Incremental tensor biased discriminant analysis: A new color-based visual tracking method

Jing Wen a, Xinbo Gao a,*, Yuan Yuan b, Dacheng Tao c, Jie Li a

a School of Electronic Engineering, Xidian University, No. 2, South Taibai Road, Xi'an 710071, China
b School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, United Kingdom
c School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 2 May 2009
Received in revised form
13 August 2009
Accepted 6 October 2009
Communicated by X. Li
Available online 20 November 2009

Keywords:
Visual tracking
Biased discriminant analysis
Tensor biased discriminant analysis
Incremental tensor biased discriminant analysis

ABSTRACT

Most existing color-based tracking algorithms utilize the statistical color information of the object as the tracking clues, without maintaining the spatial structure within a single chromatic image. Recently, the researches on the multilinear algebra provide the possibility to hold the spatial structural relationship in a representation of the image ensembles. In this paper, a third-order color tensor is constructed to represent the object to be tracked. Considering the influence of the environment changing on the tracking, the biased discriminant analysis (BDA) is extended to the tensor biased discriminant analysis (TBDA) for distinguishing the object from the background. At the same time, an incremental scheme for the TBDA is developed for the tensor biased discriminant subspace online learning, which can be used to adapt to the appearance variant of both the object and background. The experimental results show that the proposed method can track objects precisely undergoing large pose, scale and lighting changes, as well as partial occlusion.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Visual tracking is an important and essential component of visual perception, and has been an active research topic in computer vision community for decades. Influenced by the environment change and object motion, the appearance of the object takes on variety and variability, which is a challenge for describing the object effectively.

Many works have been developed for visual tracking to extract various low-level features (e.g., color [1,3,7], shape [2,6], texture and contour [4,5]), and build object appearance model (e.g., spatial histogram [1,2,7], AAMs [10], and subspace method [11–17]). Birchfield [1] and Wang et al. [2] presented the facial model with integration of shape and color. Hayashi and Fujiyoshi [3] developed color tracking method based on meanshift in luminance change. Isard and Blake [4] proposed a conditional density propagation of a parametric spine curve. Cootes et al. [6] employed the combination of shape with appearance representations for tracking. Perez et al. [7] presented multi-part color modeling to capture a rough spatial layout ignored by global histograms, without taking the background color into account. Comaniciu et al. [8] proposed a new object tracking based on kernel method and meanshift algorithm. Avidan [9] developed an ensemble tracking by using AdaBoost to distinguish the object from the background. These methods above usually utilize the low-level features of the image, but ignore the high-level semantic knowledge. Moreover, these methods often assume that the object takes on consistency and similarity in respect of the texture, gradient and so on, and obtain the segmentation-like tracking results. However, the corresponding locations/pixels within these segmentation-like tracked image patches usually have not coherent sense in the context.

For the purpose of getting more accurate results, a series of appearance-based methods were developed recently. Gross et al. [10] employed AAMs to track face efficiently in videos containing occlusion, but a complicated training process is inevitable before the tracking. The subspace-based methods were developed recently and applied to many research areas widely [11–35]. Based on the success of EigenTracking and incremental extension of the subspace learning, Ross et al. [11] presented an adaptive probabilistic visual tracking by updating subspace incrementally. Lin et al. [12] proposed a discriminative generative model for visual tracking. On the basis of Lin's work, Shen et al. [13] developed a kernelized version for tracking. Yuan et al. [32] employed the incremental principal components analysis to scene segmentation for visual surveillance. Due to the limitation of this
image-as-vector way to build subspace, a new representation based on high order tensor has drawn many researchers’ interest, and been introduced into the tracking and recognition. Tao et al. [18,33] introduced the multilinear representation for gait recognition. Moreover, Sun et al. [24] proposed incremental tensor analysis to deal with the learning of dynamic online data. Later, series of discriminant methods [21,27,28,31] for tensor analysis and their applications in retrieval [21,31], video semantic [30], gait recognition [28,33] were developed. Tao et al. [23,25] proposed a Bayesian tensor analysis method and applied it to 3-D face modeling, as well as the kernelization [22] and probabilistic [26] version. Li and Lee [14] presented a motion saliency-based visual tracking. Shao et al. [15] developed an appearance-based method using the three-dimensional trilinear tensor. Li et al. [16] employed a three-dimensional temporal tensor subspace learning for visual tracking. It should be noticed that those appearance-based tracking algorithms above have the following characteristics:

- They only make use of the intensity information [5,11–13, 15–17], which will lose chromatic knowledge in color video sequences.
- They usually regard the visual tracking as a two-class fisher discriminant problem [12,13], whereas the classification between the object and the background should be a (1+x)-class formulation.

