Robust Foreground Segmentation Based on Two Effective Background Models

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

ABSTRACT

Foreground segmentation is a common foundation for many computer vision applications such as tracking and behavior analysis. Most existing algorithms for foreground segmentation learn pixel-based statistical models, which are sensitive to dynamic scenes such as illumination change, shadow movement, and swaying trees. In order to address this problem, we propose two block-based background models using the recently developed incremental rank-(R_1, R_2, R_3) tensor-based subspace learning algorithm (referred to as *IRTSA*) [1]. These two *IRTSA*-based background models (i.e., *IRTSA-GBM* and *IRTSA-CBM* respectively for grayscale and color images) incrementally learn low-order tensor-based eigenspace representations to fully capture the intrinsic spatio-temporal characteristics of a scene, leading to robust foreground segmentation results. Theoretic analysis and experimental evaluations demonstrate the promise and effectiveness of the proposed background models.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Video surveillance, object detection

1. INTRODUCTION

Foreground segmentation is a fundamental task for many computer vision applications. Higher level operations (e.g., visual surveillance and behavior analysis) rely heavily on the information provided by foreground segmentation. In general, segmentation of foreground regions in image sequences can be accomplished by matching the learned background model with each video frame. However, it is difficult for most existing background models to detect foreground objects in dynamic scenes such as illumination change, shadow movement, and swaying trees. Consequently, effectively modeling scenes is crucial for foreground segmentation.

In recent years, much work has been done in foreground segmentation. Stauffer and Grimson [2] propose an online adaptive background model where a mixture of Gaussians is adopted to model each pixel. The model classifies each pixel by matching the pixel with the Gaussian distribution representing the pixel most effectively. Furthermore, the number of Gaussians is adjusted adaptively to best represent background processes. Sheikh and Shah [5] present an improved nonparametric model combining both temporal and spatial information. In [6], an adaptive background model for grayscale video sequences is presented. The model utilizes local spatio-temporal statistics to detect shadows and highlights. Furthermore, it can adapt to illumination changes. Haritaoglu *et al.* [3] build a statistical background model representing each pixel by three values which are its minimum intensity value, its maximum intensity value and the maximum intensity difference between consecutive frames during training. In [7], Wang *et al.* present a probabilistic method for background subtraction and shadow removal. Their method detects shadows by a combined intensity and edge measure. Tian *et al.* [9] propose an adaptive Gaussian mixture model based on a local normalized cross-correlation metric and a texture similarity metric. These two metrics are used for detecting shadows and illumination changes, respectively. Patwardhan *et al.* [22] propose a framework for coarse scene modeling and foreground detection using pixel layers. The framework allows for integrated analysis and detection in a video scene. Wang *et al.* [8] present a dynamic conditional random field model for foreground and shadow segmentation. The model utilizes a dynamic probabilistic framework based on the conditional random field (CRF) to capture spatial and temporal statistics of pixels. In [4], PCA (principal component analysis) is performed on a collection of N images to construct a background model, which is represented by the mean image and the projection matrix comprising the first p significant eigenvectors of PCA. In this way, foreground segmentation is accomplished by computing the difference between the input image and its reconstruction; then online PCA is enabled to incrementally learn the background's eigenspace representation. However, the aforementioned methods for background modeling share a problem that they are unable to fully exploit the spatio-temporal redundancies within the image ensembles. This is particularly true for those image-as-vector techniques (e.g., [4]), as the local spatial information is almost lost. Consequently, the focus has been made on developing the high-order tensor learning algorithms for effective subspace analysis. In this case, the problem of modeling the appearance of a scene is reduced to how to make tensor decomposition more accurate and efficient.

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10 ...$5.00.

