Visual Object Tracking by Structure Complexity Coefficients

Yuan Yuan, *Student Member, IEEE*, Huan Yang, Yuming Fang, and Weisi Lin, *Senior Member, IEEE*

***Abstract*—Appearance change of moving targets is a challenging problem in visual tracking. In this paper, we present a novel visual object tracking algorithm based on the observation dependent hidden Markov model (OD-HMM) framework. The observation dependency is computed by structure complexity coefficients (SCC) which is defined to predict the target appearance change. Unlike conventional methods addressing the appearance change problem by investigating different online appearance models, we handle this problem by addressing the fundamental reason of motion-related appearance change during visual tracking. Based on the analysis of motion-related appearance change, we investigate the relationship between the structure of the object surface and the appearance stability. The appearance of complex structural regions is easier to change compared with that of smooth structural regions with object moving. Based on this, we define SCC to predict the appearance stability of moving objects. Different from the standard HMM-based tracking algorithms where observations between different frames are assumed to be independent, we consider the observation dependency between consecutive frames with the information provided by SCC. Moreover, we present a novel outlier removing method in appearance model updating which helps to avoid error accumulation. Experimental results on challenging video sequences demonstrate that the proposed visual tracking algorithm with OD-HMM and SCC achieves better performance than existing related tracking algorithms.**

***Index Terms*—Appearance stability, moving target, object tracking, structure complexity coefficients (SCC).**

I. Introduction

VISUAL object tracking is an important technique in smart vision systems such as visual surveillance, robot navigation and human-computer interaction. Given the target state (e.g., location and size) in the first frame of a video clip, the aim of visual tracking is to estimate the target states in subsequent frames. In most of the recent related studies, object tracking process is modelled as a recursive loop of two steps under the HMM framework: 1) predicting some candidates of the target in the incoming frame based on the estimated target state in the previous frame, and 2) choosing the most proper candidate as the estimated target based on the appearance likelihood. These two steps are referred to motion model and appearance model respectively during the tracking process [1]–[6].

Generally, since the target appearance does not change significantly during tracking, the target could be accurately localized in most of the frames with a relatively high appearance likelihood. However, when the target appearance changes dramatically during tracking, the "target candidate" (the candidate which supposed to be the target) would be assigned with a low likelihood. This makes the "target candidate" not distinguishable with other "non-target candidates" and then the tracker might drift.

With 11 normal attributes listed in the study [7] which may affect the tracking performance, target appearance change during tracking is mainly correlated with motion related attributes (i.e., rotation, deformation and motion blur) and illumination variation. To address the problem caused by illumination variation, existing studies have investigated illumination invariant features such as "adjacent pixels color ratio" [8], sparse representation [4], locality sensitive histogram [5], etc. In this paper, we mainly focus on handling the problem of appearance change caused by target motion during visual object tracking. To deal with this problem, several approaches have been proposed such as finding stable local features [9], [10], building appearance subspace robust to multi-view [1], learning part-based appearance models [11]–[13], etc. Although there are some previous studies trying to address the appearance change problem, the fundamental reason to appearance change, namely, motion is neglected by most of these studies.

When analysing video sequences, we find that the appearance change of moving targets is caused by pixel replacement in a local region. For example, with motion blur caused by fast motion, original target pixels would be replaced by blurred pixels. Moreover, deformation and rotation of the target also result in pixel replacement by its neighbors. Based on our observation and analysis, we find that, for moving targets, the probability of appearance change in complex structural regions is higher than that in smooth structural regions. By assuming the target could be reconstructed by the appearance model, the appearance likelihood can be computed as the summation of of pixel/region-wise reconstruction error [1], [2], while large appearance change would lead to large reconstruction error. Hence, in order to enhance the likelihood of the "target candidate", we compensate the reconstruction errors with the appearance change prediction, by considering the reconstruction error in complex structural regions less than that in smooth regions. This strategy is also

Manuscript received August 25, 2014; revised January 23, 2015 and April 24, 2015; accepted May 27, 2015. Date of publication June 03, 2015; date of current version July 15, 2015. This work was supported in part by the NSF of Jiangxi under Grant 20142BAB217011 and Grant 20151BDH80003, and by the SRF for ROCS, SEM, China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris. *(Corresponding author: Yuming Fang.)*

Y. Yuan, H. Yang, and W. Lin are with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yyuan004@ntu.edu.sg; hyang3@ntu.edu.sg; wslin@ntu.edu.sg).

Y. Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2440996

similar to contrast masking of the human visual system (HVS), where the HVS is more sensitive to the contrast in smooth regions than that in complex structural regions [14], [15].

In this study, starting from analysing the motion related appearance change, we define Structure Complexity Coefficients (SCC) to predict the target appearance stability (negatively correlated to appearance change). Since target appearances in several consecutive frames are expected to be quite similar, we predict the SCC of the target with the estimated target appearance in last several frames. In the proposed tracking framework, the target appearance likelihood depends upon not only the reconstruction error computed from the appearance model, but also the appearance stability estimated based on the target appearance in the previous frames. By incorporating this dependency with the standard HMM based tracking framework, we design an Observation Dependent HMM (OD-HMM) tracking framework with SCC.

In essence, we design an SCC based tracking algorithm (*SCCT*) under OD-HMM framework where the observation dependency is computed based on SCC. An appearance model with SCC-Gaussian-Laplacian distance is built to measure the likelihood of candidates predicted by motion model. The Laplacian noise has been demonstrated to be effective in handling outliers for object tracking [2], [16], [11]. In this study, we also present a novel model updating method to remove the outliers, which can address the problem from the error accumulation in appearance model updating. Experimental results demonstrate the effectiveness of both the proposed SCC and the outlier removing methods.

II. Related Works

Visual object tracking is an important research topic in computer vision and has been studied for years. In this section, we briefly summarize the studies which are closely related to our research work. A more thorough survey can be referred to [17].

HMM-based tracking: By defining the target location as a hidden state and assuming that the observations are only related to the hidden state, HMM simplifies the tracking task to a sequential Maximum a Posterior (MAP) estimation problem. Particle Filter has been widely used in solving the sequential MAP estimation problem [1], [2], [18]. The Particle Filter provides a framework to estimate the posterior probability sequentially by several random particles and corresponding weights, regardless of the distribution of prediction and observation functions. The independent assumption of observations allows low computational complexity in appearance likelihood computation. However, these studies neglect the temporal relationship between target appearances in consecutive frames. In this study, the observation dependency is used as an important cue in computing the appearance likelihood besides the appearance model. We build a more robust Particle Filter based tracker by introducing the observation dependency into conventional HMM tracking framework.

Observation dependency in HMM: In the study [19], by expressing the multiple observation probability as a combination of individual observation probabilities, the observation dependence property is characterized by combinatorial weights. To characterize the observation dependency in visual object tracking, AR(Autoregressive)-HMM [20] was proposed where the dependency between consecutive frames is trained frame by frame based on all the previous estimated target appearances. This approach may be effective in tracking targets with repetitive movements (e.g., like shaking, running back and forth, and rotation), since the target appearance in these cases is predictable. However, in more general cases where the targets are with unpredictable movements, it is very hard to predict how the target appearance will change in the incoming frame. Moreover, the dependency training algorithm increases the computational complexity greatly. It is reported in [20] that it requires 8 second per frame in processing (implemented by C codes with Intel Q9550 2.83 GHz CPU). In this paper, the appearance stability is predicted by analysing the target structure complexity. We predict where the appearance will change rather than how the appearance will change, which is independent to the target motion. Thus, the proposed SCC is more efficient and reasonable in appearance stability estimation if we don't have the pre-knowledge of the target movement in the coming frames.

Outlier modelling in appearance model: Outliers refer to unexpected appearance data which are different greatly from the existing observed target appearance. The outlier can be introduced by occlusion, sudden illumination variation, etc. And it affects the robustness of the appearance model negatively. In [21], the appearance model updating problem is analysed and it claims that model updating with no error control would cause drifting if error accumulates during tracking. In [22], an $\mathcal{WSL}$ appearance model is proposed. The $\mathcal{L}$ refers to the 'lost' component which is used to model the outlier probability; the $\mathcal{S}$ models the temporal appearance stability; and the $\mathcal{W}$ stands for wandering model which can be used when the stable model is not available. When an observation is detected with higher probability to be outlier, the observation contributes less to the stable appearance model of the target. Semi-supervised learning [23] is applied in tracking to increase the stability of the appearance model during model updating. P-N learning is proposed in [24] and it is proved effective in controlling the error accumulation during tracking by removing the mislabeled training samples. The Gaussian-Laplacian noise is facilitating in outlier detection [2]; pixels with non-zero Laplacian noise are detected as outliers and interpolated by the mean value in model updating. In the proposed *SCCT*, a novel outlier removal approach is designed. With the assumption of SCC-weighted Gaussian-Laplacian noise in the appearance model, we decompose the target appearance into three parts: subspace (learnt from previous target appearance), Gaussian noise (valid appearance change), and Laplacian noise (invalid appearance change). Then, we only eliminate the invalid appearance change represented by Laplacian noise during appearance model updating and keep the valid appearance change represented by Gaussian noise at the same time.

