Visual Tracking Via Incremental Log-Euclidean Riemannian Subspace Learning

Xi Li[†], Weiming Hu[†], Zhongfei Zhang[‡], Xiaoqin Zhang[†], Mingliang Zhu[†], Jian Cheng[†]
[†]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[†]{lixi, wmhu, xqzhang, mlzhu, jcheng}@nlpr.ia.ac.cn
[‡]State University of New York, Binghamton, NY 13902, USA
[‡]zhongfei@cs.binghamton.edu

Abstract

Recently, a novel Log-Euclidean Riemannian metric [28] *is proposed for statistics on symmetric positive definite (SPD) matrices. Under this metric, distances and Riemannian means take a much simpler form than the widely used affine-invariant Riemannian metric. Based on the Log-Euclidean Riemannian metric, we develop a tracking framework in this paper. In the framework, the covariance matrices of image features in the five modes are used to represent object appearance. Since a nonsingular covariance matrix is a SPD matrix lying on a connected Riemannian manifold, the Log-Euclidean Riemannian metric is used for statistics on the covariance matrices of image features. Further, we present an effective online Log-Euclidean Riemannian subspace learning algorithm which models the appearance changes of an object by incrementally learning a low-order Log-Euclidean eigenspace representation through adaptively updating the sample mean and eigenbasis. Tracking is then led by the Bayesian state inference framework in which a particle filter is used for propagating sample distributions over the time. Theoretic analysis and experimental evaluations demonstrate the promise and effectiveness of the proposed framework.*

1. Introduction

For visual tracking, handling appearance variations of an object is a fundamental and challenging task. In general, there are two types of appearance variations: intrinsic and extrinsic. Pose variation and/or shape deformation of an object are considered as the intrinsic appearance variations while the extrinsic variations are due to the changes resulting from different illumination, camera motion, camera viewpoint, and occlusion. Consequently, effectively modeling such appearance variations plays a critical role in visual tracking.

Hager and Belhumeur [1] propose a tracking algorithm which uses an extended gradient-based optical flow method to handle object tracking under varying illumination conditions. In [3], curves or splines are exploited to represent the appearance of an object to develop the Condensation algorithm for contour tracking. Due to the simplistic representation scheme, the algorithm is unable to handle the pose or illumination change, resulting in tracking failures under a varying lighting condition. Zhao *et al.*[22] present a fast differential EMD tracking method which is robust to illumination changes. Silveira and Malis [17] present a new algorithm for handling generic lighting changes.

Black *et al.*[4] employ a mixture model to represent and recover the appearance changes in consecutive frames. Jepson *et al.*[5] develop a more elaborate mixture model with an online EM algorithm to explicitly model appearance changes during tracking. Zhou *et al.*[6] embed appearance-adaptive models into a particle filter to achieve a robust visual tracking. Wang *et al.*[23] present an adaptive appearance model based on the Gaussian mixture model (GMM) in a joint spatial-color space (referred to as SMOG). SMOG captures rich spatial layout and color information. Yilmaz [16] proposes an object tracking algorithm based on the asymmetric kernel mean shift with adaptively varying the scale and orientation of the kernel. Nguyen *et al.*[19] propose a kernel-based tracking approach based on maximum likelihood estimation.

Yu *et al.*[7] propose a spatial-appearance model which captures non-rigid appearance variations and recovers all motion parameters efficiently. Li *et al.*[8] use a generalized geometric transform to handle the deformation, articulation, and occlusion of appearance. Ilic and Fua [20] present a non-linear beam model for tracking large deformations. Tran and Davis [21] propose robust regional affine invariant image features for visual tracking. Grabner *et al.*[18] develop a keypoint matching-based tracking method by online learning classifier-based keypoint descriptions.

Lee and Kriegman [9] present an online learning algorithm to incrementally learn a generic appearance model for video-based recognition and tracking. Lim *et al.*[10] present a human tracking framework using robust system dynamics identification and nonlinear dimension reduction techniques. Black *et al.*[2] present a subspace learning based tracking algorithm with the subspace constancy assumption. A pre-trained, view-based eigenbasis representation is used for modeling appearance variations. However, the algorithm does not work well in the scene clutter with a large lighting change due to the subspace constancy assumption. Ho *et al.*[11] present a visual tracking algorithm based on linear subspace learning. Li *et al.*[12] propose an incremental PCA algorithm for subspace learning. In [13], a weighted incremental PCA algorithm for subspace learning is presented. Limy *et al.*[14] propose a generalized tracking

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

framework based on the incremental image-as-vector subspace learning methods with a sample mean update. In [15], Li *et al.* present a visual tracking framework based on online tensor decomposition. In [33], zhang *et al.* propose a graph embedding based discriminative learning framework for robust and efficient object tracking.

However, the above appearance-based tracking methods share a problem that their appearance models lack a competent object description criterion that captures both statistical and spatial properties of object appearance. As a result, they are usually sensitive to the variations in illumination, view, and pose. In order to tackle this problem, Tuzel *et al.* [29] and Porikli *et al.*[24] propose a covariance matrix descriptor for characterizing the appearance of an object. The covariance matrix descriptor, based on several covariance matrices of image features, is capable of fully capturing the information of the variances and the spatial correlations of the extracted features inside an object region. In particular, the covariance matrix descriptor is robust to the variations in illumination, view, and pose. Since a nonsingular covariance matrix is a symmetric positive definite (SPD) matrix lying on a connected Riemannian manifold, statistics for covariance matrices of image features may be computed through Riemannian geometry. Thus, we then give a brief review of some related work on Riemannian geometry as follows.

Tuzel *et al.*[25] present a new algorithm for human detection through classification on Riemannian manifolds. Fletcher and Joshi [26] make principal geodesic analysis on symmetric spaces in which diffusion tensors lie. Lin and Zha [27] present a Riemannian manifold learning framework which simplifies the dimension reduction problem into a classical problem in Riemannian geometry. Nevertheless, these existing algorithms for statistics on a Riemannian manifold are based on the affine-invariant Riemannian metric, under which the Riemannian mean has no closed form. Generally, an iterative numerical procedure [30] is applied to compute the Riemannian mean. Recently, Arsigny *et al.*[28] propose a novel Log-Euclidean Riemannian metric for statistics on SPD matrices. Under this metric, distances and Riemannian means take a much simpler form than the affine-invariant Riemannian metric.

Based on the Log-Euclidean Riemannian metric [28], we develop a tracking framework in this paper. The main contributions of the developed framework are as follows. First, the framework does not need to know any prior knowledge of the object. A low dimensional Log-Euclidean Riemannian eigenspace representation is learned online, and updated incrementally over the time. The framework only assumes that the initialization of the object region is provided. Second, while the Condensation algorithm [3] is used for propagating the sample distributions over the time, we develop an effective probabilistic likelihood function based on the learned Log-Euclidean Riemannian eigenspace model. Last, while R-SVD [14, 32] is applied to update both the sample mean and eigenbasis online as new data arrive, an incremental Log-Euclidean Riemannian subspace learning procedure is enabled to capture the appearance characteristics of the object during the tracking.