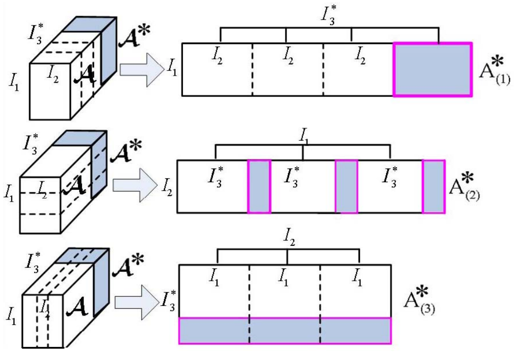

Figure 1: Illustration of the incremental rank-(R_1, R_2, R_3) tensor-based subspace learning of a 3-order tensor.

More recent work on modeling the appearance of an object focuses on using high-order tensors to construct a better representation of the object's appearance. Wang and Ahuja [10] propose a novel rank-R tensor approximation approach, which is designed to capture the spatio-temporal redundancies of tensors. In [11], an algorithm named Discriminant Analysis with Tensor Representation (DATER) is proposed. DATER is tensorized from the popular vector-based LDA algorithm. In [12, 13], the N-mode SVD, multilinear subspace analysis, is applied to constructing a compact representation of facial image ensembles factorized by different faces, expressions, viewpoints, and illuminations. Tao *et al.* [14] propose a supervised tensor learning (STL) framework to generalize convex optimization based schemes. The framework accepts nth-order tensors as inputs. He *et al.* [15] present a learning algorithm called Tensor Subspace Analysis (TSA), which learns a lower dimensional tensor subspace to characterize the intrinsic local geometric structure of the tensor space. In [16], Wang *et al.* give a convergent solution for general tensor-based subspace learning. Sun *et al.* [17] mine higher-order data streams using dynamic and streaming tensor analysis. Also in [18], Sun *et.al* present a window-based tensor analysis method for representing data streams over the time. All of these tensor-based algorithms have the same problem that they are not allowed for incremental subspace analysis for adaptively updating the sample mean and the eigenbasis.

In this paper, we propose a framework for foreground segmentation. In the framework, two background models (i.e., *IRTSA-GBM* and *IRTSA-CBM*) for grayscale and color images are developed to capture the spatio-temporal characteristics of a scene, leading to robust foreground segmentation results. These two background models are based on the recently developed incremental rank-(R_1, R_2, R_3) tensor-based subspace learning algorithm (referred to as *IRTSA*) [1]. The algorithm online constructs a low-order tensor eigenspace model, in which the sample mean and the eigenbasis are updated adaptively.

The remainder of the paper is organized as follows. An introduction to *IRTSA* [1] is given in Sec. 2. The framework for foreground segmentation is described in Sec. 3. Experimental results are reported in Sec. 4. The paper is concluded in Sec. 5.