III. Appearance Stability Prediction by Structure Complexity Coefficients

With rotation, deformation or motion blur, the target appearance would change from time to time during tracking. As target appearance changes, the reconstruction likelihood of target degrades and, hence, the tracked target would be lost. In order to enhance the reconstruction likelihood of moving targets, we

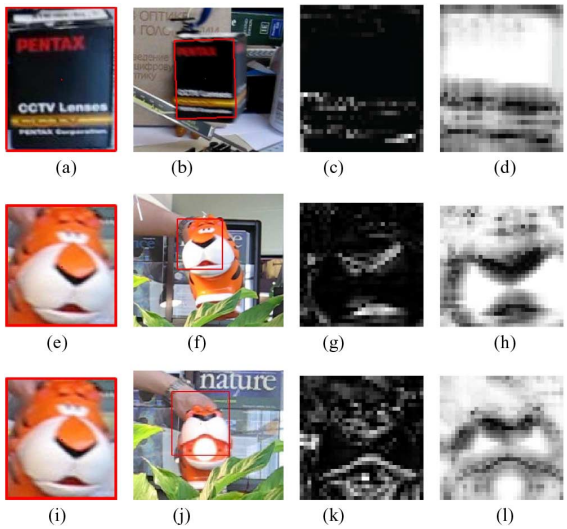


Fig. 1. Reconstruction error map and SCC with target movement: (a), (e), and (i) are targets to be tracked; (b), (f), and (j) are targets with motion blur and rotation in tth frame; (c), (g), and (k) are pixel-wise reconstruction error maps in tth frame, the dark region indicates less error and the light region indicates large error; (d), (h), and (l) are computed SCCs based on the estimated target appearance in $(t-1)$th frame, the dark region indicates complex structural region which takes high probability to change its value with motion and the light region indicates smooth structural regions which will be more stable with motion. The reconstruction error map and SCC are somehow negatively correlated, which means the SCC can be used to predict and compensate the reconstruction error caused by target motion. (a) Box. (b) Motion blur (c) Recon. error. (d) SCC. (e) Tiger1. (f) Rotation. (g) Recon. error. (h) SCC. (i) Tiger1. (j) Deformation. (k) Recon. error. (l) SCC.

propose SCC to predict the appearance stability of different regions of the target. The proposed SCC is applied to compensate the reconstruction error of the target. Thus, the likelihood of the candidate is enhanced after compensation if it is supposed to be the target. On the contrary, the likelihood of the candidate would degrade if it doesn't belong to the target.

Generally, the appearance change caused by motion can be interpreted by pixel replacement in a local region. Therefore, the reconstruction error caused by motion-related appearance change in complex structural regions is larger than that in smooth structural regions [as demonstrated in Fig. 1(c), 1(g) and 1(k)]. We here analyse appearance change from three types of target motion (i.e., rotation, deformation and motion blur) and derive SCC to predict the motion driven appearance change.

A. Appearance Change Caused by Rotation

Rotation, a kind of target-camera relative motion, occurs frequently in object tracking. It is a global motion, which means that almost all the pixels of the target shift together to the same direction with the same degree. Here, we predict the pixel value variation between consecutive frames and then model the appearance stability with target rotation.

Since rotation is a global movement of the target, by defining a certain degree and direction (i.e., motion vector defined as (m_x, m_y)) of the rotation, the target pixels will shift together and the pixel value variation at (x, y) can be computed as

$$d(m_x, m_y, x, y) = \mathbf{I}(x + m_x, y + m_y) - \mathbf{I}(x, y) \tag{1}$$

where $\mathbf{I}$ is the image patch of the target, and (m_x, m_y) is the motion vector. By assuming a uniform distribution of the movement within a maximum velocity, the variance (i.e., expected square of the pixel value variation) of the pixel value between consecutive frames can be computed as

$$\mathrm{Var}_r(\mathbf{I}) = \sum_{m_x^2 + m_y^2 \le v_m^2} \frac{d(m_x, m_y, x, y)^2}{N_d} \tag{2}$$

where v_m denotes the maximum moved distance between consecutive frames and N_d is the number of pixels within the circle region with radius of v_m. With predicted variance $\mathrm{Var}_r(\mathbf{I})$, the appearance stability $P_r(\mathbf{I})$ with target rotation can be modelled as a normal distribution

$$P_r(\mathbf{I}) = exp\left(-\frac{\mathrm{Var}_r(\mathbf{I})}{\sigma_r^2}\right) \tag{3}$$

where σ_r^2 is the parameter of Gaussian kernel.

B. Appearance Change Caused by Deformation

Another attribute of motion-related appearance change is deformation. In this case, the local motion varies between different regions of target surface. For instance, the facial expression variation is a kind of deformation and the movements of different regions of the face are independent. Since the local movement is random, a pixel is more likely to be replaced by its closer neighbors than further pixels. Thus, we model the local movement between consecutive frames as a normal distribution. With the predicted pixel value variation, we then compute the appearance stability of the target with deformation.

By assuming a two-dimensional normal distribution of the movement within a maximum velocity, the variance of the pixel value between consecutive frames can be defined as

$$\mathrm{Var}_d(\mathbf{I}) = \sum_{m_x^2 + m_y^2 \le v_m^2} d(m_x, m_y, x, y)^2 P(m_x, m_y) \tag{4}$$

where v_m denotes the maximum moved distance of pixels between consecutive frames and $P(m_x, m_y) = \mathcal{N}(0, 0, v_m^2/9, v_m^2/9, 0)$. With the predicted variance $\mathrm{Var}_d(\mathbf{I})$, the appearance stability $P_d(\mathbf{I})$ with target deformation can be modelled as a normal distribution

$$P_d(\mathbf{I}) = exp\left(-\frac{\mathrm{Var}_d(\mathbf{I})}{\sigma_d^2}\right) \tag{5}$$

where σ_d^2 is the parameter of Gaussian kernel.

C. Appearance Change Caused by Motion Blur

Motion blur caused by fast motion is also an important attribute of motion-related appearance change. With fast motion of target or camera, the blurred pixel is the fusion result of radiant energy from a relatively large region [25]. However, the blurred image cannot precisely reflect the original appearance of the target. Consider a simple movement in $x+$ direction, we can predict the blurred pixel value as [25]

$$\begin{aligned} \mathbf{I}'(x, y) &\propto \int_0^{\Delta T} \mathbf{I}(x - v(t), y) dt \\ &\propto \sum_{x'=0}^{\Delta X} \mathbf{I}(x - x', y) \end{aligned} \tag{6}$$

where $\mathbf{I}(x, y)$ is the clear image the target, $\mathbf{I}'(x, y)$ is the predicted blurred one, $v(t)$ indicates the relative velocity, ΔT is the shutter time, and ΔX is the relative moving distance during this time interval.

From (6) we can predict that, the appearance variation from motion blur in complex structural regions will be larger than that in smooth structural regions. By assuming a uniform distribution of the motion direction, for simplicity, we compute $\mathrm{Var}_b(\mathbf{I})$ based on the average of motion blurred difference in four directions ($x+$, $x-$, $y+$ and $y-$)

$$\begin{aligned}
\mathrm{Var}_b(\mathbf{I}) &= \frac{\mathrm{Var}_b^{x+}(\mathbf{I}) + \mathrm{Var}_b^{x-}(\mathbf{I}) + \mathrm{Var}_b^{y+}(\mathbf{I}) + \mathrm{Var}_b^{y-}(\mathbf{I})}{4} \\
\mathrm{Var}_b^{x+}(\mathbf{I}) &= \left[\frac{\sum_{n=0}^{v_m} \mathbf{I}(x-n, y)}{v_m + 1} - \mathbf{I}(x, y) \right]^2 \\
\mathrm{Var}_b^{x-}(\mathbf{I}) &= \left[\frac{\sum_{n=0}^{v_m} \mathbf{I}(x+n, y)}{v_m + 1} - \mathbf{I}(x, y) \right]^2 \\
\mathrm{Var}_b^{y+}(\mathbf{I}) &= \left[\frac{\sum_{n=0}^{v_m} \mathbf{I}(x, y-n)}{v_m + 1} - \mathbf{I}(x, y) \right]^2 \\
\mathrm{Var}_b^{y-}(\mathbf{I}) &= \left[\frac{\sum_{n=0}^{v_m} \mathbf{I}(x, y+n)}{v_m + 1} - \mathbf{I}(x, y) \right]^2
\end{aligned} \tag{7}$$

where v_m denotes the assumed moving distance between consecutive frames. With the predicted variance $\mathrm{Var}_b(\mathbf{I})$, the appearance stability $P_b(\mathbf{I})$ with motion blur can be modelled as a normal distribution

$$P_b(\mathbf{I}) = exp\left(-\frac{\mathrm{Var}_b(\mathbf{I})}{\sigma_b^2} \right) \tag{8}$$

where σ_b^2 is the parameter of Gaussian kernel.