In visual tracking, the object belongs to the positive sample set, while the background belongs to the negative sample set relative to the object of interest. Though the appearances of both the object and background vary with time, the variant between the object and background is so different. Since there are only pose, scale and illumination changes for the object, the changing appearances of the object are similar in some degree during the tracking, which can be regarded to a class as the blue symbols shown in Fig. 1. However, with the object moving the background changes drastically, which is shown in Fig. 1 with the orange symbols. Therefore, it is unreasonable to assign the background to one class. With this consideration, biased discriminant analysis (BDA) was developed by Zhou and Huang [19,20], which also leads to the small sample size (SSS) problem. Tao et al. [21] proposed a direct kernel biased discriminate analysis to deal with the SSS problem, and to provide a relevance feedback scheme for content-based image retrieval.

However, these discriminant analysis methods still hold the manner of the image-as-vector representation, and lose the spatial structure of the two-dimensional image. The SSS problem could be coped with by introducing the tensor representation [22]. In this paper, we propose a tensor biased discriminant analysis (TBDA) which could solve the SSS problem existed in the BDA, develop an online learning scheme for the TBDA named as incremental tensor biased discriminant analysis (ITBDA), and present a third-order color tensor-based visual tracking by employing the ITBDA to distinguish the objects from background. The contributions of this paper can be summarized as follows: (1) provide a new appearance construction method by integrating the spatial color information into a third-order tensor representation, which is more distinguishable for the appearance model; (2) propose a tensor biased discriminant analysis (TBDA), which is the generalization of the BDA for tensor representation, and able to deal with the SSS caused by the vectorization of the BDA; (3) present an incremental tensor biased discriminant analysis (ITBDA) suitable for online distinguishing the objects from the object-like background.

The reminder of this paper is organized as follows: the related previous work, say LDA and BDA, is described in Section 2. The TBDA and ITBDA are then proposed in Section 3. In Section 4, a color-based visual tracking algorithm is presented. Section 5 conducts several experiments to validate the effectiveness of the proposed tracking method undergoing large pose, scale and lighting changes for the single-object tracking, as well as the partial occlusion for the multiple-object tracking. The final section draws conclusions and future works.

2. Previous works

In this section, previous works are introduced including the linear discriminant analysis (LDA) and biased discriminant analysis (BDA).

2.1. Linear discriminant analysis (LDA)

Linear discriminant analysis is a supervised learning method, which tries to seek directions that are effective for discrimination. The objective function of LDA is to obtain a set of vectors W, which maximizing the ratio between $S_w$ and $S_b$, the within-class scatter matrix and the between-class scatter matrix,

$$W = \arg\max_w \frac{|W^T S_b W|}{|W^T S_w W|} \quad (1)$$

Suppose we have a set of $N$ samples $x_1, x_2, \ldots, x_N$, belonging to $k$ classes and each class has $N_i$ samples. Then $S_b$ and $S_w$ are defined as

$$S_b = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu)'$$
$$S_w = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N_i} (x_j - \mu_i)^T (x_j - \mu_i)'$$

where $N = \sum_{i=1}^{k} N_i$, $\mu = (1/N) \sum_{i=1}^{N} x_i$ is the mean vector of the total training set, and $\mu_i = (1/N_i) \sum_{j=1}^{N_i} x_j$ is the mean vector for the individual class $k_i$. $x_j$ is the $j$th samples belonging to class $k_i$. Therefore, $W$ can be computed from the eigenvectors of $S_w^{-1}S_b$.

Given two classes, i.e., $k=2$, LDA is reduced to Fisher discriminant analysis (FDA) as shown in Fig. 2; otherwise, multiple discriminant analysis (MDA).

![Fig. 1. Data stream for classification.](image)

![Fig. 2. FDA versus BDA, the left is FDA and the right is BDA.](image)
2.2. Biased discriminant analysis (BDA)

Zhou and Huang [19,20] proposed the biased discriminant analysis, which defines the \((1+\epsilon)\)-class classification problem. This method is applicable for the cases that there are an unknown number of classes (the symbols in orange color in Fig. 2) but the user only concern with one class (the symbols in blue color in Fig. 2), i.e., the user is biased toward one class (in blue).

The biased discriminant analysis tries to seek the subspace to discriminate the positive samples (the only class of concern to the user) from the negative samples (unknown number of classes). The objective aims to find a set of vector \(W\) maximizing the ratio between \(S_p\) and \(S_n\), the positive-scatter\-class matrix and the negative-class scatter matrix,

\[
W = \arg\max_W \frac{\|W^TS_pW\|}{\|W^TS_nW\|}
\]

where the scatter matrix can be obtained by

\[
\begin{align*}
S_p &= \sum_{i=1}^{N_p} (\mathbf{x}_i - \mathbf{\mu}_p)(\mathbf{x}_i - \mathbf{\mu}_p)^T \\
S_n &= \sum_{i=1}^{N_n} (\mathbf{y}_i - \mathbf{\mu}_n)(\mathbf{y}_i - \mathbf{\mu}_n)^T
\end{align*}
\]

where \(\mathbf{x}_i\) and \(\mathbf{y}_i\) denote the positive and negative samples, respectively. \(N_p\) and \(N_n\) are the numbers of the positive and negative samples, and \(\mathbf{\mu}_p = 1/N_p \sum_{i=1}^{N_p} \mathbf{x}_i\) is the mean vector of the positive samples. \(W\) can be computed from the eigenvectors of \(S_p^{-1}S_n\), BDA minimizes the variance of the positive samples and maximizes the scatter distance of all negative samples from the center of the positive samples. However, BDA still has the SSS problem, since the valid dimensionality is determined by the minimum of the positive and negative sample, while for the FDA, it is only one [20].