Input:
CVD($\mathrm{A}_{(k)}$) of the mode-k unfolding matrix $\mathrm{A}_{(k)}$, i.e. $\mathrm{U}^{(k)}\mathrm{D}^{(k)}\mathrm{V}^{(k)^T}$ $(1 \leq k \leq 3)$ of an original tensor $\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$, newly-added tensor $\mathcal{F} \in \mathcal{R}^{I_1 \times I_2 \times I_3^{'}}$, column mean $\bar{L}^{(1)}$ of $\mathrm{A}_{(1)}$, column mean $\bar{L}^{(2)}$ of $\mathrm{A}_{(2)}$, row mean $\bar{L}^{(3)}$ of $\mathrm{A}_{(3)}$ and R_1, R_2, R_3.
Output:
CVD($\mathrm{A}^*_{(i)}$) of the mode-i unfolding matrix $\mathrm{A}^*_{(i)}$, i.e. $\hat{\mathrm{U}}^{(i)}\hat{\mathrm{D}}^{(i)}\hat{\mathrm{V}}^{(i)^T}(1 \leq i \leq 3)$ of $\mathcal{A}^* = (\mathcal{A} \mid \mathcal{F}) \in \mathcal{R}^{I_1 \times I_2 \times I_3^*}$ where $I_3^* = I_3 + I_3^{'}$, column mean $\bar{L}^{(1)^*}$ of $\mathrm{A}^*_{(1)}$, column mean $\bar{L}^{(2)^*}$ of $\mathrm{A}^*_{(2)}$ and row mean $\bar{L}^{(3)^*}$ of $\mathrm{A}^*_{(3)}$.
Algorithm:

```
1. $\mathrm{A}^*_{(1)} = \big(\mathrm{A}_{(1)} | F_{(1)}\big)$;
2. $\mathrm{A}^*_{(2)} = \big(\mathrm{A}_{(2)} | F_{(2)}\big) \cdot \mathrm{P} = \mathrm{B} \cdot \mathrm{P}$, where P is defined in (1);
3. $\mathrm{A}^*_{(3)} = \left(\frac{\mathrm{A}_{(3)}}{F_{(3)}}\right)$;
4. $[\hat{\mathrm{U}}^{(1)}, \hat{\mathrm{D}}^{(1)}, \hat{\mathrm{V}}^{(1)}, \bar{L}^{(1)^*}]$=R-SVD$(\mathrm{A}^*_{(1)}, \bar{L}^{(1)}, R_1)$;
5. $[\hat{\mathrm{U}}^{(2)}, \hat{\mathrm{D}}^{(2)}, \widetilde{\mathrm{V}}_2, \bar{L}^{(2)^*}]$=R-SVD$(\mathrm{B}, \bar{L}^{(2)}, R_2)$;
6. $\hat{\mathrm{V}}^{(2)} = \mathrm{P}^T \cdot \widetilde{\mathrm{V}}_2$;
7. $[\widetilde{\mathrm{U}}_3, \widetilde{\mathrm{D}}_3, \widetilde{\mathrm{V}}_3, \widetilde{L}_3]$=R-SVD$((\mathrm{A}^*_{(3)})^T, (\bar{L}^{(3)})^T, R_3)$;
8. $\hat{\mathrm{U}}^{(3)} = \widetilde{\mathrm{V}}_3$, $\hat{\mathrm{D}}^{(3)} = (\widetilde{\mathrm{D}}_3)^T$, $\hat{\mathrm{V}}^{(3)} = \widetilde{\mathrm{U}}_3$, $\bar{L}^{(3)^*} = (\widetilde{L}_3)^T$.
```

Figure 2: The incremental rank-(R_1, R_2, R_3) tensor-based subspace analysis algorithm (*IRTSA*). R-SVD$((\mathbb{C} \mid \mathbb{E}), L, R)$ represents that the first R dominant eigenvectors are used in R-SVD for the matrix $(\mathbb{C}|\mathbb{E})$ with $\mathbb{C}$'s column mean being L.