Before starting the discussion on the proposed tracking framework, we first give a brief review of the related background, including the covariance matrix descriptor in Sec.2 and Riemannian geometry for SPD matrices in Sec. 3.

2. Covariance matrix descriptor

Tuzel *et al.* [29] propose a novel covariance matrix descriptor with details described as follows. Denote I as a $W \times H$ one-dimensional intensity or three-dimensional color image, and F as the $W \times H \times d$ dimensional feature image extracted from I.

$$F(x, y) = \psi(I, x, y), \tag{1}$$

where ψ is a function for extracting image features such as intensity, color, gradients, and filter responses. For a given rectangular region $R \subset I$, denote $\{f_i\}_{i=1,\ldots,L}$ as the d-dimensional feature points obtained by ψ within R. Consequently, the image region R can be represented as a $d \times d$ covariance matrix:

$$\mathbf{C}_R = \frac{1}{L-1}\sum_{i=1}^{L}(f_i - \mu)(f_i - \mu)^T \tag{2}$$

where μ is the mean of $\{f_i\}_{i=1\ldots L}$. For the tracking issue in this paper, the mapping function $\psi(I, x, y)$ is defined as $(x, y, (\mathbb{E}_i)_{i=1\ldots N})$ where (x, y) is the pixel location; N is the number of I's color channels; and $\mathbb{E}_i$ is formulated as:

$$\left(I^i, |I^i_x|, |I^i_y|, \sqrt{\left(I^i_x\right)^2 + \left(I^i_y\right)^2}, |I^i_{xx}|, |I^i_{yy}|, \arctan\frac{|I^i_y|}{|I^i_x|}\right) \tag{3}$$

where I^i_x, I^i_{xx}, I^i_y, and I^i_{yy} are intensity derivatives in the ith color channel, I^i is the intensity value in the ith color channel, $|\cdot|$ is a function returning the absolute value of its argument, and the last term stands for the first-order gradient orientation. If I is a grayscale image, $F(x, y)$ is an 9-dimensional feature image (i.e., $N = 1$ and d=9); otherwise, $F(x, y)$ is a 23-dimensional feature image (i.e., $N = 3$ and d=23). Consequently, the covariance matrix descriptors of a grayscale and a color image regions are 9×9 and 23×23 symmetric matrices, respectively.

3. Riemannian geometry for SPD matrices

Hereinafter, denote $Sym(n)$ as the space of real $n \times n$ symmetric matrices, and $Sym^+(n)$ as the space of real $n \times n$ SPD matrices.

In $Sym(n)$, there are two fundamental operations—matrix exponential and logarithm, which can be computed easily as follows. Given a symmetric matrix $\mathbf{A} \in Sym(n)$, the singular value decomposition (SVD) of $\mathbf{A}$ is denoted as $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{U}$ is an an orthonormal matrix, and $\mathbf{\Sigma} = \text{Diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of the eigenvalues. Hence, the matrix exponential exp($\mathbf{A}$) is formulated as:

$$\begin{aligned}\exp(\mathbf{A}) &= \textstyle\sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \\ &= \mathbf{U}\cdot\mathrm{Diag}\left(\exp(\lambda_1), \ldots, \exp(\lambda_n)\right)\cdot\mathbf{U}^T.\end{aligned} \tag{4}$$

Similarly, the matrix logarithm $\log(\mathbf{A})$ has the following form:

$$\begin{aligned}\log(\mathbf{A}) &= \textstyle\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(\mathbf{A}-\mathbf{I_n})^k \\ &= \mathbf{U}\cdot\mathrm{Diag}\left(\log(\lambda_1), \ldots, \log(\lambda_n)\right)\cdot\mathbf{U}^T,\end{aligned} \tag{5}$$

where $\mathbf{I_n}$ is an n-by-n identity matrix. In particular, the matrix exponential always exists, whereas the matrix logarithm is only available for SPD matrices.

In $Sym^+(n)$, SPD matrices lie on a connected Riemannian manifold. Consequently, Riemannian metrics should be used for statistics on SPD matrices. Typically, there exist two invariant Riemannian metrics in $Sym^+(n)$. One is the widely used affine-invariant Riemannian metric, and the other is the recently introduced Log-Euclidean Riemannian metric.

Under the affine-invariant Riemannian metric, there is no closed form for the Riemannian mean of several SPD matrices. Generally, an iterative numerical procedure [30] is applied to compute the Riemannian mean. Furthermore, the distance between two points $\mathbf{X}$ and $\mathbf{Y}$ in $Sym^+(n)$ under the affine-invariant Riemannian metric is computed by $\|\log(\mathbf{X}^{-1/2}\cdot\mathbf{Y}\cdot\mathbf{X}^{-1/2})\|$.

Under the Log-Euclidean Riemannian metric, SPD matrices lie in a Lie group $\mathcal{G}$. The tangent space at the identity element in $\mathcal{G}$ forms a Lie algebra $\mathcal{H}$ which is a vector space. Consequently, the Riemannian mean μ of several elements in $\mathcal{H}$ is simply an arithmetic mean of matrix logarithms. Correspondingly, the Riemannian mean $\boldsymbol{\mu}^*$ of several elements in $\mathcal{G}$ is computed by mapping back the Riemannian mean μ with the matrix exponential $\exp(\cdot)$. For example, given N SPD matrices $\{\mathbf{X}_i\}_{i=1}^N$, the mean μ corresponding to the Lie algebra $\mathcal{H}$ is explicitly computed by $\mu = \frac{1}{N}\sum_{i=1}^N \log(\mathbf{X}_i)$, and the mean corresponding to the Lie group $\mathcal{G}$ is obtained from $\boldsymbol{\mu}^* = \exp\left(\frac{1}{N}\sum_{i=1}^N \log(\mathbf{X}_i)\right)$. For more details of Lie groups and Lie algebras, refer to [31]. Moreover, the distance between two points $\mathbf{X}$ and $\mathbf{Y}$ in $Sym^+(n)$ under the Log-Euclidean Riemannian metric is easily calculated by $\|\log(\mathbf{Y}) - \log(\mathbf{X})\|$.

Clearly, distances and Riemannian means under the Log-Euclidean metric take a much simpler form than those under the affine-invariance metric. See more details of these two metrics in [28]. In this paper, statistics on SPD matrices are made in the Lie algebra $\mathcal{H}$.