However, it is pointed out that when the number of the training measurements is limited, the vectorization operation always leads to the SSS problem [22]. For this reason, these vectorized discriminant methods often use the regularization method, which has been reported to be not a good choice by many papers on face recognition [34,35]. Therefore, we propose a new method based on tensor representation which could cope with the problem in some extent in the next section.

3. Incremental tensor biased discriminant analysis

In this section, a new supervised subspace method—tensor biased discriminant analysis (TBDA), the generalization of the biased discriminant analysis (BDA) [19] by introducing tensor representation, is developed to distinguish the positive and negative classes and mainly focus on the class of interest. Due to tensor representation makes full use of the structure information of the object, which is a reasonable constraint [22] to reduce the number of the unknown parameters used to represent a learning model, the proposed TBDA could cope with the SSS problem [22] in some extent by introducing tensor representation. In addition, an incremental tensor biased discriminant analysis (ITBDA) algorithm for distinguishing the positive from the negative class adaptively when the data are available incrementally/online. Moreover, the definition of the tensor algebra operation involved in this section could be referred to [18,22].

3.1. Tensor biased discriminant analysis

Before presenting the TBDA, we introduce the tensor LDA in advance. A tensor can be regarded as a multidimensional matrix. An \(M\)-order tensor \(\mathbf{X}\) is an \(M\)-dimensional matrix. Here, we denote a tensor sample as \(\mathbf{X} \in \mathbb{R}^{1 \times \cdots \times M}\). The traditional LDA aims to find a projection which is optimal for separating different classes in a low-dimensional space. Similarly, tensor LDA is to obtain the optimal discriminative projections for the tensor data in each order,

\[
W_d = \arg\max_W \frac{\|W^T S^{(d)}_d W\|}{\|W^T S^{(d)}_n W\|}, \quad d = 1, \ldots, M
\]

where

\[
\begin{align*}
S^{(d)}_p &= \frac{1}{N} \sum_{i=1}^{N} N_i (\mathbf{X}_i^{(d)} - \mathbf{X}_i^{(d)}\mathbf{X}_i^{(d)})^T \\
S^{(d)}_n &= \frac{1}{N} \sum_{i=1}^{N} N_i (\mathbf{X}_i^{(d)} - \mathbf{X}_i^{(d)}\mathbf{X}_i^{(d)})^T, \quad \mathbf{X}_i^{(d)} \in k_i
\end{align*}
\]

in which the symbol \(\mathbf{X}_i^{(d)}\) means the mode-\(d\) matricizing unfolding [18]. \(\mathbf{X}_i^{(d)}\) and \(\mathbf{X}_i^{(d)}\) are the mean of the total samples, the mean for the individual class \(k_i\), respectively. \(\mathbf{X}_i^{(d)}\) is the \(j\)th sample belonging to class \(k_i\). \(W_d\), which is the mode-\(d\) discriminant projection, can be computed from the eigenvectors of \(S^{(d)}_p^{-1}S^{(d)}_n\).

Similarly, for the case there are unknown numbers of classes while we only care for one class, tensor biased discriminant analysis (TBDA) is proposed for necessity. Accordingly, the positive and negative variance matrices can be defined as follows:

\[
\begin{align*}
S^{(d)}_X &= \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{X}_i^{(d)} - \mathbf{X}_i^{(d)}\mathbf{X}_i^{(d)}^T \\
S^{(d)}_Y &= \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbf{Y}_i^{(d)} - \mathbf{Y}_i^{(d)}\mathbf{Y}_i^{(d)}^T \\
S^{(d)}_N &= \frac{1}{N} \sum_{i=1}^{N} N_i (\mathbf{X}_i^{(d)} - \mathbf{X}_i^{(d)}\mathbf{X}_i^{(d)})^T
\end{align*}
\]

where \(N_p\) and \(N_n\) are the numbers of the positive and the negative samples, respectively. We can compute the biased discriminant projection in each mode by the following objective function:

\[
W_d = \arg\max_W \frac{\|W^T S^{(d)}_X W\|}{\|W^T S^{(d)}_Y W\|}, \quad d = 1, \ldots, M
\]

The eigenvectors are computed from \(S^{(d)}_X^{-1}S^{(d)}_Y\).

3.2. Incremental tensor biased discriminant analysis

We extend TBDA to an incremental learning fashion, named as incremental tensor biased discriminant analysis (ITBDA), so as to adapt to the situation that the data cannot be obtained at one time, especially in the context of visual tracking, the appearance model of the object should be updated to adapt to the change of both the object and the background with time drift. For visual tracking, the object measurement is the positive class which is concerned only, while the background belongs to the negative classes with unknown numbers of classes as illustrated in Fig. 1. Due to the appearances of both the object and background vary with time, the two variance matrices \(S^{(d)}_X\) and \(S^{(d)}_n\) should be updated as the arrival of new data (video frames).