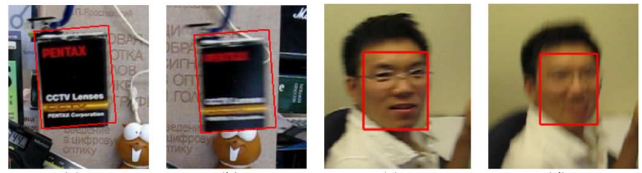
D. Structure Complexity Coefficients

As analyzed in above sections, rotation results in pixel shift by global motion; deformation results in pixel shift by local motion; and blur results in pixel integration by fast global motion. Assuming these movements are independent, we formulate the appearance stability of the moving target based on the joint probability

$$\begin{aligned}
P(\mathbf{I}) &= P_r(\mathbf{I}) * P_d(\mathbf{I}) * P_b(\mathbf{I}) \\
&= exp\left(-\frac{\mathrm{Var}_r(\mathbf{I})}{\sigma_r^2} - \frac{\mathrm{Var}_d(\mathbf{I})}{\sigma_d^2} - \frac{\mathrm{Var}_b(\mathbf{I})}{\sigma_b^2} \right)
\end{aligned} \tag{9}$$

where P is the pixel-wise appearance variation probability of moving target. With the larger value in $P(\mathbf{I})$, the related pixel or region is more stable with target movement.

1) Motion Blur Detection: Visually, motion blur only occurs with fast motion, and it can be detected by image analysis algorithms [26], [27]. In this study, the parameter σ_b is calculated based on the blur detection results. With lager motion blur, σ_b is set to be smaller to make P_b contributing more to P, and vice versa.



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

Fig. 2. Illustration of α_k and q_k of a box and a face within the red bounding box in different blur degrees. (a) $\alpha_k = 1.9$, $q_k = 0.26$. (b) $\alpha_k = 2.9$, $q_k = 0.90$. (c) $\alpha_k = 3.1$, $q_k = 0.24$. (d) $\alpha_k = 4.1$, $q_k = 0.64$.

Power Spectrum Slope has been proved promising in image blur detection, due to the fact that a blurred image usually has a large slope of power spectrum, while an unblurred image always corresponds to a small slope of power spectrum [28], [29]. Before computing power spectrum slope, we first compute the power spectrum of the resized target patch (size of $N \times N$) by taking the squared magnitude after Discrete Fourier transform (DFT)

$$S(u, v) = \frac{1}{N^2} |\mathbf{I}_f(u, v)|^2 \tag{10}$$

where $\mathbf{I}_f(u, v)$ denotes the DFT coefficients of the target patch. By representing $S(u, v)$ in polar coordinates, we can obtain $S(f, \theta)$. Then, $S(f)$ is approximated by summing the power spectrum S over all directions θ

$$S(f) = \sum_{\theta} S(f, \theta) \simeq A / f^{-\alpha} \tag{11}$$

where A is an amplitude scaling factor for each orientation and α is the frequency exponent, called Power Spectrum Slope [26].

Since the texture of different objects varies from each other, even at the same degree of motion blur, α may differs among different objects. In order to get a unique metric for blur degree among different objects, by assuming the objects are not blurred in the first frame, we derive q_k by comparing α_k against α_1 (the Power Spectrum Slope of the labeled target in frame k and the first frame, respectively)

$$q_k = \frac{\alpha_k - \alpha_1}{\alpha_1}. \tag{12}$$

Fig. 2 gives illustration of α_k and q_k of a box and a face with different blur degrees. Since q_k can well represent the blur degree of targets despite object diversity, we compute σ_b inversely proportional to q_k

$$\sigma_b \propto \frac{1}{\max(q_k, \tau)} \tag{13}$$

where τ is a small value to prevent dividing by zero or negative value (we set τ to be 0.0001).

In this study, we set the values of σ_r and σ_d empirically in experiments.

2) Normalization of SCC: The proposed SCC denotes the appearance stability of each pixel. When we apply it to compensate the reconstruction error of the target, a normalization step is necessary in order to make it comparable with conventional

reconstruction error computation methods. The pixel-wise probability $P(\mathbf{I})$ of appearance stability is normalized as

$$\mathbf{C}(\mathbf{I}) = \frac{N_P \times P(\mathbf{I})}{\sum P(\mathbf{I})} \tag{14}$$

where $\mathbf{C}(\mathbf{I})$ is SCC of the image patch $\mathbf{I}$, and N_P is the number of elements in $P(\mathbf{I})$. Fig. 1(d), 1(h) and 1(l) illustrate the computed SCC, where dark region indicates small values in $\mathbf{C}(\mathbf{I})$.

IV. Structure Complexity Coefficients Tracker

In this section, we incorporate the proposed SCC with a Gaussian-Laplacian (GL) noise based appearance model under OD-HMM framework. Experimental results in Section V demonstrate that the appearance model is more robust to object moving by compensating the GL noise with the proposed SCC.

A. Gaussian-Laplacian Noise Compensated by SCC

Laplacian noise assumption has been proved to be effective in handling outliers in object tracking, especially in occlusion scenarios [2], [16], [11]. With the same concern, we model the target appearance vector $\mathbf{y}$ (the target image patch) as a linear model with Gaussian-Laplacian noise in this study

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \mathbf{n} + \mathbf{s} \tag{15}$$

where $\mathbf{y} \in \mathbb{R}^{d\times 1}$ is a d-dimensional vector, $\mathbf{a} \in \mathbb{R}^{m\times 1}$ denotes the estimated m-dimensional parameter vector and $\mathbf{B} \in \mathbb{R}^{d\times m}$ represents the input data matrix (the row vector is subspace basis in this study). Specifically, $\mathbf{n} \in \mathbb{R}^{d\times 1}$ indicates the Gaussian noise term and $\mathbf{s} \in \mathbb{R}^{d\times 1}$ indicates the Laplacian noise term. The Gaussian noise is used to model small dense noise, while the Laplacian noise aim to deal with outliers.

Based on the Gaussian-Laplacian reconstruction error assumption, given an appearance vector $\mathbf{y}$ and the subspace basis $\mathbf{B}$, the parameter vector $\mathbf{a}$ and Laplacian noise $\mathbf{s}$ is computed by maximizing the joint likelihood $p(\mathbf{y}, \mathbf{a}, \mathbf{s})$ [2], [11]

$$\begin{aligned} p(\mathbf{y}, \mathbf{a}, \mathbf{s}) &= K \exp\left\{-\frac{1}{\sigma_N^2}[\frac{1}{2}\|\mathbf{n}\|_2^2 + \lambda\|\mathbf{s}\|_1]\right\} \\ &= K \exp\left\{-\frac{1}{\sigma_N^2}[\frac{1}{2}\|\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s}\|_2^2 + \lambda\|\mathbf{s}\|_1]\right\} \end{aligned} \tag{16}$$

where σ_N^2 is the variance of the Gaussian error term, λ is regularization constant of Laplacian error terms and K is the normalization constant.

As shown in (16), each element in Gaussian or Laplacian noises is treated equally in both ℓ_1 and ℓ_2 normalization. However, as analysed in Section III, regions with different structure complexities are with different appearance stabilities. If one pixel is predicted to be unstable in motion, the noise correlated to this pixel should be less considered during the calculation of the appearance likelihood. Thus, the joint reconstruction likelihood can be rewritten as follows by considering SCC:

$$p(\mathbf{y}, \mathbf{a}, \mathbf{s}, \mathbf{C}) = K \exp\left\{-\frac{1}{\sigma_N^2}[\frac{1}{2}\mathbf{n}^\top \mathbf{W}_\mathbf{n}\mathbf{n} + \lambda\|\mathbf{s}\mathbf{W}_\mathbf{s}\|_1]\right\} \tag{17}$$

where $\mathbf{W}_\mathbf{n}$ and $\mathbf{W}_\mathbf{s}$ are two diagonal matrixes, which denote the error tolerance of each element in $\mathbf{y}$. Larger values in these two matrixes indicate less error tolerance of the related element in $\mathbf{y}$. We define $\mathbf{W}_\mathbf{n}$ and $\mathbf{W}_\mathbf{s}$

$$\mathbf{W}_\mathbf{n} = \mathtt{diag}(\mathbf{C}), \tag{18}$$

$$\mathbf{W}_\mathbf{s} = \sqrt{\mathtt{diag}(\mathbf{C})} \tag{19}$$

where $\mathbf{C}$ is the computed SCC in Section III-D and $\mathtt{diag}(\mathbf{C})$ converts $\mathbf{C}$ to a diagonal matrix.

