference*, 2002.

[30] D. Yu and D. Frincke, "Alert Confidence Fusion in Intrusion Detection Systems with Extended Dempster-Shafer Theory," in *Proc. ACM the 43rd Annual Southeast Regional Conference*, 2005.

[31] M. Yang, Z. Fan, J. Fan, and Y. Wu, "Tracking Non-stationary Visual Appearances by Data-driven Adaptation," *IEEE Trans. on Image Processing,* Vol.18, Iss.7, pp.1633-1644, 2009.

[32] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Transactions on Neural Networks*, Vol. 21, Iss. 10, pp. 1610-1623, 2010.

[33] X. Li, C. Shen, Q. Shi, A. Dick, and A. van den Hengel, "Non-sparse linear representations for visual tracking with online reservoir metric learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 1760-1767, 2012.

[34] N. Jiang, W. Liu, and Y. Wu, "Adaptive and discriminative metric differential tracking," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 1161-1168, 2011.

[35] N. Jiang, W. Liu, and Y. Wu, "Learning Adaptive Metric for Robust Visual Tracking," *IEEE Trans. on Image Processing*, Iss. 99, 2011.

[36] X. Wang, G. Hua, and T. Han, "Discriminative tracking by metric learning," in *Proc. Eur. Conf. Comp. Vis.*, pp. 200–214, 2010.

[37] F. Smarandache and J. Dezert, "Advances and applications of DSmT for information fusion: collected works," *American Research Press*, Vol. 2, 2006.

TABLE I

THE QUANTITATIVE COMPARISON RESULTS OF THE NINE TRACKERS IN AVERAGE TLEWHS ON THE TWENTY VIDEO SEQUENCES.

	FragT	VTD	MILT	OAB1	OAB5	IPCA	L1T	CSVM	**STWDS**
trellis70	1.2141	0.8257	0.9667	1.7182	2.0395	0.7493	0.4465	1.5802	**0.0982**
car11	0.8796	0.6891	0.9613	0.4784	0.2439	0.0636	0.5561	0.8228	**0.0795**
animal	0.5744	1.1982	0.2551	0.5009	0.3388	1.0205	0.2466	0.4680	**0.0134**
cubicle	0.7460	0.5106	1.9058	1.5001	0.4650	1.0906	0.8558	0.9352	**0.2968**
woman	1.4729	1.6843	1.7781	1.5739	1.5724	1.5685	1.7375	1.4374	**0.0413**
seq-simultaneous	0.2168	1.0770	0.5538	0.2454	0.4836	1.1447	0.2804	0.2862	**0.1138**
video-bus2	2.3276	2.0346	5.4240	2.4682	6.1899	0.0990	0.4065	0.4726	**0.0248**
football2	0.1258	1.4699	0.1279	0.1444	3.1110	0.0535	0.5650	3.0895	**0.0357**
seq-jd	0.2218	0.2266	0.6471	0.7800	0.3935	0.8328	0.7256	1.1010	**0.1138**
girl	0.2320	0.1041	0.3443	0.4621	0.5990	0.2010	0.1156	0.2771	**0.1028**
car4	0.3046	0.6611	0.1431	0.2067	0.8936	**0.0523**	0.0658	0.4978	0.0564
shaking	1.5942	3.3748	0.1422	0.1683	0.3764	2.3024	3.7228	1.2529	**0.1368**
pktest02	1.9902	0.1113	**0.0816**	0.1317	1.5471	0.0897	0.1192	0.2741	0.1391
surfer	2.2915	1.7367	0.0559	0.4735	2.5624	1.0473	0.6792	0.8806	**0.0301**
singer2	0.2579	**0.0860**	0.2041	0.6139	1.0540	0.0726	0.2936	0.3236	0.0875
CamSeq01	0.1090	**0.0775**	0.1635	0.1052	0.1297	0.1574	0.4693	0.2253	0.1650
davidin300	0.7871	0.3030	0.2137	0.5370	0.5303	0.0623	0.1741	0.3749	**0.0754**
pedxing-seq2	0.8350	0.0751	0.0795	0.0610	0.1863	0.0468	0.0716	0.0934	**0.0614**
Distortion	0.3852	0.6536	0.0503	0.0600	0.4559	0.3829	0.1982	0.4015	**0.0245**
Pedestrians	**0.0403**	0.2868	0.2941	0.2870	0.3668	0.2945	0.2919	0.3117	0.0423
Mean	0.8303	0.8593	0.7196	0.6258	1.1770	0.5666	0.6011	0.7553	**0.0869**
Std	0.7242	0.8343	1.2021	0.6481	1.4097	0.6153	0.8063	0.6817	**0.0640**