Notice the difference of the computation of \(S^{(d)}_X\) from \(S^{(d)}_X\), \(S^{(d)}_n\) is the deviation of the negative samples from the mean of the positive class, while \(S^{(d)}_n\) is the variance matrix of the positive samples. We describe the online updating on \(S^{(d)}_X\) at first, which is similar to [17].

Suppose there are a set of old tensor data \(\{\mathbf{X}_i, i = 1, \ldots, n_k\}\), and the new arrival tensor data \(\{\mathbf{X}_j, j = 1, \ldots, m_k\}\), where \(n_k\) and \(m_k\) are
the old and new arrival data numbers. The total variance matrix for the mode-d tensor can be defined as follows:

\[
\mathbf{S}_X^{(d)} = \sum_{i=1}^{n_\text{old}} (\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})^\top
\]

where

\[
\bar{\mathbf{X}}^{(d)} = \frac{n_\text{old} \mathbf{X}_{\text{old}}^{(d)} + m_\text{new} \mathbf{X}_{\text{new}}^{(d)}}{n_\text{old} + m_\text{new}}
\]

and

\[
\begin{cases}
\mathbf{X}_{\text{old}}^{(d)} = \frac{1}{n_\text{old}} \sum_{i=1}^{n_\text{old}} \mathbf{X}_i^{(d)} \\
\mathbf{X}_{\text{new}}^{(d)} = \frac{1}{m_\text{new}} \sum_{i=1}^{m_\text{new}} \mathbf{X}_i^{(d)}
\end{cases}
\]

where \(\mathbf{X}^{(d)}\) is the total mean of the old and new tensor data, and \(\mathbf{X}_{\text{old}}^{(d)}\) and \(\mathbf{X}_{\text{new}}^{(d)}\) are the old and new mean, respectively. By defining the old and new variance matrices,

\[
\begin{cases}
\mathbf{S}_{X_{\text{old}}}^{(d)} = \sum_{i=1}^{n_\text{old}} (\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})^\top \\
\mathbf{S}_{X_{\text{new}}}^{(d)} = \sum_{i=1}^{m_\text{new}} (\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})^\top
\end{cases}
\]

we have Eq. (9) as

\[
\mathbf{S}_X^{(d)} = \sum_{i=1}^{n_\text{old}} (\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_i^{(d)} - \bar{\mathbf{X}}^{(d)})^\top + m_\text{new} (\mathbf{X}_{\text{new}}^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_{\text{new}}^{(d)} - \bar{\mathbf{X}}^{(d)})^\top
\]

subscribe Eq. (10) into Eq. (13), and we get

\[
\mathbf{S}_X^{(d)} = \mathbf{S}_{X_{\text{old}}}^{(d)} + \mathbf{S}_{X_{\text{new}}}^{(d)} + \frac{n_\text{old} m_\text{new}}{n_\text{old} + m_\text{new}} (\mathbf{X}_{\text{old}}^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_{\text{new}}^{(d)} - \bar{\mathbf{X}}^{(d)})^\top
\]

By analyzing Eq. (14), we can come out to the conclusion that when the new data are received, the updated variance matrix is not the simple sum of the variance matrix of the old data \(\mathbf{S}_{X_{\text{old}}}^{(d)}\) and the new arrival data \(\mathbf{S}_{X_{\text{new}}}^{(d)}\), there also exist a new term \((n_\text{old} m_\text{new} / (n_\text{old} + m_\text{new})) (\mathbf{X}_{\text{old}}^{(d)} - \bar{\mathbf{X}}^{(d)})(\mathbf{X}_{\text{new}}^{(d)} - \bar{\mathbf{X}}^{(d)})^\top\). It should be mentioned that \(\mathbf{S}_{X_{\text{old}}}^{(d)}\) is stored as \(\mathbf{U}_X^{(d)} \mathbf{X}_{\text{old}}^{(d)} \mathbf{U}_X^{(d)\top}\) \(\mathbf{S}_{X_{\text{old}}}^{(d)}\) for the sake of reducing the memory requirement and facilitating the computation of the likelihood in Section 5, where \(\mathbf{U}_X^{(d)}\) is an unitary matrix composed of basis vectors and \(\mathbf{D}_X^{(d)}\) is a diagonal matrix with nonnegative real numbers on the diagonal corresponding to the singular-value decomposition (SVD) of \(\mathbf{S}_{X_{\text{old}}}^{(d)}\).

Notice that the computation of the matrix \(\mathbf{S}_{Y_{\text{old}}}^{(d)}\) is different from that of \(\mathbf{S}_{X_{\text{old}}}^{(d)}\). We only need sum up the two matrix \(\mathbf{S}_{X_{\text{old}}}^{(d)}\) and \(\mathbf{S}_{X_{\text{new}}}^{(d)}\).