In order to maximize the joint likelihood in (17), the objective function below is minimized:

$$\mathcal{L}(\mathbf{a}, \mathbf{s}) = \frac{1}{2}(\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s})^\top \mathbf{W}_\mathbf{n}(\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s}) + \lambda\|\mathbf{s}\mathbf{W}_\mathbf{s}\|_1, i \tag{20}$$

to get the optimal solution: $[\hat{\mathbf{a}}, \hat{\mathbf{s}}] = \arg\min_{\mathbf{a},\mathbf{s}} \mathcal{L}(\mathbf{a}, \mathbf{s})$. We solve the above object function based on weighted least square criterion and ℓ_1 regularization term on $\mathbf{s}$. Alg. 1 provides a detailed description of the optimization process.

Algorithm 1: ℓ_1 regularization with SCC

Input : An observation vector $\mathbf{y}$, matrix $\mathbf{B}$, diagonal matrixes $\mathbf{W}_\mathbf{n} = \mathtt{diag}(\mathbf{C})$ and $\mathbf{W}_\mathbf{s} = \sqrt{\mathtt{diag}(\mathbf{C})}$, a small constant λ and max iteration times i_{max}

Output: Estimated optimal $\hat{\mathbf{a}}$ and $\hat{\mathbf{s}}$

Initialize $\mathbf{s}_0 = \mathbf{0}$ and $i = 0$;
$\mathbf{P} = (\mathbf{B}^\top \mathbf{W}_\mathbf{n}\mathbf{B})^{-1}\mathbf{B}^\top \mathbf{W}_\mathbf{n}$;
while *Not convergence or* $i < i_{max}$ **do**
    Compute $\mathbf{a}_{i+1}$ via $\mathbf{a}_{i+1} = \mathbf{P}(\mathbf{y} - \mathbf{s}_i)$;
    Compute $\mathbf{s}_{i+1}$ via $\mathbf{s}_{i+1} = \mathcal{S}_{\lambda/\mathbf{W}_\mathbf{s}}([\mathbf{y} - \mathbf{B}\mathbf{a}_{i+1}])$;
    $i = i + 1$
end
$\hat{\mathbf{a}} = \mathbf{a}_i$, $\hat{\mathbf{s}} = \mathbf{s}_i$;

B. SCC-Gaussian-Laplacian Versus Gaussian-Laplacian

Given a subspace basis $\mathbf{B}$ and several predicted noisy target candidates, we have to estimate the most reliable one based on the distances between candidates' appearance and the subspace. The distance is usually defined to be inversely proportional to the maximized joint likelihood with respect to the estimated coefficient

$$d(\mathbf{y}; \mathbf{B}) \propto -\log \max_\mathbf{a} p(\mathbf{y}, \mathbf{a}, \mathbf{s}) \tag{21}$$

where $\mathbf{y}$ denotes the appearance vector of the candidate and $\mathbf{a}$ is the estimated parameter.

With the SCC-GL noise assumption, we define the SCC-GL distance from observation vector to the appearance model as follows:

$$\begin{aligned} d_{SCC} &= -\log \max_{\mathbf{a},\mathbf{s}} p(\mathbf{y}, \mathbf{a}, \mathbf{s}) \\ &= \min_{\mathbf{a},\mathbf{s}} \frac{1}{2}(\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s})^\top \mathbf{W}_\mathbf{n}(\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s}) + \lambda\|\mathbf{s}\mathbf{W}_\mathbf{s}\|_1. \end{aligned} \tag{22}$$

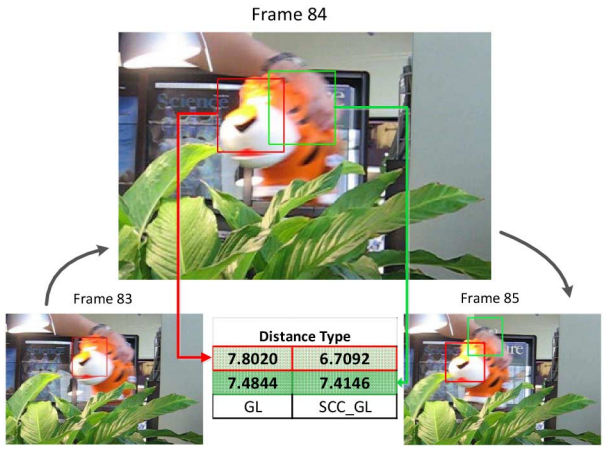


Fig. 3. Moving tiger toy tracking illustration with GL and SCC-GL distances: Before Frame 83, both GL and SCC-GL return good tracking results and the same appearance model. In Frame 84, if GL distance is applied, the green bounding box (bad candidate) returns a smaller distance value; if we use SCC-GL distance, the red bounding box (good candidate) returns a smaller distance value.

Moreover, in Least Soft-threshold Squares Tracking [2] and Online NMF tracking [11], the Gaussian Laplacian (GL) distance is defined

$$d_{GL} = \min_{\mathbf{a},\mathbf{s}} \frac{1}{2}\|\mathbf{y} - \mathbf{B}\mathbf{a} - \mathbf{s}\|_2^2 + \lambda\|\mathbf{s}\|_1. \tag{23}$$

Fig. 3 illustrates a tiger toy tracking example. In Frame 84 where a moving tiger toy exists, the SCC-GL distance of the good candidate (red bounding box) is smaller than the bad candidate (green bounding box), while the GL distance of the good candidate is larger than the bad one.

C. Observation Dependent HMM

HMM is widely used in object tracking algorithms. Let $\mathbf{x}_t$ be the state variable of the target and $\mathbf{z}_t$ be the image observation (in principle, the entire image frame) at time t.

Given $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_t\}$ from the first frame to the t-th frame, the aim is to estimate $\mathbf{x}_t$ by MAP estimation

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p(\mathbf{x}_t|\mathbf{z}_{1:t}). \tag{24}$$

Taking the Bayes rule into consideration, the posterior probability can be decomposed as

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{z}_{1:t}) &= \frac{p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1})p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \\ &\propto p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1})p(\mathbf{x}_t|\mathbf{z}_{1:t-1}). \end{aligned} \tag{25}$$

In (25), since the observation $\mathbf{z}_t$ and $\mathbf{z}_{1:t-1}$ are given at time t and $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ is independent to the state variable $\mathbf{x}_t$, we treat $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ as an normalization constant to all the possible $\mathbf{x}_t$.

In traditional HMM based tracking studies [1], [2], [18], the observations between different frames are assumed to be independent, and the appearance model $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1})$ is simplified to be $p(\mathbf{z}_t|\mathbf{x}_t)$ in these studies. However, since the appearance variation between consecutive frames is very limited due to the short time interval, the appearance is somehow correlated within certain number of frames. So, in order to derive the appearance observation dependency between consecutive frames, we model the tracking process as an Observation Dependent HMM (OD-HMM). Specifically, based on the target appearance in the previous frames, SCC is calculated to predict the target appearance change at the t-th frame. Thus, within the OD-HMM tracking framework, the posterior probability can be computed as

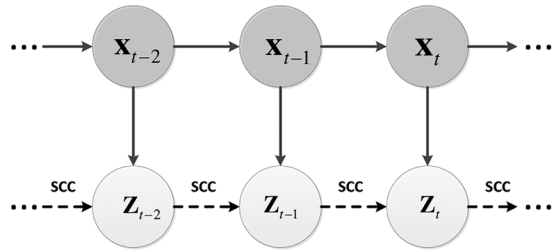


Fig. 4. First-order observation-dependent HMM based on structure complexity coefficients.

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1}) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) \tag{26}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ denotes the motion model that predicts prior probability of candidate state, and $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{1:t-1})$ is the appearance model which is used to estimate the probability of the observation given the target state and observations in previous frames. In our current implementation, we only use the 1st-order OD-HMM for simplicity. The 1st-order OD-HMM is illustrated in Fig. 4. When modeling the probability of $\mathbf{z}_t$, we consider not only the relationship between the target state and the observation at the current frame, but also the relationship between the observations at consecutive frames.