4. The framework for visual tracking

In Sec.4.1, we first give an overview of the proposed tracking framework. Then, our object representation method is introduced in Sec. 4.2. Subsequently, our proposed incremental Log-Euclidean Riemannian subspace learning algorithm (*IRSL*) is detailedly described in Sec. 4.3. Finally, we discuss how to make Bayesian state inference for visual tracking in Sec. 4.4.

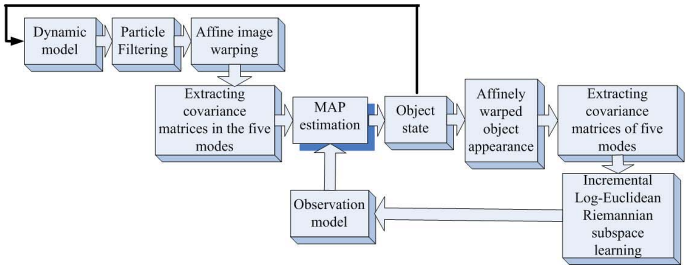
Dynamic model
Particle Filtering
Affine image warping
Extracting covariance matrices in the five modes
MAP estimation
Object state
Affinely warped object appearance
Extracting covariance matrices of five modes
Incremental Log-Euclidean Riemannian subspace learning
Observation model

Figure 1. **The architecture of the tracking framework.**

4.1. Overview of the framework

The tracking framework includes two stages: (a) Log-Euclidean Riemannian subspac learning; and (b) Bayesian inference for visual tracking. In the first stage, a low dimensional Log-Euclidean Riemannian eigenspace model is learned online. The model uses the proposed incremental Log-Euclidean Riemannian subspace learning algorithm (called *IRSL*) to find the dominant projection subspaces of the Log-Euclidean unfolding matrices in the five modes. In the second stage, the object locations in consecutive frames are obtained by maximum a posterior (MAP) estimation within the Bayesian state inference framework in which a particle filter is applied to propagate sample distributions over the time. After MAP estimation, we just use the Log-Euclidean covariance matrices of image features inside the affinely warped image region associated with the highest weighted hypothesis to update the Log-Euclidean Riemannian eigensapace model. These two stages are executed repeatedly as time progresses. Moreover, the framework has a strong adaptability in the sense that when new image data arrive, the Log-Euclidean Riemannian eigenspace model follows the updating online. The architecture of the framework is shown in Figure 1.

4.2. Object representation

In our tracking framework, an object is represented by five covariance matrices (from Eq. (2)) of the image features inside the object region. These five covariance matrices correspond to the five modes of the object appearance, respectively. We call the covariance matrix of the i-th mode as the mode-i covariance matrix for $1 \leq i \leq 5$, as exemplified in the upper parts of Figure 2(b)-(f). As time progresses, the mode-i covariance matrices $\{\mathbf{C}_{(i)}^t \in Sym^+(d)\}_{t=1,2,\ldots,N}$ constitute a mode-i covariance tensor $\mathcal{A}_{(i)} \in \mathcal{R}^{d\times d\times N}$, as exemplified in the lower parts of Figure 2(b)-(f). If $\mathbf{C}_{(i)}^t$ is a singular matrix, we replace $\mathbf{C}_{(i)}^t$ with $\mathbf{C}_{(i)}^t + \epsilon\mathbf{I}_d$, where ϵ is a very small positive constant ($\epsilon = 1\mathrm{e} - 18$ in the experiments), and $\mathbf{I}_d$ is a $d \times d$ identity matrix. By the Log-Euclidean mapping (5), we transform the mode-i covariance tensor $\mathcal{A}_{(i)}$ into a new one:

$$\mathcal{LA}_{(i)} = \{\log(\mathbf{C}_{(i)}^t)\}_{t=1,2,\ldots,N} \tag{6}$$

which is called the Log-Euclidean covariance tensor. Due to the vector space structure of $\log(\mathbf{C}_{(i)}^t)$ under the Log-Euclidean Riemannian metric, $\log(\mathbf{C}_{(i)}^t)$ is unfolded into a d^2-dimensional vector $\mathbf{vec}_{(i)}^t$ which is formulated as:

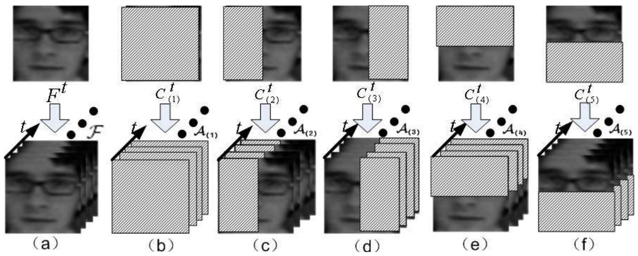

Figure 2. **Object representation using covariance matrices and tensors in the five modes. A face image F^t at time t is shown in the upper part of (a) while a 3-order face tensor $\mathcal{F} = \{F^t\}_{t=1,2,...}$ (i.e., face image ensemble) is displayed in the lower one of (a). The upper parts of (b)-(f) correspond to the covariance matrices (i.e., $C^t_{(i)}$ for $1 \le i \le 5$) of image features in the five modes while the lower ones are associated with the corresponding covariance tensors (i.e., $\mathcal{A}_{(i)}$ for $1 \le i \le 5$).**

$$\mathbf{vec}^t_{(i)} = \mathrm{UT}(\log(\mathbf{C}^t_{(i)})) = (c_1^{t(i)}, c_2^{t(i)}, \ldots, c_{d^2}^{t(i)})^T \quad (7)$$

where $\mathrm{UT}(\cdot)$ is an operator unfolding a matrix into a column vector. The unfolding process can be illustrated by Figure 3(a), where the left part displays the mode-i covariance tensor $\mathcal{A}_{(i)} \in \mathcal{R}^{d \times d \times N}$ for $1 \le i \le 5$, the middle part corresponds to the mode-i Log-Euclidean covariance tensor $\mathcal{LA}_{(i)}$ $(1 \le i \le 5)$, and the right part is associated with the mode-i Log-Euclidean unfolding matrix $\mathrm{LA}_{(i)}$ $(1 \le i \le 5)$ with the t-th column being $\mathbf{vec}^t_{(i)}$ for $1 \le t \le N$. As a result, $\mathrm{LA}_{(i)}$ is formulated as:

$$\mathrm{LA}_{(i)} = \left(\mathbf{vec}^1_{(i)}\ \mathbf{vec}^2_{(i)}\ \cdots\ \mathbf{vec}^t_{(i)}\ \cdots\ \mathbf{vec}^N_{(i)}\right). \quad (8)$$

In the next section, we discuss the proposed incremental Log-Euclidean Riemannian subspace learning algorithm (*IRSL*) for the mode-i unfolding matrix $\mathrm{LA}_{(i)}$ $(1 \le i \le 5)$. *IRSL* applies the online learning technique (R-SVD [14, 32]) to find the dominant projection subspaces of $\mathrm{LA}_{(i)}$.