The reason for this is that the only interested class for the tracking task is the object, while the background observation also changes at the same time. We do not need to update the negative (background observation) mean, and should keep the biased center at the updated positive (object observation) mean only, with the following equation:

\[
\mathbf{S}_{Y_{\text{old}}}^{(d)} = \mathbf{S}_{X_{\text{old}}}^{(d)} + \mathbf{S}_{X_{\text{new}}}^{(d)}
\]
4. ITBDA-based visual tracking

Visual tracking could be formulated as a classification problem between the object and background observations, which are encoded with the locations or motion parameters of the objects through the unobservable states, and the task is to infer the unobservable states from the observed images over time.

4.1. Bayesian inference

We regard the visual tracking as a state estimation problem in a way similar to [17]. Denote \( z_t \) as an image region observed at time \( t \) and \( Z_t = (z_1, \ldots, z_t) \) is a set of image regions observed from the beginning to time \( t \). A visual tracking is a process to infer state \( g_t \) from the observation \( z_t \). Due to the motion of the object from one frame to the next can be modeled by a one-order Markov model, this inference problem is defined as a recursive equation:

\[
p(g_t | Z_t) = \kappa p(z_t | g_t) \int p(g_t | Z_{t-1}) p(g_{t-1} | Z_{t-1}) \, dg_{t-1}
\]

(17)

where \( \kappa \) is a constant, \( p(z_t | g_t) \) and \( p(g_t | Z_{t-1}) \) correspond to the observation model and dynamic model, respectively. \( p(g_{t-1} | Z_{t-1}) \) is the state estimation given all the prior observation up to time \( t-1 \), and \( p(z_t | g_t) \) is the likelihood of the observed image \( z_t \) at \( g_t \). Then the posterior estimation \( p(g_t | Z_t) \) at time \( t \) could be computed efficiently by Eq. (17). Therefore, the key steps for the visual tracking based on Bayesian inference are the computation of the dynamical model and observation model.

For visual tracking, an ideal distribution \( p(g_t | Z_t) \) should be convex at \( z_t \), i.e., \( g_t \) matching the observed object’s location \( z_t \). While the integral in Eq. (17) predicts the regions where objects is likely to appear given all the prior observation, the observation model \( p(z_t | g_t) \) could determine the most likely state that matches the observation at time \( t \).

**Dynamic model**: The location of an object in an image can be described by an affine image warp with six parameters. In this paper, the state at time \( t \) consists of the six parameters of an affine transformation \( g_t = [a_t, b_t, \theta_t, s_t, x_t, y_t] \), where \( a_t, b_t, \theta_t, s_t, x_t, y_t \), denote the vertical, horizontal transition, rotation angle, scale, aspect ratio, and skew direction at time \( t \).

\[
p(g_t | Z_{t-1}) = N(g_t; \mu_{g_t}, \Sigma_t)
\]

(18)

where \( \Psi = \text{a diagonal covariance matrix whose elements are the corresponding variances of affine parameters, i.e., } \sigma_{a_t}^2, \sigma_{b_t}^2, \sigma_{\theta_t}^2, \sigma_{s_t}^2, \sigma_{x_t}^2, \sigma_{y_t}^2. \) Each parameter in \( g_t \) is modeled independently by a Gaussian distribution around its counterpart in \( g_{t-1} \). And the warp of the observations patches between the successive frames is an affine transformation. In order to get the most likely state, a sampling scheme is adopted to collect a set of states \( \{g_{t-1}^i, i=1, \ldots, k\} \) as many as possible. Then the optimal state can be estimated by maximizing the likelihood \( p(z_t | g_t) \) among these states.

**Observation model**: Given a set of samples \( I = \{I^1, \ldots, I^p\} \) where \( I \) is the appearance tensor data (as shown in Fig. 3) collected in \( z_{t-1} \) based on \( \{g_{t-1}^i, i=1, \ldots, k\} \). That is, we draw a set samples parameterized by \( \{g_{t-1}^i, i=1, \ldots, k\} \) in \( z_{t-1} \) that have large \( p(z_t | g_t) \), but the low posterior \( p(g_t | Z_{t-1}) \). \( p(z_t | g_t) \) measures the probability of \( z_t^k \) as samples being generated by the object class.

\[
p(c_t^k | g_t) \propto \exp\left(-\|d_t^k - X_d^d\|^2 \right) \prod_{d=1}^M \times d \left(U^d X^d_t \right)^2, \quad i = 1, \ldots, k
\]

(19)

where \( \| \cdot \|^2 \) is Frobenius norm, and \( \times d \) is the mode product in tensor algebra [18].

**Feature selection for tracking**: We keep the best estimated sample at every frame parameterized by the optimal state as the positive sample. For the negative class, the samples with medium likelihood usually confuse the estimation, due to their appearances are close to the object class and containing both the observation of the object and background. Therefore, we will keep both the background samples and object-like samples, which have low values in Eq. (20) as negative samples.