Motion Model: In this work, we consider the affine transformation $\mathbf{x}_t = (x_t, y_t, \theta_t, s_t, \alpha_t, \phi_t)$ to be the state variable, where $x_t, y_t, \theta_t, s_t, \alpha_t, \phi_t$, denote x, y translation, rotation angle, scale, aspect ratio, and skew direction at time t. The transition between consecutive frames is modelled as Gaussian distribution with a diagonal covariance matrix $\mathbf{\Sigma_x}$

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}, \mathbf{\Sigma_x}). \tag{27}$$

Appearance Model: Given the predicted target state $\mathbf{x}_t$, we define $\mathbf{y}_t$ as the corresponding target appearance vector in the image observation $\mathbf{z}_t$. Thus, $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{t-1})$ is equivalent to the likelihood of $\mathbf{y}_t$ belonging to the target given the SCC predicted from $\mathbf{z}_{t-1}$.

$$p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{z}_{t-1}) = p(\mathbf{y}_t|\mathbf{z}_{t-1}) \tag{28}$$

By assuming the target appearance is generated from a PCA subspace, with Gaussian-Laplacian noise assumption, the target appearance $\mathbf{y}$ is formulated

$$\mathbf{y} = \mu + \mathbf{U}\mathbf{c} + \mathbf{n} + \mathbf{s} \tag{29}$$

where $\mathbf{n}$ and $\mathbf{s}$ are Gaussian and Laplacian noise term respectively. The PCA subspace which represents the target appearance is spanned by $\mathbf{U}$ and centered at μ. We can define $\bar{\mathbf{y}} = \mathbf{y} - \mu$. Given a target appearance candidate $\mathbf{y}_t^i$ corresponding

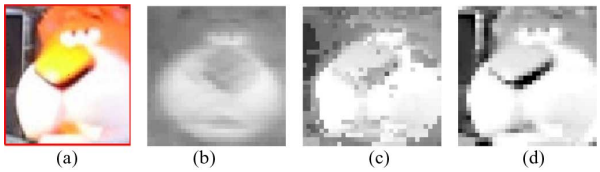


Fig. 5. Outlier elimination methods comparison: (a) the estimated target in the current frame; (b) the mean of appearance vectors in the previous frame which is a part of the observation model; (c) the result of replacing L-noise pixels with existing appearance vector mean; and (d) the result of the proposed method by subtracting L-noise. (a) Estimated observation. (b) Observations mean. (c) Replace L-noise pixels. (d) Subtract L-noise.

to a predicted state variable $\mathbf{x}_t^i$, based on the discussions in Sections IV-A and IV-B, we first solve the optimization problem

$$[\hat{\mathbf{c}}^i, \hat{\mathbf{s}}^i] = \arg\min_{\hat{\mathbf{c}}^i, \hat{\mathbf{s}}^i} \frac{1}{2}(\bar{\mathbf{y}}^i - \mathbf{U}\mathbf{c}^i - \mathbf{s}^i)^\top \mathbf{W_n}(\bar{\mathbf{y}}^i - \mathbf{U}\mathbf{c}^i - \mathbf{s}^i) + \lambda\|\mathbf{s}^i\mathbf{W_s}\|_1 \quad (30)$$

where i denotes the i-th candidate predicted by motion model, $\mathbf{W_n}$ and $\mathbf{W_s}$ are obtained by (18) based on $\mathbf{C}(\hat{\mathbf{y}}_{t-1})$, where $\hat{\mathbf{y}}_{t-1}$ is the estimated target appearance in the previous frame.

After the optimized $\hat{\mathbf{c}}^i$ and $\hat{\mathbf{s}}^i$ are obtained, the SCC-GL distance can be calculated

$$d(\mathbf{y}_t^i; \mathbf{U}, \mu) = \frac{1}{2}(\bar{\mathbf{y}}_t^i - \mathbf{U}\hat{\mathbf{c}}_t^i - \hat{\mathbf{s}}_t^i)^\top \mathbf{W_n}(\bar{\mathbf{y}}_t^i - \mathbf{U}\hat{\mathbf{c}}_t^i - \hat{\mathbf{s}}_t^i) + \lambda\|\hat{\mathbf{s}}_t^i\mathbf{W_s}\|_1 \quad (31)$$

and the observation likelihood of the y_t^i belonging to the target is measured by

$$p(\mathbf{y}_t^i|\mathbf{z}_{t-1}) = exp\left(-\frac{d(\mathbf{y}_t^i; \mathbf{U}, \mu)}{\sigma^2}\right) \quad (32)$$

where σ is a constant controlling the variance of the Gaussian Kernel.

Model Updating: As analysed in [2], [16], [11], the Laplacian noise term represents the outliers. Thus, in order to eliminate the negative effect of outlier, we reconstruct the estimated target appearance vector $\hat{\mathbf{y}}_t$ by subtracting its related Laplacian noise $\hat{\mathbf{s}}_t$ before model updating

$$\mathbf{y}_{recon}^t = \hat{\mathbf{y}}_t - \hat{\mathbf{s}}_t \quad (33)$$

where $\mathbf{y}_{recon}^t$ is the reconstructed appearance vector in t-th frame and is used for updating the subspace center μ and basis matrix $\mathbf{U}$. The updating of μ and $\mathbf{U}$ is based on incremental principal component analysis (IPCA) [1].

Compared with outlier elimination method by replacing the outlier pixels with mean values of the former appearance vectors in LSST [2], our method makes the reconstructed appearance vector more smooth and retains some variable information from the Gaussian term as well. As shown in Fig. 5(c), all the pixels with non-zero value in Laplacian term are treated as outlier pixels and are replaced by existing mean values, where the Gaussian noise information in all outlier pixels are missing. The proposed method by subtracting the appearance vector with Laplacian noise keeps the varying appearance information which is represented by the Gaussian noise term and removes the effect the Laplacian noise as well.

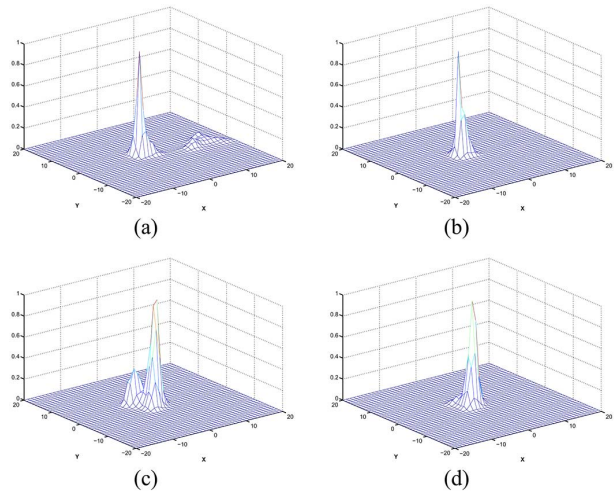


Fig. 6. SCC-GL and GL likelihood maps around the target in Tiger1 sequence. (a) and (b) are taken from Frame 82, while (c) and (d) are taken from Frame 92. (a) GL likelihood. (b) SCC-GL likelihood. (c) GL likelihood. (d) SCC-GL likelihood.

V. Experiments

To evaluate the performance of OD-HMM and SCC in visual object tracking, we implement a Structure Complexity Coefficients Tracker (*SCCT*) in MATLAB R2012b. The regularization constant λ is fixed to be 0.1 in the experiments. Simultaneously, the variance σ_r^2 and σ_d^2 are fixed to be 0.1, while σ_b^2 varies based on the result from motion blur detection. Each image observation is resized to be a 32×32 image patch and then reshaped to a 1×1024 dimension observation vector. The maximum moving distance between consecutive frames v_m is empirically fixed to be 4 in all the experiments. 16 eigenvectors are used to represent the PCA subspace of the target appearance. In order to obtain both efficiency and accuracy, the particle number is set as 600 and the PCA subspace is updated every 5 frames.

In this Section, we have conducted comparison experiments between the proposed *SCCT* and eleven recent state-of-art algorithms, including *LSST* [2], *IVT* [1], *MTT* [16], *SCM* [30], *MIL* [10], *LOT* [31], *TLD* [24], *BLUT* [32], *DT* [33], *Struck* [34], *LSHT* [5] and *ART* [20]. Moreover, to verify the contribution of SCC alone, we implement *SCCT-CI* by fixing the SCC matrix $\mathbf{C}$ to an identity matrix $\mathbf{I}$ where the observation dependency is not considered. All these algorithms are tested on thirteen published benchmark sequences with motion-related challenges like rotation, deformation and motion blur. And these tested sequences cover most tracking scenarios including both indoor and outdoor, and various target types such as face, human, vehicles, small objects, etc. Specifically, the sequences and ground truth are collected from Visual Tracker Benchmark [7], *BLUT* dataset [32], *PROST* dataset [35], *LSST* dataset [2] and *BoBoT* dataset [36].