2. INCREMENTAL RANK-(R_1, R_2, R_3) TENSOR-BASED SUBSPACE LEARNING (*IRTSA*)

Based on R-SVD [19, 20], *IRTSA* [1] identifies the dominant projection subspaces of 3-order tensors, and is capable of incrementally updating these subspaces when new data arrive. Given the CVD($\mathrm{A}_{(k)}$) of the mode-k unfolding matrix $\mathrm{A}_{(k)}(1 \leq k \leq 3)$ for a 3-order tensor $\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$, *IRTSA* is able to efficiently compute the CVD($\mathrm{A}^*_{(i)}$) $= \hat{\mathrm{U}}^{(i)}\hat{\mathrm{D}}^{(i)}\hat{\mathrm{V}}^{(i)^T}$ of the mode-i unfolding matrix $\mathrm{A}^*_{(i)}(1 \leq i \leq 3)$ for $\mathcal{A}^* = (\mathcal{A} \mid \mathcal{F}) \in \mathcal{R}^{I_1 \times I_2 \times I_3^*}$ where $\mathcal{F} \in \mathcal{R}^{I_1 \times I_2 \times I_3^{'}}$ is a new 3-order subtensor and $I_3^* = I_3 + I_3^{'}$. To facilitate the description, Fig. 1 is used for illustration. In the left half of Fig. 1, three identical tensors are unfolded in three different modes. For each tensor, the white regions represent the original subtensor while the dark regions denote the newly added subtensor. The three unfolding matrices corresponding to the three different modes are shown in the right half of Fig. 1, where the dark regions represent the unfolding matrices of the newly added subtensor $\mathcal{F}$. With the emergence of the new data subtensors, the column spaces of $\mathrm{A}^*_{(1)}$ and $\mathrm{A}^*_{(2)}$ are extended at the same time when the row space of $\mathrm{A}^*_{(3)}$ is extended. Consequently, *IRTSA* needs to track the changes of these three unfolding spaces, and needs to identify the dominant projection subspaces for a compact representation of the tensor. It is noted that $\mathrm{A}^*_{(2)}$ can be decomposed as: $\mathrm{A}^*_{(2)} = \big(\mathrm{A}_{(2)} \mid \mathrm{F}_{(2)}\big) \cdot \mathrm{P} = \mathrm{B} \cdot \mathrm{P}$, where $\mathrm{B} = \big(\mathrm{A}_{(2)} \mid \mathrm{F}_{(2)}\big)$ and P is an orthonormal matrix obtained by column exchange and transpose operations on an $(I_1 \cdot I_3^*)$-order identity matrix G. Let $G = (\overbrace{E_1}^{I_3} | \overbrace{Q_1}^{I_3^{'}} | \overbrace{E_2}^{I_3} | \overbrace{Q_2}^{I_3^{'}} | \cdots | \cdots | \overbrace{E_{I_1}}^{I_3} | \overbrace{Q_{I_1}}^{I_3^{'}})$ which is generated by partitioning G into $2I_1$ blocks in the column dimension. Consequently, the orthonormal matrix P is formulated as:

$$\mathrm{P} = (E_1|E_2| \cdots | E_{I_1}|Q_1|Q_2| \cdots |Q_{I_1})^T. \tag{1}$$

In this way, CVD($\mathrm{A}^*_{(2)}$) is efficiently computed on the basis of P and CVD(B) obtained by applying R-SVD to B. Furthermore, CVD($\mathrm{A}^*_{(1)}$) is efficiently obtained by performing R-SVD on the matrix $\big(\mathrm{A}_{(1)} \mid \mathrm{F}_{(1)}\big)$. Similarly, CVD($\mathrm{A}^*_{(3)}$) is efficiently obtained

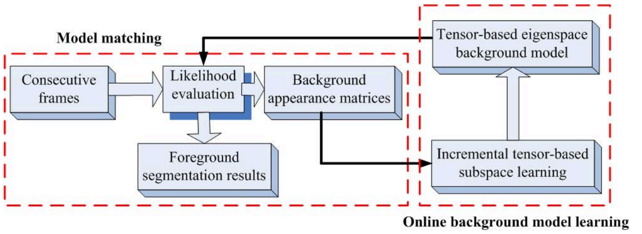
Model matching
Consecutive frames
Likelihood evaluation
Background appearance matrices
Foreground segmentation results
Tensor-based eigenspace background model
Incremental tensor-based subspace learning
Online background model learning

Figure 3: The architecture of the foreground segmentation framework.

by performing R-SVD on the matrix $\left(\frac{\mathrm{A}_{(3)}}{\mathrm{F}_{(3)}}\right)^T$. For a compact eigenspace representation of the mode-i unfolding matrix $\mathrm{A}^*_{(i)}$ $(1 \leq i \leq 3)$, we just maintain the first R_i principal eigenvectors in R-SVD. The specific procedure of *IRTSA* [1] is listed in Fig. 2. The main computational cost of *IRTSA* [1] is to compute the SVDs of unfolding matrices in different modes. Please see the detailed quantitative complexity analysis of R-SVD in [20].