4.2. Proposed tracking algorithm

In this paper, we adopt the RGB space as the appearance value, since color images can be thought as real valued third-order tensors as shown in Fig. 3. Both the object and background observations are the warped images, obtained by using affine transform with the six state parameters. The warped images are kept with their original chromatic structures, as third-order tensors with the first two orders as the location and the third order as the RGB color components as shown in the right of Fig. 3. The proposed tracking algorithm is based on maximum likelihood estimate given all the observations up to that time instance.

\[
g_t^g = \arg \max_{g_t} \left[p(Z_t | g_t) \right]
\]

(21)

The proposed tracking method can be initialized by manually labeling the object of interest or automatically detecting the object to be tracked through the available detection method. In this paper, we initialized the tracking by manually labeling the object in the first frame. The proposed tracking method is a propagation process for the conditional density of the object observation based on the affine state parameters. Firstly, according to the dynamical model in Eq. (18), the tracker draws \( k \) samples which represent the possible locations of the object at time \( t-1 \) based on \( \{g_{t-1}^i, i=1, \ldots, k\} \), and obtains \( k \) “particle” observations by warping the color image observations \( I = \{I^1, \ldots, I^p\} \) through the affine transformation based on these state parameters. Then, discriminate the object and background samples by the ITBDA among all these \( k \) image observations, and
compute the likelihood in Eq. (20) for the positive samples as the weights for propagation of the conditional density by the observation model. Thirdly, the optimal object sample and the background samples are assigned to positive and negative samples set, respectively, for updating the discriminative projection by ITBDA. The detailed tracking flow is given in Fig. 4.

5. Experimental results

In order to evaluate the performance of the proposed tracking algorithm, we collected eight videos with human face and pedestrians as the tracking objects, where the first three video sequences are captured indoor undergoing large pose variant and drastic illumination, the last five video sequences are recorded in shopping center in Portugal.

We test five and three video sequences for the single- and multiple-objects tracking, respectively, to verify the tracking performance of the proposed method. For the single object tracking in our experimental results, the red box shows the maximum likelihood of the observation. Moreover, the mean and original tracked images are shown in the bottom row of each panel, as well as the error image and reconstructive image based on the positive subspace.

Fig. 4. The proposed tracking flowchart.

Fig. 5. The results for tracking a human face mainly undergoing large pose variant with the frame numbers of the results #7, #23, #45, #66, #78, #100, #115, #130, #159, #181, #216, #230, #252, #299, #376, #381, #420, #457, #473, and #494, in order.

5.1. Tracking a human face

In this subsection, we desire two experiments to evaluate the tracking performance mainly undergoing pose and drastic light variant, in Examples 1 and 2, respectively.

**Example 1.** In this video sequence, the object, i.e., the human face, walks fast in the moving camera, undergoing large pose (#7, #45, #78, #130, #230, and #299), scale (#7, #78, #181, #252, #376, and #494) variant as shown in Fig. 5. Our method is able to efficiently learn a compact color representation while the color histogram method usually gives an inaccurate location for the object. All the eigenbases are constructed automatically from the RGB-space images and constantly updated to model the appearance of the object, as well as the background, both of which will be combined to compute the discriminant projection for the distinguishment at next frame.

**Example 2.** In this experiment, we aim to test the performance of the proposed tracking algorithm under different light conditions. In Figs. 6 and 7, the human faces are both under the environment of illumination change, as well as slight pose changing. In Fig. 6, the human face comes close to the camera, with a slow change of the light condition from the bright to the dark (#9, #168, and #384), with slight pose (#9, #48, and #248), expression (#9, #48, and #326) and scale variant (#9 and #287) during the light darkening. In Fig. 7, the human face undergoes a drastic illumination, walking from a very dark environment to a bright place (#6, #129, and #213), with slight pose (#27, #50, #109, and #170) during the light darkening. Our method makes full use of the color

---

Fig. 6. A person undergoing drastic light change (from the light to the dark), slight pose and expression, as well as the scale variant. The frame numbers of the results are #9, #48, #168, #205, #248, #287, #326, and #384, in order.

Fig. 7. A person undergoing drastic light change (from the dark to the light), and slight pose variant. The frame numbers of the results are #6, #27, #50, #70, #109, #129, #149, #170, #196, and #213, in order.

Fig. 8. A pedestrian moves far away from the camera with another one, with the other pedestrian occlusion. The frame numbers of the results are #50, #89, #106, #116, #126, #190, #202, #269, #325, and #361, in order.
information, which is combined with spatial structure in the third-order tensor data, could get good results under the various light conditions.

5.2. Tracking a pedestrian

In this subsection, the object is a pedestrian as shown in Figs. 8 and 9. The appearances of the pedestrians do not keep constant variance in spatial structure as the human face any more. Though the changing of the pedestrian observation seems periodic, the appearance among a period is so different from each other and usually accompanies with the out-of-plane rotation, which results in the nonperiodicity of the appearance changing.

The object in Fig. 8 experiences twice short-time occlusions in about #106, #116 frame and #190 frame, the object in red clothes is different in color from the two different occlusions, could be tracked well after the occlusions removed.

The object in Fig. 9 has been occluded two times by the same occlusion pedestrian in about frame #106 and #378. Though the colors for the two pedestrians are very similar, the spatial constraint for the color information could keep the tracker performs well, while the methods based on histogram or statistical model often lose the track by following the occlusion pedestrian away.