4.3. Incremental Log-Euclidean Riemannian subspace learning

4.3.1 Introduction to R-SVD

The classic R-SVD algorithm [32] efficiently computes the singular value decomposition (SVD) of a dynamic matrix with newly added columns or rows, based on the existing SVD. However, the R-SVD algorithm [32] is based on the zero mean assumption, leading to the failure of tracking subspace variabilities. Based on [32], Limy *et al.* [14] extends the R-SVD algorithm to computing the eigenbasis of a scatter matrix with the mean update. The following is an introduction to the operator $\mathrm{CVD}(\cdot)$ used hereinafter. Given a matrix $H = \{\mathrm{K}_1, \mathrm{K}_2, \ldots, \mathrm{K}_g\}$ and its column mean K, we let $\mathrm{CVD}(H)$ denote the SVD of the matrix $\{\mathrm{K}_1 - \mathrm{K}, \mathrm{K}_2 - \mathrm{K}, \ldots, \mathrm{K}_g - \mathrm{K}\}$.

4.3.2 Incremental Log-Euclidean Riemannian subspace learning)

Based on the extended R-SVD [14], *IRSL* presented below efficiently identifies the dominant projection subspaces of the mode-i Log-Euclidean unfolding matrix for $1 \le i \le 5$, and is capable of incrementally updating these subspaces when new data arrive. Given the $\mathrm{CVD}(\mathrm{LA}_{(i)}) = \mathrm{U}_{(i)}\mathrm{D}_{(i)}\mathrm{V}^T_{(i)}$ of the mode-i Log-Euclidean unfolding matrix $\mathrm{LA}_{(i)}(1 \le i \le 5)$ for a Log-Euclidean covariance tensor $\mathcal{LA}_{(i)} \in \mathcal{R}^{d \times d \times N}$, *IRSL* is able to efficiently compute the $\mathrm{CVD}(\mathrm{LA}^*_{(i)}) = \mathrm{U}^*_{(i)}\mathrm{D}^*_{(i)}\mathrm{V}^{*T}_{(i)}$ of the mode-i Log-Euclidean unfolding matrix $\mathrm{LA}^*_{(i)}$ for $\mathcal{LA}^*_{(i)} = \left(\mathcal{LA}_{(i)} \mid \mathcal{LF}_{(i)}\right) \in \mathcal{R}^{d \times d \times N^*}$ where $\mathcal{LF}_{(i)} \in \mathcal{R}^{d \times d \times N'}$ is a newly-added Log-Euclidean covariance subtensor and $N^* = N + N'$. To facilitate the description, Figure 3(b) is used for illustration. In the left part of Figure 3(b), the mode-i covariance tensor is shown. The white regions represent the original covariance subtensor $\mathcal{A}_{(i)}$ while the dark regions denote the newly added covariance subtensor $\mathcal{F}_{(i)}$. In the middle part of Figure 3(b), the Log-Euclidean covariance tensor is displayed. The mode-i unfolding matrix is displayed in the right part of Figure 3(b), where the dark regions represent the mode-i unfolding matrix $\mathrm{LF}_{(i)}$ of the newly added Log-Euclidean covariance subtensor $\mathcal{LF}_{(i)}$. With the emergence of the new data subtensors, the column space of $\mathrm{LA}^*_{(i)}$ is extended. Consequently, *IRSL* needs to track the change of the column space of $\mathrm{LA}^*_{(i)}$, and needs to identify the dominant projection subspace for a compact representation of $\mathrm{LA}^*_{(i)}$. Based on the CVD of the $\mathrm{LA}_{(i)}$, the CVD of $\mathrm{LA}^*_{(i)}$ is efficiently obtained by performing R-SVD on the matrix $\left(\mathrm{LA}_{(i)} \mid \mathrm{LF}_{(i)}\right)$. The specific procedure of *IRSL* is listed in Table 1.

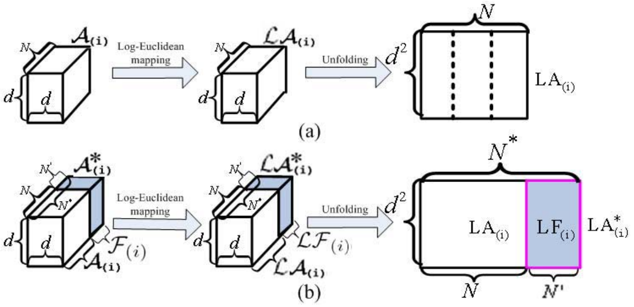

Figure 3. **Illustration of mode-i Log-Euclidean unfolding and the proposed *IRSL*. (a) shows the generative process of the mode-i Log-Euclidean unfolding matrix; (b) displays the incremental learning process of the proposed *IRSL*.**

In real tracking applications, it is necessary for a subspace analysis-based algorithm to evaluate the likelihood between the test sample and the learned subspace. In *IRSL*, the criterion for the likelihood evaluation is given as follows. Given the mode-i covariance matrix $\mathrm{TC}_{(i)} \in \mathcal{R}^{d \times d}$ $(1 \le i \le 5)$ of features inside a test image $\mathcal{T}$, and the learned Log-Euclidean Riemannian eigenspace (i.e., $\mathrm{LA}_{(i)}$'s column mean $\bar{L}_{(i)}$ and $\mathrm{CVD}(\mathrm{LA}_{(i)}) = \mathrm{U}_{(i)}\mathrm{D}_{(i)}\mathrm{V}^T_{(i)}$ for $1 \le i \le 5$), the likelihood can be determined by the sum of the reconstruction error norms of the five modes:

$$RE = \sum_{i=1}^{5} \omega_i \cdot \|(\mathbf{vec}_{(i)} - \bar{L}_{(i)}) - U_{(j)} \cdot U^T_{(j)} \cdot (\mathbf{vec}_{(i)} - \bar{L}_{(i)})\|^2 \quad (9)$$

where ω_i is the mode-i weight ($\sum_{i=1}^{5} \omega_i = 1$, and $\omega_i = 0.2$

Input:

CVD of the unfolding matrix $\mathrm{LA}_{(i)}$, i.e., $\mathrm{U}_{(i)}\mathrm{D}_{(i)}\mathrm{V}_{(i)}^T$ of a mode-i Log-Euclidean covariance tensor $\mathcal{LA}_{(i)} \in \mathcal{R}^{d\times d\times N}$, newly-added covariance tensor $\mathcal{F}_{(i)} \in \mathcal{R}^{d\times d\times N'}$, column mean $L_{(i)}$ of $\mathrm{LA}_{(i)}$, the maintained dimension R_i of the mode-i eigenspace, and $1 \leq i \leq 5$.