3. THE FRAMEWORK FOR FOREGROUND SEGMENTATION

3.1 Overview of the framework

The foreground segmentation framework based on *IRTSA* includes two stages: (a) online background model learning; and (b) model matching. In the first stage, a low dimensional tensor-based eigenspace background model is online learned by *IRTSA* as new data arrive. In the second stage, consecutive frames are matched with the learned tensor-based eigenspace background model to detect moving regions over the time. These two steps are executed repeatedly as time progresses. The architecture of the foreground segmentation framework is shown in Fig. 3.

3.2 Problem formulation for foreground segmentation

Denote $\mathcal{G} = \{BM_q \in \mathcal{R}^{M\times N}\}_{q=1,2,\ldots,t}$ as a scene's background appearance sequence with the q-th frame being BM_q. For convenience, we rename $\mathcal{G} = \{BM_q \in \mathcal{R}^{M\times N}\}_{q=1,2,\ldots,t}$ as a background appearance tensor (i.e., a background appearance multidimensional matrix). Denote p_{uv} as the u-th and v-th pixel of the scene. We just use a K-neighbor background appearance subtensor $\mathcal{A} = \{BM^{uv}_q \in \mathcal{R}^{I_1\times I_2\times t}\}_{q=1,2,\ldots,t}$ (i.e., the spatio-temporal K-neighborhood of p_{uv}, and $K = I_1{\cdot}I_2 - 1$) to capture the spatio-temporal interactions between the u-th and v-th pixel and its neighbor pixels. In this paper, K is chosen to be 24 (i.e., the spatio-temporal 24-neighborhood of p_{uv}). Consequently, effectively mining the spatio-temporal statistical properties of the subtensor $\mathcal{A}$ is crucial for robust foreground segmentation. The aforementioned formulations are illustrated by Fig. 4. Subsequently, the proposed *IRTSA* is enabled to make tensor-based subspace analysis over $\mathcal{A}$ for effectively mining the statistical properties of $\mathcal{A}$.

Now we are ready to discuss the two proposed background models (*IRTSA-GBM* and *IRTSA-CBM*) respectively in the next two sections 3.3 and 3.4.

3.3 Grayscale background model (*IRTSA-GBM*)

The tensor-based eigenspace model for an existing tensor $\mathcal{A} = \{BM^{uv}_q \in \mathcal{R}^{I_1\times I_2\times t}\}_{q=1,2,\ldots,t}$ ($I_1 = I_2 = 5$ in the experiments) consists of the maintained eigenspace dimensions (R_1, R_2, R_3) corresponding to three tensor unfolding modes, the mode-n column projection matrices $U^{(n)} \in \mathcal{R}^{I_n\times R_n}$ $(1 \leq n \leq 2)$, the mode-3 row projection matrix $V^{(3)} \in \mathcal{R}^{(I_1 I_2)\times R_3}$, the column means $\bar{L}^{(1)}$ and $\bar{L}^{(2)}$ of the mode-$(1,2)$ unfolding matrices $\mathrm{A}_{(1)}$ and $\mathrm{A}_{(2)}$, and the row mean $\bar{L}^{(3)}$ of the mode-3 unfolding matrix $\mathrm{A}_{(3)}$. Given the K-neighbor image region $\mathcal{J}^{uv}_{t+1} \in \mathcal{R}^{I_1\times I_2\times 1}$ centered at the u-th and v-th pixel p_{uv} of a new frame $\mathcal{J}_{t+1} \in \mathcal{R}^{M\times N\times 1}$, the distance RM_{uv} (determined by the three reconstruction error norms of the three modes) between $\mathcal{J}^{uv}_{t+1}$ and the learned tensor-based eigenspace model is formulated as:

$$\begin{aligned} RM_{uv} &= \sqrt{(\omega_1{\cdot}\|Q_1\|^2 + \omega_2{\cdot}\|Q_2\|^2 + \omega_3{\cdot}\|Q_3\|^2)\,/(I_1{\cdot}I_2)}; \\ Q_n &= (\mathcal{J}^{uv}_{t+1} - \mathcal{M}_n) - (\mathcal{J}^{uv}_{t+1} - \mathcal{M}_n) \times_n (U^{(n)} {\cdot}\, U^{(n)^T}), \quad n = 1,2; \\ Q_3 &= (\mathrm{J}^{uv}_{(3)} - \mathrm{M}_3) - (\mathrm{J}^{uv}_{(3)} - \mathrm{M}_3) \cdot (V^{(3)} {\cdot}\, V^{(3)^T}); \end{aligned} \tag{2}$$

where $\times_n$ is the mode-n tensor product (detailed in [1]), $\|\cdot\|$ is the Frobenius norm, ω_k is the mode-k weight ($\sum_{k=1}^{3} \omega_k = 1$ s.t. $\omega_k \geq 0$, and $\omega_k = \frac{1}{3}$ in the experiments), $\mathrm{J}^{uv}_{(3)}$ is the mode-3 unfolding matrix of $\mathcal{J}^{uv}_{t+1}$, $\mathrm{M}_3 = \bar{L}^{(3)}$ which is the row mean of the mode-3 unfolding matrix $\mathrm{A}_{(3)}$, $\mathcal{M}_1$ and $\mathcal{M}_2$ are defined as:

$$\begin{aligned} \mathcal{M}_1 &= (\, \overbrace{\bar{L}^{(1)}, \ldots, \bar{L}^{(1)}}^{I_2} \,) \in \mathcal{R}^{I_1\times I_2\times 1} \\ \mathcal{M}_2 &= (\, \overbrace{\bar{L}^{(2)}, \ldots, \bar{L}^{(2)}}^{I_1} \,)^T \in \mathcal{R}^{I_1\times I_2\times 1} \end{aligned} \tag{3}$$

where $\bar{L}^{(1)}$ and $\bar{L}^{(2)}$ are the column means of the mode-$(1,2)$ unfolding matrices $\mathrm{A}_{(1)}$ and $\mathrm{A}_{(2)}$, respectively. In this way, the criterion for foreground segmentation is defined as:

$$p_{uv} \in \begin{cases} \text{background} & \text{if } \exp\left(-\frac{RM^2_{uv}}{2\sigma^2}\right) > T_{gray} \\ \text{foreground} & \text{otherwise,} \end{cases} \tag{4}$$

where p_{uv} is the u-th and v-th pixel of the scene, σ is a scaling factor, and T_{gray} denotes a threshold. Thus, the entry $BM_{t+1}(u,v)$ of the background appearance matrix BM_{t+1} (referred in Sec. 3.2) at time $t+1$ is defined as:

$$BM_{t+1}(u,v) = \begin{cases} \mathcal{H}_{uv} & \text{if } p_{uv} \in \text{foreground} \\ \mathcal{J}_{t+1}(u,v) & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{H}_{uv} = (1-\alpha^*)\mathfrak{BM}_t(u,v) + \alpha^*\mathcal{J}_{t+1}(u,v)$, α^* is a learning rate factor, and $\mathfrak{BM}_t$ with the entry $\mathfrak{BM}_t(u,v)$ is the mean matrix of $BM_{1:t}$ at time t, i.e., $\mathfrak{BM}_t = \frac{1}{t}\sum_{k=1}^{t} BM_k$. Typically, $\mathfrak{BM}_t$ is computed recursively as: $\mathfrak{BM}_t = \frac{t-1}{t}\mathfrak{BM}_{t-1} + \frac{1}{t}BM_t$. Subsequently, *IRTSA* is applied to incrementally update the tensor-based eigenspace model of the K-neighbor background appearance subtensor $BM^{uv}_{1:t}$ of $BM_{1:t}$ as t increases. In the next section 3.4,