It should be pointed out that, in the bottom row of each panel in Figs. 5–9, the tracked image regions have the same structure for the appearance, even for the pedestrian tracking. This character of the proposed tracking method could provide reasonable tracked blob to the recognition task.

5.3. Qualitative comparison

In this subsection, two tracking methods are run, as a qualitative benchmark, on three video sequences: In [2], a spatial color mixture of Gaussians (SMOG) appearance model for particle filters were proposed, by considering both the colors in a region

Fig. 9. A pedestrian moves far away from the camera, with another pedestrian occlusion. The frame numbers of the results are #171, #185, #203, #221, #233, #281, #378, #448, #516, and #558, in order.

Fig. 10. A comparison of our tracking method (the bottom row) with the SMOG tracker [2] (the top row), the ITSL tracker (the second row) and the ICTSL tracker (the third row) on the video sequence of a person moving fast in the camera, with pose variant, cast shadow and scale changes.
and the spatial layout of the colors. However, the spatial information in [2] used the color histogram combined spatial information partly, which is irreversible feature extraction, that is, the original image structure cannot be reconstructed by the features. In order to maintain the true object appearance, incremental tensor subspace learning (ITSL) method [16] is also implemented in the experiments, which utilized only gray information, by integrating the temporal information to a third-order tensor incrementally. Then, our proposed method without discriminant analysis is also run for comparative objective by using the likelihood measurement in Eq. (19), named as incremental chromatic tensor subspace learning (ICTSL) here.

As shown in Figs. 10–12, our method provides comparable performance to the three methods above. The SMOG tracker (listed in the top row) is much more robust to the case that the appearance variant is slow as in Fig. 12 at about frame #151 and #181. However, it is not steady in Fig. 10 (#41, #151, and #271), Fig. 11 (#181 and #213) and Fig. 12 (#211 and #241), and loses the tracking object at the earlier frame number than the other three methods, due to the fast motion, view and scale changes, cast shadow and drastic illumination.

The ITSL tracker (the second row) gets better tracking performance than the SMOG tracker when there are the small pose changes in Fig. 11. However, it is sensitive to the appearance variant resulting from the fast motion and occlusion as shown in Figs. 10 and 12, and loses the tracking object due to the drastic light condition in Fig. 11.

The tracking performance of our proposed method without discriminant analysis (the third row) is much better than the above two methods, especially in Fig. 12, because it does not only hold the structure information of the appearance in the tensor representation, but also combine the color information with the tensor. However, without the discriminant analysis between the object and the background, the tracking results are not satisfactory in Fig. 10 (#369 and #384) and Fig. 11 (#433 and #451), due to the non-rigid motion and large light variant.

The proposed tracking method based on the ITBDA (the bottom row) could track the object effectively undergoing the various appearance changes caused by both the instinct and extrinsic factors such as pose (Fig. 10), light (Fig. 11) and scale variant (Figs. 10 and 12). The reason for that is our method could not only keep the color information in the structure representation, but also distinguish the object from the background by using the color structural information. Moreover, the particles number can be reduced in some degree, when the biased discriminant analysis is used to “filter” the particles. Thus, it is the true object observation, but the arbitrary image patches (maybe containing the background patches), whose conditional densities are propagated frame by frame.

5.4. Quantitative comparison

To evaluate the tracking precision quantitatively, we conducted a face tracking experiment on in the first and second video sequences to compare our tracking error results with the two available methods, i.e., SMOG and ITSL. We computed the location error between the estimated facial locations and the manually labeled “ground truth”.

The tracking errors are plotted for each frame in Figs. 13(a) and 14(a) for the first and second videos, respectively. For most frames our tracking results match the ground truth well in the first video as shown in Fig. 10, the largest errors occurring during large scale and fast pose changes. In Fig. 13, the SMOG tracker performs poorly, experiencing significant drift off the target object. This can be attributed to the appearance model of the SMOG tracker, based on the spatial histograms of the pixel intensities, which is not stable to the variance of the scale and pose, and does not adapt over time. The rest three tracking methods could get similar performance before about #330. The large errors occur during large scale and fast pose changes simultaneously. After about frame #330, the ITSL tracker loses the

---

**Fig. 11.** A comparison of our tracking method (the bottom row) with the SMOG tracker [2] (the top row), the ITSL tracker (the second row) and the ICTSL tracker (the third row) on the video sequence of a person moving close to the camera, from the bright place to the dark one.
object, and could not recover to the right position, because of the less of the color and discriminative information. After about frame #369, the ICTSL tracker loses the track and drift off. However, the proposed method could keep steady tracking performance.