Output:

Column mean $\bar{L}_{(i)}^*$ of $\mathrm{LA}_{(i)}^*$, and CVD of the unfolding matrix $\mathrm{LA}_{(i)}^*$, i.e., $\mathrm{U}_{(i)}^*\mathrm{D}_{(i)}^*\mathrm{V}_{(i)}^{*T}$ of $\mathcal{LA}_{(i)}^* = \left(\mathcal{LA}_i \mid \mathcal{LF}_{(i)}\right) \in \mathcal{R}^{d\times d\times N^*}$ where $N^* = N + N'$ and $\mathcal{LF}_{(i)}$ represents the corresponding Log-Euclidean covariance tensor of $\mathcal{F}_{(i)}$.

Algorithm:

1. Obtain $\mathcal{LF}_{(i)}$ by transforming $\mathcal{F}_{(i)}$ through (6);
2. Unfold $\mathcal{LF}_{(i)}$ into $\mathrm{LF}_{(i)}$ by (8);
3. $[\mathrm{CVD}(\mathrm{LA}_{(i)}^*), \bar{L}_{(i)}^*]$=R-SVD$(\mathrm{CVD}(\mathrm{LA}_{(i)}), \mathrm{LF}_{(i)}L_{(i)}, R_i)$.

Table 1. **The incremental Log-Euclidean Riemannian subspace learning algorithm (*IRSL*). R-SVD$(\cdot,\cdot,\cdot,R_i)$ represents that the first R_i dominant eigenvectors are used in R-SVD [14].**

in the experiments), and $\mathbf{vec}_{(i)} = \mathrm{UT}(\log(\mathrm{TC}_{(i)}))$ obtained from Eq. (7). The smaller the RE, the larger the likelihood.

4.4. Bayesian state inference for visual tracking

For visual tracking, a Markov model with a hidden state variable is generally used for motion estimation. In this model, the object motion between two consecutive frames is usually assumed to be an affine motion. Let X_t denote the state variable describing the affine motion parameters (the location) of an object at time t. Given a set of observed images $\mathcal{O}_t = \{O_1, \ldots, O_t\}$, the posterior probability is formulated by Bayes' theorem as:

$$p(X_t|\mathcal{O}_t) \propto p(O_t|X_t) \int p(X_t|X_{t-1})p(X_{t-1}|\mathcal{O}_{t-1})dX_{t-1} \quad (10)$$

where $p(O_t|X_t)$ denotes the observation model, and $p(X_t| X_{t-1})$ represents the dynamic model. $p(O_t \mid X_t)$ and $p(X_t \mid X_{t-1})$ decide the entire tracking process. A particle filter [3] is used for approximating the distribution over the location of the object using a set of weighted samples.

In the tracking framework, we apply an affine image warping to model the object motion of two consecutive frames. The six parameters of the affine transform are used to model $p(X_t \mid X_{t-1})$ of a tracked object. Let $X_t = (x_t, y_t, \eta_t, s_t, \beta_t, \phi_t)$ where $x_t, y_t, \eta_t, s_t, \beta_t, \phi_t$ denote the x, y translations, the rotation angle, the scale, the aspect ratio, and the skew direction at time t, respectively. We employ a Gaussian distribution to model the state transition distribution $p(X_t|X_{t-1})$. Also the six parameters of the affine transform are assumed to be independent. Consequently, $p(X_t|X_{t-1})$ is formulated as:

$$p(X_t|X_{t-1}) = \mathcal{N}(X_t; X_{t-1}, \Sigma) \quad (11)$$

where Σ denotes a diagonal covariance matrix with diagonal elements: $\sigma_x^2, \sigma_y^2, \sigma_\eta^2, \sigma_s^2, \sigma_\beta^2, \sigma_\phi^2$, respectively. The observation model $p(O_t \mid X_t)$ reflects the probability that a sample is generated from the subspace. In this paper, RE, defined in (9), is used to measure the distance from the sample to the center of the subspace. Consequently, $p(O_t|X_t)$ is formulated as:

$$p(O_t|X_t) \propto exp(-RE) \quad (12)$$

After maximum a posterior (MAP) estimation, we just use the Log-Euclidean covariance matrices of features inside the affinely warped image region associated with the highest weighted hypothesis to update the Log-Euclidean Riemannian eigensapace model.

5. Experiments

In order to evaluate the performance of the proposed tracking framework, six videos are used in the experiments. Videos 1 and 3 are taken from stationary cameras in different scenes while Videos 2, 4, 5, and 6 are recorded with moving cameras. The first four videos consist of 8-bit gray scale images while the last two are composed of 24-bit color images. Video 1 consists of dark and motion-blurring gray scale images, where many motion events take place, including wearing and taking off the glasses, head shaking, and hands occluding the face from time to time. In Video 2, a girl changes her facial pose over the time with varying lighting conditions. Besides, the girl's face is severely occluded by a man in the middle of the video stream. In Video 3, a pedestrian as a small object moves down a road in a dark and blurry scene. In Video 4, a man changes his pose and facial expression over the time with hands occluding the face from time to time. Moreover, each frame in Video 4 contains seven benchmark points, which characterize the location and the shape of his face. In Video 5, a hand moves in an indoor scene with a red notebook occluding the hand from time to time. In the last video, a girl's face is occluded partially by her hand from time to time. For the Log-Euclidean Riemannian eigenspace representation, the size of each object region is normalized to 20×20 pixels. The settings of the ranks R_i $(1 \leq i \leq 5)$ in *IRSL* are obtained from the experiments. The Log-Euclidean Riemannian subspace is updated every three frames. For the particle filtering in the visual tracking, the number of particles is set to be 200. The six diagonal elements $(\sigma_x^2, \sigma_y^2, \sigma_\eta^2, \sigma_s^2, \sigma_\beta^2, \sigma_\phi^2)$ of the covariance matrix Σ in (11) are assigned as $(5^2, 5^2, 0.03^2, 0.03^2, 0.005^2, 0.001^2)$, respectively.