A. Distracter Elimination With SCC

Generally, the reason of losing tracking of targets is that the estimated likelihood of distracter candidate is larger than that of the target in cases of fast motion, appearance variation, etc. In Fig. 6, we plot the likelihood map around the target candidate in different frames, where X and Y direction means number of pixels shifting at the target position. It can be seen that with

TABLE I
AVERAGE PER-FRAME CENTER LOCATION ERROR (IN PIXEL), THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE, AND GREEN FONTS

Sequence	SCCT	SCCT-CI	LSST	IVT	MTT	SCM	MIL	LOT	TLD	BLUT	DT	Struck	LSHT
Tiger1	*20.48*	43.99	68.27	69.70	56.76	155.32	31.17	160.30	38.56	138.23	**9.62**	76.38	148.81
Box	**8.89**	*12.62*	16.31	11.99	94.36	116.73	182.06	255.60	123.57	202.12	106.18	**8.62**	109.68
Lemming	**11.79**	**12.82**	*15.20*	41.54	169.53	81.87	77.34	29.01	156.58	171.07	166.32	36.44	84.28
Car11	**1.53**	1.70	**1.60**	2.00	2.26	**1.60**	40.90	50.52	26.58	76.26	18.67	1.74	9.46
Body	14.30	**8.61**	**12.96**	113.27	26.17	13.70	176.03	89.37	42.22	14.63	259.16	113.82	*13.04*
Car4	**4.50**	19.60	*7.97*	8.98	19.15	**6.92**	190.46	28.27	18.22	13.45	18.53	120.78	27.46
Face	**6.28**	*6.39*	7.44	6.35	124.94	8.45	174.79	30.40	15.18	**5.97**	75.64	110.65	34.81
FaceOcc2	**5.06**	**5.13**	*5.41*	8.30	9.50	6.06	14.96	14.79	9.48	36.78	7.88	5.99	8.01
Jumping	*4.70*	20.24	4.82	35.33	85.28	**4.50**	10.42	5.91	**4.48**	76.35	67.08	5.92	86.27
CupTable	**1.90**	2.39	**2.13**	2.44	2.22	*2.23*	11.55	2.51	7.61	82.45	17.73	3.55	2.54
David	*2.91*	3.02	**2.78**	**2.82**	12.69	11.83	48.15	55.30	11.20	102.95	20.36	52.00	5.96
David2	**1.07**	**1.39**	5.03	1.40	1.59	3.59	11.27	4.10	6.46	136.04	17.29	**1.39**	1.75
Caviar2	**1.72**	4.64	**2.35**	5.30	64.45	*2.53*	69.82	11.78	39.87	13.78	24.48	8.14	60.82

TABLE II
AVERAGE PER-FRAME OVERLAP RATIO, THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE, AND GREEN FONTS

Sequence	SCCT	SCCT-CI	LSST	IVT	MTT	SCM	MIL	LOT	TLD	BLUT	DT	Struck	LSHT
Tiger1	**0.604**	0.431	0.301	0.295	0.330	0.133	0.451	0.124	*0.463*	0.209	**0.751**	0.194	0.105
Box	**0.725**	0.650	0.680	*0.703*	0.197	0.254	0.051	0.092	0.134	0.132	0.337	**0.817**	0.339
Lemming	**0.717**	**0.713**	*0.696*	0.598	0.139	0.421	0.370	0.580	0.083	0.142	0.171	0.566	0.403
Car11	**0.885**	*0.836*	**0.838**	0.804	0.789	0.815	0.238	0.146	0.383	0.013	0.497	0.827	0.545
Body	0.631	**0.706**	*0.682*	0.080	0.603	**0.712**	0.082	0.308	0.517	0.666	0.097	0.182	0.677
Car4	**0.886**	0.735	0.834	*0.839*	0.724	**0.840**	0.060	0.553	0.722	0.769	0.719	0.264	0.707
Face	**0.853**	*0.845*	0.835	**0.849**	0.255	0.841	0.181	0.518	0.712	0.827	0.330	0.269	0.597
FaceOcc2	**0.783**	*0.780*	0.771	0.699	0.749	*0.780*	0.652	0.461	0.683	0.251	0.771	**0.786**	0.761
Jumping	**0.661**	0.398	0.643	0.277	0.080	**0.669**	0.512	0.603	0.628	0.016	0.111	*0.660*	0.100
CupTable	**0.851**	0.818	**0.852**	0.803	0.810	*0.840*	0.438	0.790	0.573	0.166	0.417	0.711	0.730
David	**0.773**	**0.777**	0.757	**0.773**	0.525	0.593	0.193	0.100	0.502	0.045	0.500	0.394	0.682
David2	0.815	0.806	0.516	0.737	*0.841*	0.740	0.454	0.624	0.611	0.015	0.547	**0.883**	**0.857**
Caviar2	*0.752*	0.606	**0.804**	0.614	0.376	**0.811**	0.259	0.599	0.429	0.625	0.420	0.574	0.334

GL likelihood alone, a second peak occurs around the target in likelihood maps. By taking SCC into consideration, single peak likelihood maps are produced where the second peak is flattened, which means the negative effects from the distracter are reduced.

B. Quantitative Evaluation

Two evaluation criteria are used in performance evaluation of the proposed *SCCT*: center location error and overlap ratio. Both of them are computed against the published manually labelled ground truth. Table I reports the average per-frame center location error $\tilde{e}$ (in pixel) calculated as follows:

$$\tilde{e} = \frac{\sum_{t=1}^{N_q} \|(\hat{x}_t - x_t^{gt}, \hat{y}_t - y_t^{gt})\|}{N_q} \tag{34}$$

where N_q is the total number of frames, $(\hat{x}_t, \hat{y}_t)$ is the center of estimated bounding box and (x_t^{gt}, y_t^{gt}) is center of the manually labeled bounding box. In Table II, the average per-frame overlap ratio $\tilde{r}$ is demonstrated

$$\tilde{r} = \frac{\sum_{t=1}^{N_q} \frac{area(R_T \bigcap R_G)}{area(R_T \bigcup R_G)}}{N_q} \tag{35}$$

where N_q is the total number of frames, R_T and R_G are the estimated and ground truth bounding boxes respectively. It is clear that in most sequences with motion-related challenges, the proposed *SCCT* can obtain better performance than other existing algorithms (in top three best performance among the compared algorithms for most video sequences).

Since *ART* [20] is the most related work which considers the observation dependency cue in HMM as well, we provide further comparison between *SCCT* and *ART* on the public sequences mentioned in [20].[1] The compared results of average centre location error are shown in Table III. With the eight evaluated video sequences, the proposed *SCCT* performs much better on four sequences, while *ART* performs better on the rest four. If we look into *Tiger1*, *Tiger2*, *Sylvester* and *Shaking* on which *ART* shows better performance, all the targets in these sequences are with repetitive movements.

ART can obtain good performance on these video sequences by learning and clustering the previous target appearance. The proposed *SCCT* provides a more general way in predicting the target appearance change with no pre-knowledge of the target movement in the coming frames.

[1]Since the code of *ART* is not released, comparison results of the algorithms ART [20], VTD [37], MIL [10], IVT [1], and FRAGT [38] are obtained from the study [20].