TABLE II

THE QUANTITATIVE COMPARISON RESULTS OF THE NINE TRACKERS IN TRACKING SUCCESS RATES ON THE TWENTY VIDEO SEQUENCES.

	FragT	VTD	MILT	OAB1	OAB5	IPCA	L1T	CSVM	**STWDS**
trellis70	0.2974	0.4072	0.3493	0.2295	0.0339	0.3593	0.4591	0.0519	**0.9581**
car11	0.4020	0.4326	0.1043	0.3181	0.2799	0.9211	0.5700	0.1018	**0.9389**
animal	0.1408	0.0845	0.6761	0.3099	0.5352	0.1690	0.5352	0.4507	**1.0000**
cubicle	0.7255	0.9020	0.2353	0.4706	0.8627	0.7255	0.6863	0.6863	**0.9608**
woman	0.2852	0.2004	0.2058	0.2148	0.1859	0.2148	0.2509	0.3755	**1.0000**
seq-simultaneous	0.6829	0.3171	0.2927	0.6829	0.6585	0.3171	0.5854	0.6098	**0.9268**
video-car	0.4711	0.6353	0.1550	0.4225	0.0578	**1.0000**	0.9058	0.9574	**1.0000**
football2	0.7667	0.4667	0.6233	0.6667	0.0467	0.9967	0.3633	0.2900	**1.0000**
seq-jd	0.8020	0.7723	0.5545	0.5446	0.3168	0.6634	0.2277	0.2178	**0.9406**
girl	0.6335	0.9044	0.2211	0.1773	0.1633	0.8466	0.8845	0.4303	**0.9462**
car4	0.6793	0.6210	0.7959	0.7464	0.3819	**1.0000**	**1.0000**	0.4227	**1.0000**
shaking	0.1534	0.2767	**0.9918**	0.9890	0.8438	0.0110	0.0411	0.2959	**0.9918**
pktest02	0.1667	**1.0000**	**1.0000**	**1.0000**	0.2333	**1.0000**	**1.0000**	0.7667	**1.0000**
surfer	0.2128	0.4149	0.9894	0.3112	0.0399	0.4069	0.2766	0.4309	**0.9920**
singer2	0.9304	**1.0000**	**1.0000**	0.3783	0.2087	**1.0000**	0.6739	0.8217	**1.0000**
CamSeq01	**1.0000**	0.9901	0.9703	**1.0000**	0.9703	0.9901	0.5446	0.8713	**1.0000**
davidin300	0.4545	0.7900	0.9654	0.3550	0.4762	**1.0000**	0.8550	0.5887	**1.0000**
pedxing-seq2	0.2975	**1.0000**	**1.0000**	**1.0000**	0.7911	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Distortion	0.1969	0.1890	**1.0000**	**1.0000**	0.3228	0.2677	0.7638	0.1969	**1.0000**
Pedestrians	**1.0000**	0.5065	0.4870	0.5065	0.4416	0.4935	0.5065	0.4805	**1.0000**
Mean	0.5149	0.5955	0.6309	0.5662	0.3925	0.6691	0.6065	0.5023	**0.9828**
Std	0.2875	0.2990	0.3383	0.2900	0.2899	0.3417	0.2765	0.2717	**0.0255**

[38] J. Straub, "Evaluation and comparison of Dempster-Shafer, weighted Dempster-Shafer, and probability techniques in decision making," in *Proc. International Conference on Machine Vision*, 2011.

[39] X. Li, A. Dick, C. Shen, A. van den Hengel, and H. Wang, "Incremental Learning of 3D-DCT Compact Representations for Robust Visual Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.