Five experiments are conducted to demonstrate the claimed contributions of the proposed *IRSL*. In these five experiments, we compare tracking results of *IRSL* with those of a state-of-the-art Riemannian metric based tracking algorithm [24], referred to here as *CTMU*, in different scenarios including scene blurring, small object tracking, object pose variation, and occlusion. *CTMU* is a represen-

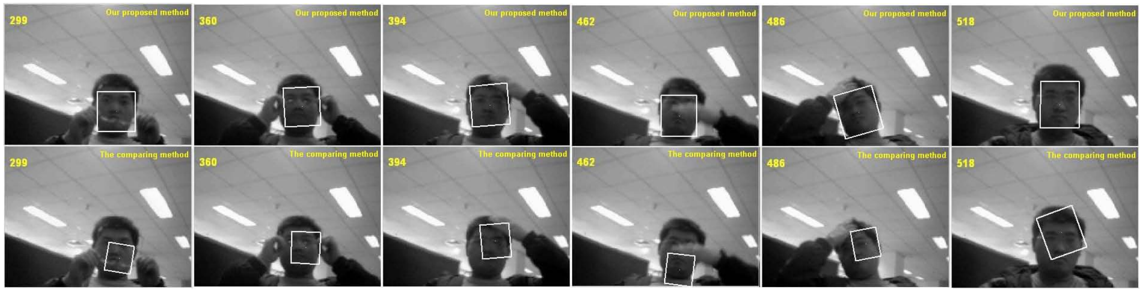

Figure 4. **The tracking results of *IRSL* (row 1) and *CTMU* (row 2) over representative frames under partial occlusions and scene blurring.**

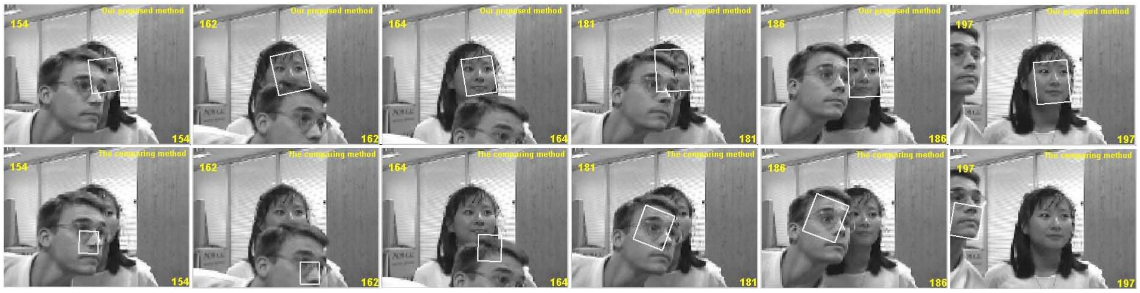

Figure 5. **Tracking results of *IRSL* (row 1) and *CTMU* (row 2) over representative frames in the scenario of a severe occlusion.**

tative Riemannian metric based tracking algorithm which uses the covariance matrix of features for object representation. By using a model updating mechanism, *CTMU* adapts to the undergoing object deformations and appearance changes, resulting in a robust tracking result. In contrast to *CTMU*, *IRSL* relies on Log-Euclidean Riemannian subspace learning to reflect the appearance changes of an object. Consequently, it is interesting and desirable to make a comparison between *IRSL* and *CTMU*. Furthermore, *CTMU* does not need additional parameter settings since *CTMU* computes the covariance matrix of image features as the object model. More details of *CTMU* are given in [24].

The first experiment is to compare the performances of the two methods *IRSL* and *CTMU* in handling partial occlusions and scene blurring using Video 1. In this experiment, R_i $(1 \leq i \leq 5)$ in *IRSL* is set as 5. Some samples of the final tracking results are demonstrated in Figure 4, where rows 1 and 2 are for *IRSL* and *CTMU*, respectively, in which six representative frames (299, 360, 394, 462, 486, and 518) of the video stream are shown. From Figure 4, we see that *IRSL* is capable of tracking the object all the time even though the object is occluded partially from time to time in a poor lighting condition. In comparison, *CTMU* is lost in tracking from time to time.

The second experiment is for a comparison between *IRSL* and *CTMU* on tracking a girl's face in the scenario of severe occlusions using Video 2. In this experiment, R_i $(1 \leq i \leq 5)$ in *IRSL* is set as 6. Some samples of the final tracking results are demonstrated in Figure 5, where rows 1 and 2 correspond to *IRSL* and *CTMU*, respectively, in which six representative frames (154, 162, 164, 181, 186, and 197) of the video stream are shown. From Figure 5, it is clear that *IRSL* is capable of tracking the object successfully in the case of severe occlusion while *CTMU* gets lost in tracking the object after severe occlusions.

The third experiment aims to compare the tracking performance of *IRSL* with that of *CTMU* in handling scene blurring and small object scenarios using Video 3. R_i $(1 \leq i \leq 5)$ in *IRSL* is set as 4. We show some samples of the final tracking results for *IRSL* and *CTMU* in Figure 6, where the first and the second rows correspond to the performances of *IRSL* and *CTMU*, respectively, in which six representative frames (495, 498, 505, 522, 547, and 550) of the video stream are shown. Clearly, *IRSL* succeeds in tracking the object while *CTMU* fails.

The fourth experiment is to make a quantitative comparison between *IRSL* and *CTMU* in the scenarios of partial occlusions and pose variations using Video 4. In this experiment, R_i $(1 \leq i \leq 5)$ in *IRSL* is set as 8. Some samples of the final tracking results are shown in Figure 7, where rows 1 and 2 correspond to *IRSL* and *CTMU*, respectively, in which six representative frames (49, 108, 117, 185, 289, and 292) of the video stream are shown. From Figure 7, it is clear that *IRSL* is capable of tracking the object successfully while *CTMU* is almost lost in tracking the object. During the tracking, seven validation points, corresponding to the seven benchmark points, are obtained according to the object's affine motion parameters at each frame. We use the location deviation (also called tracking error) between the validation points and the benchmark ones to quantitatively evaluate the tracking performance. The quantitative comparison results are displayed in Figure 8, from which we see that the tracking error of *IRSL* is always lower than that of *CTMU*.

The last experiment is to compare the tracking perfor-

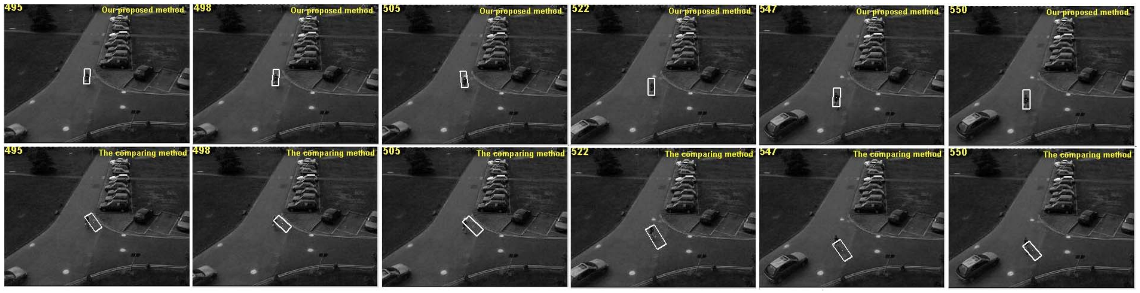

Figure 6. **The tracking results of *IRSL* (row 1) and *CTMU* (row 2) over representative frames in the scenarios of scene blurring and small object.**