TABLE III
AVERAGE PER-FRAME CENTER LOCATION ERROR (IN PIXEL), THE BEST TWO RESULTS ARE SHOWN IN RED AND BLUE FONTS

Sequence	SCCT	ART	VTD	MIL	IVT	FRAGT
Girl	**5.9**	**10.6**	14.6	33.1	55.6	20.4
David	**2.9**	**3.3**	46.9	24.9	42.8	27.5
Tiger1	**20.5**	**4.9**	44.5	30.3	55.7	24.8
Tiger2	**29.8**	**5.4**	53.1	11.9	48.9	36.7
FaceOcc1	**7.5**	**8.1**	9.8	35.4	10.9	9.3
FaceOcc2	**5.1**	**6.0**	54.1	15.5	12.8	63.9
Sylvester	**11.6**	**5.9**	23.1	14.8	120.9	15.7
Shaking	**20.1**	**7.7**	78.2	42.7	96.5	194.4

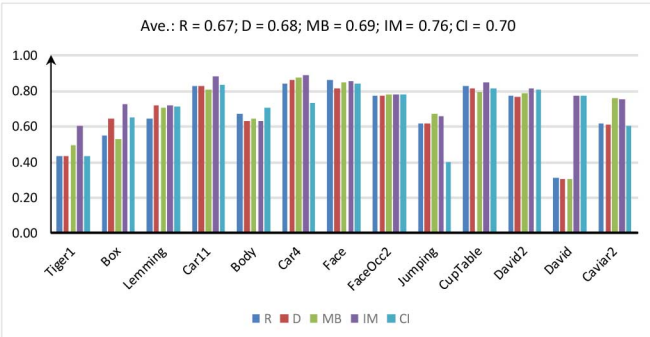


Fig. 7. Tracking performance with separate attribute in SCC. R, D, and MB denote rotation, deformation, and motion blur modelled SCC, respectively. IM denotes the integrated SCC model which consider all the three listed motion attributes and CI denotes that the SCC is set to be an identity matrix where the observation dependency is not considered. Average of the OR on all tested sequences are on the top of the figure.

The processing time of the proposed *SCCT* is tested on a PC with Intel E5-1650 CPU (3.2 GHz) and 16 GB memory. It runs at 5 frames per second on average.

C. Separated Model Tested in SCC

In this section, we test the proposed tracking algorithm by model SCC with three motion attributes (rotation, deformation and motion blur) separately to clarify the contribution of each attribute to the final SCC. We conduct three more experiments where the appearance stability $P(\mathbf{I})$ in (9) are computed only based on $P_r(\mathbf{I})$ (rotation), $P_d(\mathbf{I})$ (deformation) and $P_b(\mathbf{I})$ (motion blur) respectively.

The comparison result is shown in Fig. 7. Overlap ratio (OR) is used to measure the tracking performance. We can see that the SCC with integrated models results in the best tracking performance in most video sequences and obtain the highest average OR. By only using rotation, deformation or motion blur in SCC alone, the performance in tracking is worse than that using the integrated model and even slightly worse than that without using the SCC. This justifies that the integration of three attributes in SCC is necessary in visual tracking. Since all the three attributes of rotation, deformation and motion blur are very common in video sequences, only considering one attribute in modelling would possibly make the SCC sensitive to other attributes and thus results in poor performance.

D. Qualitative Evaluation

In Fig. 8, we show some sample frames of the comparison experiment between the proposed *SCCT* and some relevant algorithms.[2] For presentation clarity, we only draw the result of some well performed relevant algorithms, i.e., *LSST* [2], *IVT* [1], *SCM* [30], *TLD* [24], *BLUT* [32], *Struck* [34] and *LSHT* [5].

Rotation: The sequences used in comparison experiments with target rotation include both in-plane-rotation (e.g., *David2* and *FaceOcc2*) and out-plane-rotation (e.g., *Box*, *Lemming*, *David*, *Tiger1*, *Car11* and *CupTable*). In can be seen from Fig. 8 that, it is easy for most compared trackers to localize the targets in the in-plane-rotation sequences. This is because the in-plane-rotation can be treated as a 2-D motion. And with orientation invariant features, the target appearance in the bounding box doesn't change much. However, for out-of-plane rotation cases, the target appearance in the bounding box would change and new appearance will be introduced. It is difficult for the existing trackers to capture the target accurately. It is demonstrated that in most rotation sequences, the proposed *SCCT* can obtain accurate tracking results by predicting the appearance change caused by rotation. Since out-of-plane rotation only changes the target appearance in some local regions between consecutive frames, with local sparse representation, *Struck* and *LSHT* work well in some rotation sequences like *CupTable* and *Box*. *LSST* and *IVT* also perform well in some rotation sequences. The reason is that the PCA representation is somehow robust in tracking multi-view objects. *DT* performs quite well on *Tiger* due to the benefits from its large basin of attraction and model updating frame by frame. There is no model updating in *BLUT* and thus it loses tracking targets in most rotation sequences.

Motion Blur: As indicated in Fig. 8, in sequences of *Jumping*, *Face*, *Box*, *Lemming*, *Body* and *Car4*, the targets are blurred with different degrees by fast motion of either target or camera, which makes the target indistinguishable against its surroundings. The proposed *SCCT* can obtain more accurate and stable tracking results on these sequences, since the SCC is able to predict the appearance change caused by motion blur. *LSST* is also able to localize the target in most video frames. *SCM* performs well in some blurred sequences benefiting from its discriminative model which makes the blurred target more distinguish against the background. However, it loses tracking when there are other challenges co-existing, like occlusion and rotation in *Lemming* and *Box*. *BLUT* is specifically designed for tracking blurred objects, and it is demonstrated to perform well on *Face* and *Body*. However, without model updating, it loses tracking on other sequences with appearance variation.

Deformation: In sequences of *Body* and *Caviar2* in Fig. 8, it can be seen that the pose variation during walking causes the deformation of target pedestrians. Only the proposed *SCCT*, *SCM* and *LSST* can localize the target in most frames from these two video sequences. *Struck* and *IVT* can well localize the target pedestrian in *Caviar2* since their appearance models can adjust to the appearance variation caused by deformation. However, they lose tracking on *Body* when motion blur also exists.

[2][Online]. Available: https://sites.google.com/site/sorsyuanyuan/home/odt-project-page

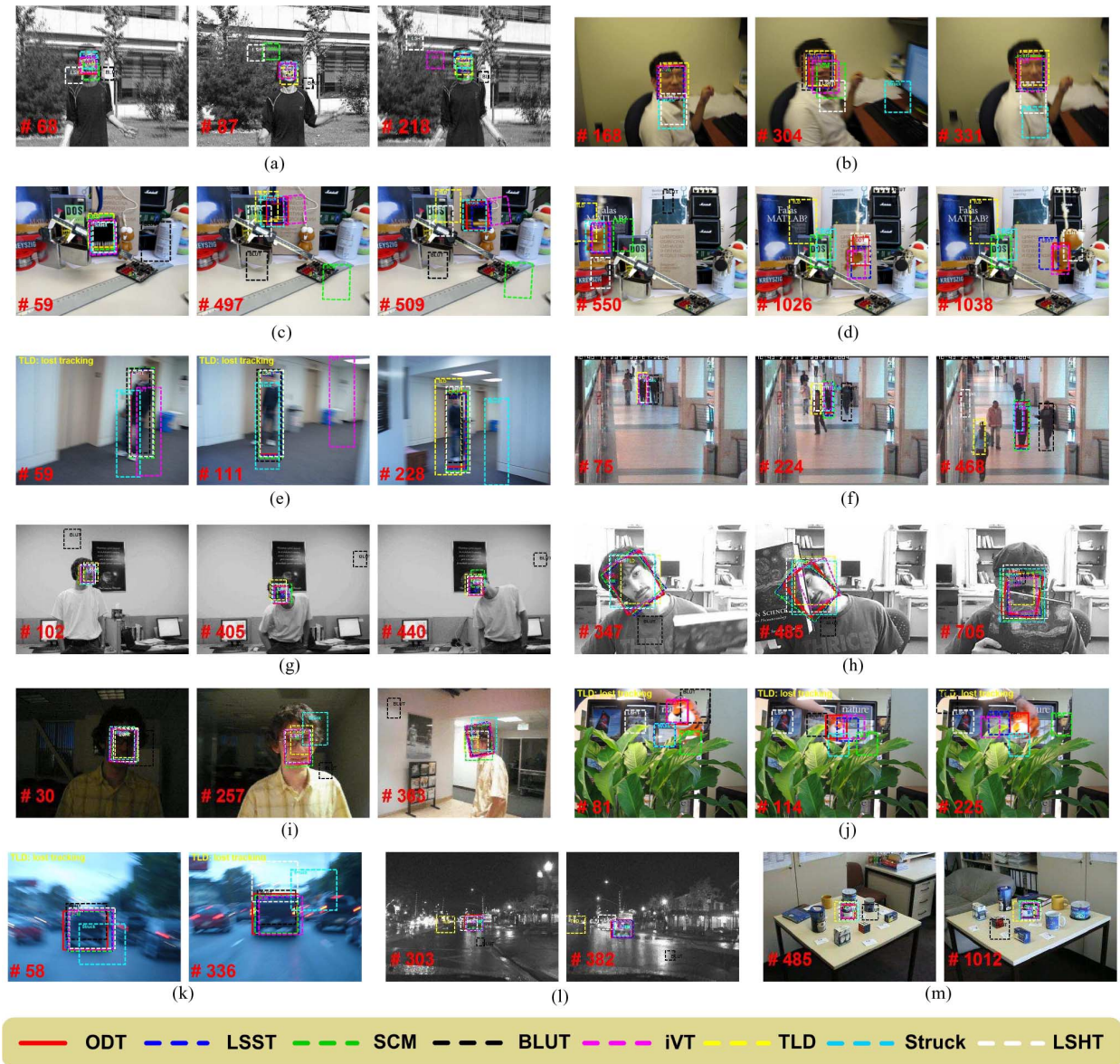


Fig. 8. Visual illustration of tracking results 13 sequences with motion blur, deformation, rotation, occlusion, illumination variation, etc. (a) *Jumping* with *motion blur*. (b) *Face* with *motion blur*. (c) *Box* with *motion blur*, *rotation* and *occlusion*. (d) *Lemming* with *motion blur*, *rotation* and *occlusion*. (e) *Body* with *motion blur* and *deformation*. (f) *Caviar2* with *occlusion* and *deformation*. (g) *David2* with *rotation*. (h) *FaceOcc2* with *rotation* and *occlusion*. (i) *David* with *rotation*, and *illumination variation*. (j) *Tiger1* with *rotation*, *occlusion* and *illumination variation*. (k) *Car4* with *motion blur*. (l) *Car11* with *illumination variation* and *rotation*. (m) *CupTable* with *rotation*.