Due to the scale and motion change gently in the second video (Fig. 11) compared to the first video (Fig. 10), the tracking errors of the four methods are also smaller in Fig. 14(a) than those in Fig. 13(a). The same conclusion could be drawn by comparing the location errors in Figs. 13(b) and 14(b). The location error in Fig. 10 is more than that in Fig. 11. Moreover, since the location and scale variant is small for the example in Fig. 11, the mean (symbol ‘*’)) and derivation of the location error in Fig. 14(b) would be smaller than that in Fig. 13(b), compared to the example in Fig. 10, which is reasonable theoretically. However, in the second video sequence, the object appearance is characterized by the drastic illumination, which would result in large variant of the pixel intensity value and lost track as seen in Fig. 14(a). The SMOG tracker still takes on instable tracking result respect to the ground truth. The tracking error of the ITSL tracker could get steady performance compared to the SMOG tracker, but the error is still a little larger than the rest two trackers, due to only use the illumination intensity and lost the chromatic spatial information. The ICTSL could get more accurate tracking performance compared with the two methods mentioned before, but the tracker still lost due to it does not take the background into account. The

![Figure 12](image)

**Fig. 12.** A comparison of our tracking method (the bottom row) with the SMOG tracker [2] (the top row), the ITSL tracker (the second row) and the ICTSL tracker (the third row) on the video sequence of a person moving close to the camera, from the bright to the dark place.

![Figure 13](image)

**Fig. 13.** The comparison for the tracking errors by the four methods shown in Fig. 10: (a) the tracking error between the estimated locations and the manually labeled location and (b) the comparison of the location precision and derivation by the four methods.
The proposed tracking method is able to get rather satisfied results as shown in Fig. 14(a).

In addition, our simulation environment is Matlab 7.7 on a PC (2.0 GHz Intel Core Duo, 1 GB RAM). The computational time of the tracking algorithm is about 0.245 fps, with 300 particle numbers, for the single-object tracking. It will cost as the objects number times as the single object, due to the objects are independent to each other for the multiple-objects tracking.

5.5. Tracking multiple objects

All the experimental results above are the single-object tracking examples. Here, we use the rest three video sequences to verify the multiple objects tracking ability of the proposed method in Figs. 15–17, respectively.

In Fig. 15, there are two objects (pedestrians) of interest, labeled by white and cyan rectangle, respectively. The two
object walk from the bottom right scene into the shop, experiencing the scale and shape change, as well as the partial occlusion before they go into the shop. The object labeled by cyan rectangles have similar color. It is the similar color that results in the tracking failure of the proposed method. So, the proposed tracking method would lose the object of interest when there is the full-body occlusion, especially the occluding object has the same or similar color as the occluded object.

6. Conclusion and future work

This paper proposes a new color-clue-based visual tracking method, which can incrementally learn the object color structure information and discriminate the object from the background in online way. For this application purpose, we extend the biased discriminant analysis from the vector-based method to tensor-based way in order to keep the structure information well combine with the color information, present batch tensor biased discriminant analysis (TBDA) and its incremental version, ITBDA, for distinguishing the object and background. Moreover, the TBDA could be able to deal with the SSS problems, as well as give a good talk on the extension of vector-vise method to the high order one. The experimental results show the robustness of the proposed algorithm to tracking object undergoes large pose variant, illumination change and occlusions.

Although our method could track object well for most cases, we have to point out the proposed method is a little sensitive to the very drastic illumination, and not adaptive to the full-body occlusion whose color is same as the object’s. Moreover, the proposed tracking algorithm cannot be implemented in real time, which will be one of the future works.

Acknowledgements

We want to thank the helpful comments and suggestions from the anonymous reviewers. This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China.

References


Jing Wen received the B.Sc. degree in Electronic Information Science and Technology from Shanxi University, Taiyuan, China, in 2003, and the M.Eng. degree in Signal and Information Processing from Xidian University, Xi’an, China, in 2006. Since August 2006, she has been pursuing her Ph.D. degree in Pattern Recognition and Intelligent System at Xidian University. Her research interests include pattern recognition and computer vision.

Xinbo Gao received the B.Sc., M.Sc. and Ph.D. degrees in signal and information processing from Xidian University, China, in 1994, 1997 and 1999, respectively. From 1997 to 1998, he was a research fellow in the Department of Computer Science at Shizuoka University, Japan. From 2000 to 2001, he was a postdoctoral research fellow in the Department of Information Engineering at the Chinese University of Hong Kong. Since 2001, he has been with the School of Electronic Engineering at Xidian University. Currently, he is a Professor of Pattern Recognition and Intelligent System, and Director of the VIPS Lab, Xidian University. His research interests are computational intelligence, machine learning, computer vision, pattern recognition and artificial intelligence. In these areas, he has published 4 books and over 100 technical articles in refereed journals and proceedings including IEEE TIP, TCSVT, TNN, TSMC, etc. He is on the editorial boards of journals including EURASIP Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as general chair/co-chair or program committee chair/co-chair or PC member for around 30 major international conferences.

Jie Li received the B.Sc., M.Sc. and Ph.D. degrees in Circuit and System from Xidian University, China, in 1995, 1998 and 2005, respectively. Since 1998, she joined the School of Electronic Engineering at Xidian University. Currently, she is an Associate Professor of Xidian University. Her research interests include computational intelligence, machine learning, and image processing. In these areas, she has published over 30 technical articles in refereed journals and proceedings including IEEE TCSVT, IJPS, etc.