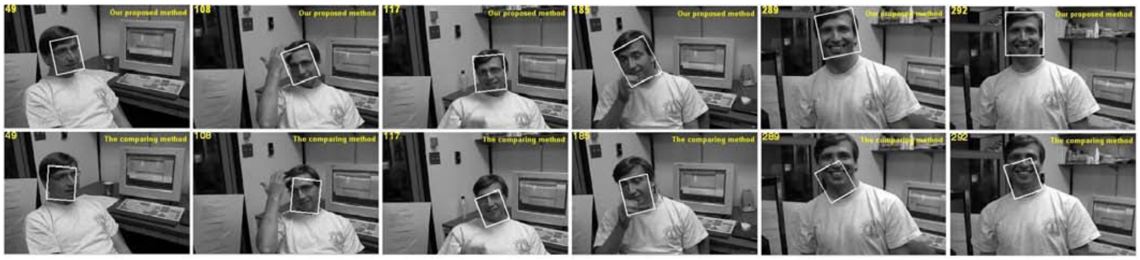

Figure 7. **The tracking results of *IRSL* (row 1) and *CTMU* (row 2) over representative frames in the scenarios of partial occlusions**

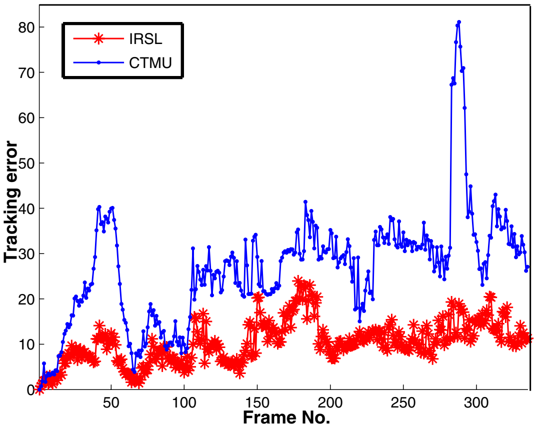

Figure 8. **The quantitative comparison results of *IRSL* and *CTMU* using video 4.**

mance of *IRSL* with that of *CTMU* in the color scenario with a partial occlusion using Videos 5 and 6. The RGB color space is used in this experiment. R_i $(1 \leq i \leq 5)$ for Videos 5 and 6 are set as 6 and 8, respectively. We show some samples of the final tracking results for *IRSL* and *CTMU* in Figure 9, where the first and the second rows correspond to the performances of *IRSL* and *CTMU* over Video 5, respectively, in which six representative frames (73, 79, 94, 125, 128, and 132) of the video stream are shown, while the third and the last rows correspond to the performances of *IRSL* and *CTMU* over Video 6, respectively, in which six representative frames (389, 390, 393, 396, 399, and 400) of the video stream are shown. Clearly, *IRSL* succeeds in tracking for both Video 5 and Video 6 while *CTMU* fails.

In summary, we observe that *IRSL* outperforms *CTMU* in the scenarios of blurring scenes, small objects, pose variations, and occlusions. *IRSL* makes a full use of the spatial correlation information of object appearance in the five modes. The dominant subspace information of the five modes is incorporated into *IRSL*. Even if the subspace information of some modes is partially lost or drastically varies, *IRSL* is capable of recovering the information using the cues of the subspace information from other modes. In comparison, *CTMU* only captures the statistical properties of object appearance in one mode, resulting in the loss of the local spatial correlation information inside the object region. In particular, *IRSL* constructs a robust Log-Euclidean Riemannian eigenspace representation of an object appearance. The representation fully explores the distribution information of covariance matrices of image features under the Log-Euclidean Riemannian metric, whereas *CTMU* relies heavily on an intrinsic mean in the Lie group structure without considering the distribution information of the covariance matrices of image features. Consequently, *IRSL* is an effective online subspace learning algorithm which performs well in modeling appearance changes of an object in many complex scenarios.

6. Conclusion

In this paper, we have developed a visual tracking framework based on the novel Log-Euclidean Riemannian metric. In this framework, the Log-Euclidean covariance matrices of image features in the five modes have been used to represent object appearance. Further, a novel online Log-Euclidean Riemannian subspace learning algorithm *IRSL*, which enables subspace analysis under a Log-Euclidean Riemannian metric, has been proposed to reflect the appearance changes of an object. Moreover, a novel criterion for the likelihood evaluation, based on the Log-Euclidean Riemannian subspace reconstruction error norms in the five modes, has been proposed to measure the similarity between the test image and the learned subspace model during the tracking. Compared with the state-of-art Riemannian metric-based tracking method *CTMU*, the proposed *IRSL* is

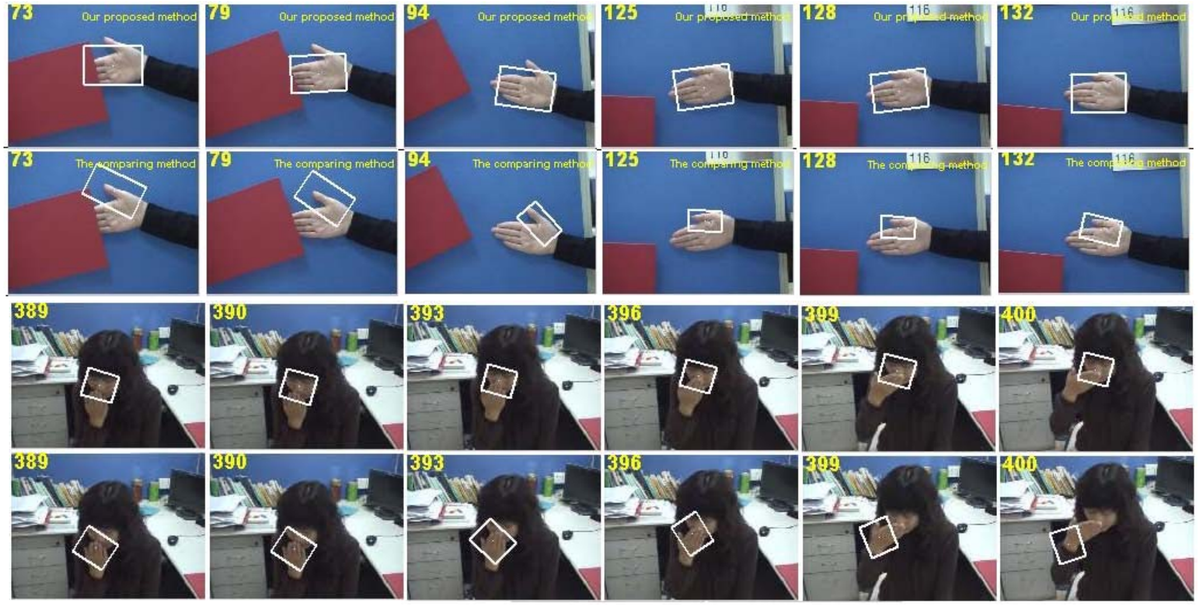

Figure 9. **The tracking results of *IRSL* and *CTMU* over representative frames in the color scenarios of partial occlusions. Rows 1 and 2 show the tracking results of *IRSL* and *CTMU* for Video 5, respectively. Rows 3 and 4 display the tracking results of *IRSL* and *CTMU* for Video 6, respectively.**

more robust to occlusion, scene blurring, small object, and object pose variation. Experimental results have demonstrated the robustness and promise of the proposed framework.