Occlusion: Occlusion is one of the most common challenges in visual object tracking. It is easy for trackers to drift if the appearance model is not robust enough to occlusion. We test the tracking algorithms on sequences (e.g., *Box*, *Lemming*, *FaceOcc2*, *Tiger1* and *Caviar2*) that the target is occluded within some time during tracking. We also provide some comparison samples in Fig. 8. It can be seen that the proposed *SCCT* is able to obtain accurate tracking results on these sequences, since the Laplacian noise used in *SCCT* can eliminate the outliers well during appearance model updating. *LSST* can also obtain good results on some sequences. However, when there are challenges like rotation and motion blur (e.g., *Box* and *Tiger1*), it can not perform well stably. Since the local features are effective in dealing with partial occlusion, it can be seen that *SCM*, *Struck* and *LSHT* can also obtain good tracking results on some sequences with partial occlusion (e.g., *Box*, *Caviar2* and *FaceOcc2*).

Illumination Variation: In Fig. 8, *David*, *Tiger1* and *Car11* are three test sequences with illumination variation. Since the IPCA representation is robust to illumination variation, the proposed *SCCT* obtains accurate tracking results on these three sequences. Similarly, *LSST* and *IVT* also produce good results on *David* and *Car11*. However, both of them lose tracking when there is motion-related challenges in sequence *Tiger*. Other algorithms perform fairly on some sequences, but no one can get stable tracking results on all the video sequences.

VI. Conclusion

In this paper, we have proposed an OD-HMM based visual object tracking algorithm by SCC to address the motion related appearance change problems. The SCC is defined to predict the appearance stability of moving targets. With SCC, we compensate the reconstruction error while estimating the likelihood of candidates with the appearance model, and the motion related appearance change problem during tracking is addressed. Additionally, an effective model updating mechanism is investigated to remove outliers. Both qualitative and quantitative comparison experiments demonstrate the better performance of the proposed OD-HMM based visual object tracking algorithm against state-of-art tracking algorithms by providing more stable and accurate results on the challenging video sequences.

References

[1] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 125–141, 2008.

[2] D. Wang, H. Lu, and M.-H. Yang, "Least soft-threshold squares tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2371–2378.

[3] Y. Yuan, E. Sabu, F. Yuming, and L. Weisi, "Visual object tracking based on backward model validation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1898–1910, Nov. 2014.

[4] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[5] S. He, Q. Yang, R. W. Lau, J. Wang, and M.-H. Yang, "Visual tracking via locality sensitive histograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2427–2434.

[6] P. Cui, L.-F. Sun, F. Wang, and S.-Q. Yang, "Contextual mixture tracking," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 333–341, Feb. 2009.

[7] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2411–2418.

[8] T.-K. Kim, I.-M. Cho, and J.-H. Lee, "Illumination-invariant object tracking method and image editing system using the same," U.S. Patent 7,171,023, Jan. 30, 2007.

[9] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf.*, 2006, vol. 1, pp. 47–56.

[10] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 983–990.

[11] D. Wang and H. Lu, "On-line learning parts-based representation via incremental orthogonal projective non-negative matrix factorization," *Signal Process.*, vol. 93, no. 6, pp. 1608–1623, 2013.

[12] J. Kwon and K. M. Lee, "Highly nonrigid object tracking via patch-based dynamic appearance modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2427–2441, Oct. 2013.

[13] C.-T. Chu, J.-N. Hwang, H.-I. Pai, and K.-M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1602–1615, Nov. 2013.

[14] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 467–476, Dec. 1995.

[15] A. Ginsburg, "Pattern recognition techniques suggested from psychological correlates of a model of the human visual system," *Proc. NAECON 73*, pp. 309–316, 1973.

[16] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2042–2049.

[17] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surveys*, vol. 38, no. 4, p. 13, 2006.

[18] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[19] X. Li, M. Parizeau, and R. Plamondon, "Training hidden Markov models with multiple observations-a combinatorial method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 371–377, Apr. 2000.

[20] D. W. Park, J. Kwon, and K. M. Lee, "Robust visual tracking using autoregressive hidden Markov model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1964–1971.

[21] L. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 810–815, Jun. 2004.

[22] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1296–1311, Oct. 2003.

[23] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 234–247.

[24] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[25] M. Potmesil and I. Chakravarty, "Modeling motion blur in computer-generated images," *ACM SIGGRAPH Comput. Graphics*, vol. 17, no. 3, pp. 389–399, Jul. 1983.

[26] R. Liu, Z. Li, and J. Jia, "Image partial blur detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[27] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[28] P. J. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *J. Opt. Soc. Amer. A*, vol. 19, no. 6, pp. 1096–1106, 2002.

[29] D. J. Field and N. Brady, "Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes," *Vis. Res.*, vol. 37, no. 23, pp. 3367–3383, 1997.

[30] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1838–1845.

[31] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1940–1947.

[32] Y. Wu *et al.*, "Blurred target tracking by blur-driven tracker," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1100–1107.

[33] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1910–1917.

[34] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 263–270.

[35] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 723–730.

[36] D. A. Klein, D. Schulz, S. Frintrop, and A. B. Cremers, "Adaptive real-time video-tracking for arbitrary objects," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 772–777.

[37] J. Kwon and K. Lee, "Visual tracking decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1269–1276.

[38] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 798–805.

Yuan Yuan (S'14) received the B.E. degree in electronic engineering from the Beijing University of Post and Telecommunication, Beijing, China, in 2011, and is currently working toward the Ph.D. degree at the School of Computer Engineering, Nanyang Technological University, Singapore.

His current research interests include visual surveillance, object tracking, and visual attention.

Huan Yang received the B.S. degree in computer science from the Heilongjiang Institute of Technology, Harbin, China, in 2007, the M.S. degree in computer science from Shandong University, Jinan, China, in 2010, and is currently working toward the Ph.D. degree at the School of Computer Engineering, Nanyang Technological University, Singapore.

Her current research interests include object detection/recognition, image and video processing, and visual quality assessment.

Yuming Fang received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree in computer engineering from the Nanyang Technological University, Singapore.

He is currently an Associate Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He was previously a Visiting Postdoc Research Fellow with the IRCCyN Laboratory, Polytech Nantes, University of Nantes, Nantes, France, the University of Waterloo, Waterloo, ON, Canada, and the Nanyang Technological University, Singapore. His current research interests include multimedia processing and computer vision.

Weisi Lin (S'91–M'92–SM'00) received the Ph.D. degree from Kings College London, London, U.K.

He was previously the Lab Head of Visual Processing with the Institute for Infocomm Research, Singapore. Currently, he is an Associate Professor with the School of Computer Engineering, Institute for Infocomm Research. He has authored or coauthored over 120 journal papers and 200 conference papers, filed 7 patents, and authored 2 books. His current research interests include image processing, perceptual signal modeling, video compression, and multimedia communication.

Dr. Lin is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, and the *Journal of Visual Communication and Image Representation*, and was previously an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA (2011–2013). He also served as a Guest Editor for six special issues of international journals. He has been a Technical Program Chair for IEEE ICME 2013, PCM 2012, and QoMEX 2014. He chaired the IEEE MMTC Special Interest Group on QoE (2012–2014). He has been an invited/panelist/keynote/tutorial speaker at over 10 international conferences, as well as a Distinguished Lecturer of the Asia-Pacific Signal and Information Processing Association (2012–2013).