7. Acknowledgment

This work is partly supported by NSFC (Grant No. 60520120099, 60672040 and 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453). ZZ is partly supported by NSF (IIS-0535162), AFRL (FA8750-05-2-0284), and AFOSR (FA9550-06-1-0327).

References

[1] G. Hager and P. Belhumeur, “Real-time tracking of image regions with changes in geometry and illumination,” in *Proc. CVPR,* pp.430-410, 1996.

[2] M. J. Black and A. D. Jepson, “Eigentracking: Robust matching and tracking of articulated objects using view-based representation,” in *Proc. ECCV,* pp.329-342, 1996.

[3] M. Isard and A. Blake, “Contour tracking by stochastic propagation of conditional density,” in *Proc. ECCV,* Vol. 2, pp.343-356, 1996.

[4] M. J. Black, D. J. Fleet, and Y. Yacoob, “A framework for modeling appearance change in image sequence,” in *Proc. ICCV,* pp.660-667, 1998.

[5] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, “Robust Online Appearance Models for Visual Tracking,” in *Proc. CVPR,* Vol. 1, pp.415-422, 2001.

[6] S. K. Zhou, R. Chellappa, and B. Moghaddam, “Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters,” *IEEE Trans. on Image Processing,* Vol. 13, pp.1491-1506 , November 2004.

[7] T. Yu and Y. Wu, “Differential Tracking based on Spatial-Appearance Model(SAM),” in *Proc. CVPR,* Vol. 1, pp.720-727, June 2006.

[8] J. Li, S. K. Zhou, and R. Chellappa, “Appearance Modeling under Geometric Context,” in *Proc. ICCV,* Vol. 2, pp.1252-1259, 2005.

[9] K. Lee and D. Kriegman, “Online Learning of Probabilistic Appearance Manifolds for Video-based Recognition and Tracking,” in *Proc. CVPR,* Vol. 1, pp.852-859 , 2005.

[10] H. Lim, V. I. Morariu3, O. I. Camps, and M. Sznaier1, “Dynamic Appearance Modeling for Human Tracking,” in *Proc. CVPR,* Vol. 1, pp.751-757, 2006.

[11] J. Ho, K. Lee, M. Yang and D. Kriegman, “Visual Tracking Using Learned Linear Subspaces ,” in *Proc. CVPR,* Vol. 1, pp.782-789, 2004.

[12] Y. Li, L. Xu, J. Morphett and R. Jacobs, “On Incremental and Robust Subspace Learning,” *Pattern Recognition,* 37(7), pp. 1509-1518, 2004.

[13] D. Skocaj, A. Leonardis, “Weighted and Robust Incremental Method for Subspace Learning,” in *Proc. ICCV,* pp.1494-1501, 2003.

[14] J. Limy, D. Ross, R. Lin and M. Yang, “Incremental Learning for Visual Tracking,” *NIPS,* pp.793-800, MIT Press, 2005.

[15] X. Li, W. Hu, Z. Zhang, X. Zhang, and G. Luo, “Robust Visual Tracking Based on Incremental Tensor Subspace Learning,” in *Proc. ICCV,* 2007.

[16] A. Yilmaz, “Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection,” in *Proc. CVPR,* 2007.

[17] G. Silveira and E. Malis, “Real-time Visual Tracking under Arbitrary Illumination Changes,” in *Proc. CVPR,* 2007.

[18] M. Grabner, H. Grabner, and H. Bischof, “Learning Features for Tracking,” in *Proc. CVPR,* 2007.

[19] Q. A. Nguyen, A. Robles-Kelly, and C. Shen, “Kernel-based Tracking from a Probabilistic Viewpoint,” in *Proc. CVPR,* 2007.

[20] S. Ilić and P. Fua, “Non-Linear Beam Model for Tracking Large Deformations,” in *Proc. ICCV,* 2007.

[21] S. Tran and L. Davis, “Robust Object Tracking with Regional Affine Invariant Features,” in *Proc. ICCV,* 2007.

[22] Q. Zhao, S. Brennan, and H. Tao, “Differential EMD Tracking,” in *Proc. ICCV,* 2007.

[23] H. Wang, D. Suter, K. Schindler, and C. Shen, “Adaptive Object Tracking Based on an Effective Appearance Filter,” *IEEE Trans. on PAMI.,* Vol. 29, Iss. 9, pp.1661-1667, 2007.

[24] F. Porikli, O. Tuzel, and P. Meer, “Covariance Tracking using Model Update Based on Lie Algebra,” in *Proc. CVPR,* Vol. 1, pp. 728-735, 2006.

[25] O. Tuzel, F. Porikli, and P. Meer, “Human Detection via Classification on Riemannian Manifolds,” in *Proc. CVPR,* 2007.

[26] P. T. Fletcher and S. Joshi, “Principal Geodesic Analysis on Symmetric Spaces: Statistics of Diffusion Tensors,” in *Proc. Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis,* pp. 87-98, 2004.

[27] T. Lin and H. Zha, “Riemannian Manifold Learning,” *IEEE Trans. on PAMI.,* Iss. 99, 2007.

[28] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, “Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices,” *SIAM Journal on Matrix Analysis and Applications,* 2006.

[29] O. Tuzel, F. Porikli, and P. Meer, “Region Covariance: A Fast Descriptor for Detection and Classification,” in *Proc. ECCV,* 2006.

[30] X. Pennec, P. Fillard, and N. Ayache, “A Riemannian Framework for Tensor Computing,” *IJCV,* pp.41-66, 2006.

[31] W. Rossmann, “Lie Groups: An Introduction Through Linear Group,” Oxford Press, 2002.

[32] A. Levy and M. Lindenbaum, “Sequential Karhunen-Loeve Basis Extraction and Its Application to Images,” *IEEE Trans. on Image Processing,* Vol. 9, pp.1371-1374, 2000.

[33] X. Zhang, W. Hu, S. Maybank, and X. Li, “Graph Based Discriminative Learning for Robust and Efficient Object Tracking,” in *Proc. ICCV,* 2007.