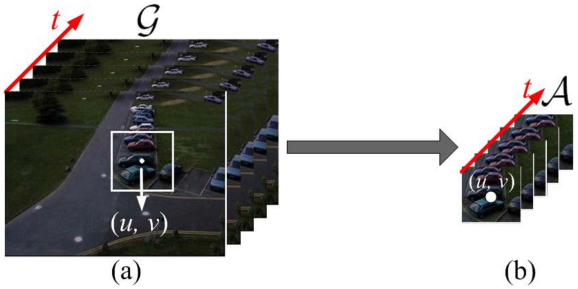

Figure 4: Illustration of the problem formulations for foreground segmentation.

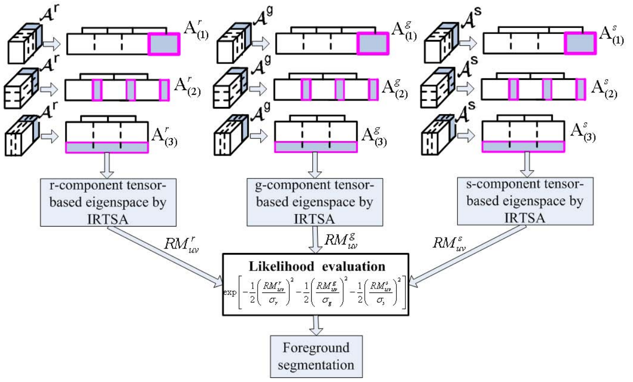
$\mathcal{A}^r$
$\mathrm{A}^r_{(1)}$
$\mathrm{A}^r_{(2)}$
$\mathrm{A}^r_{(3)}$
$\mathcal{A}^g$
$\mathrm{A}^g_{(1)}$
$\mathrm{A}^g_{(2)}$
$\mathrm{A}^g_{(3)}$
$\mathcal{A}^s$
$\mathrm{A}^s_{(1)}$
$\mathrm{A}^s_{(2)}$
$\mathrm{A}^s_{(3)}$
r-component tensor-based eigenspace by IRTSA
g-component tensor-based eigenspace by IRTSA
s-component tensor-based eigenspace by IRTSA
RM^r_{uv}
RM^g_{uv}
RM^s_{uv}
Likelihood evaluation
$\exp\left[-\frac{1}{2}\left(\frac{RM^r_{uv}}{\sigma_r}\right)^2-\frac{1}{2}\left(\frac{RM^g_{uv}}{\sigma_g}\right)^2-\frac{1}{2}\left(\frac{RM^s_{uv}}{\sigma_s}\right)^2\right]$
Foreground segmentation

Figure 5: Illustration of the foreground segmentation process using *IRTSA-CBM*.

we discuss the proposed color background model, which is an extension to the proposed *IRTSA-GBM*.

3.4 Color background model (*IRTSA-CBM*)

In *IRTSA-CBM*, the RGB color space is transformed into the scaled one (r, g, s), where $r = R/(R+G+B)$, $g = G/(R+G+B)$, and $s = (R+G+B)/3$. Let $\mathcal{A}^r \in \mathcal{R}^{I_1 \times I_2 \times t}$ be the r-component image ensemble composed of t background appearance matrices $BM^r_{1:t}$, $\mathcal{A}^g \in \mathcal{R}^{I_1 \times I_2 \times t}$ be the g-component image ensemble composed of t background appearance matrices $BM^g_{1:t}$, $\mathcal{A}^s \in \mathcal{R}^{I_1 \times I_2 \times t}$ be the s-component image ensemble composed of t background appearance matrices $BM^s_{1:t}$, $\mathcal{J}^r_{t+1} \in \mathcal{R}^{I_1 \times I_2 \times 1}$ be the r-component frame at time $t+1$, $\mathcal{J}^g_{t+1} \in \mathcal{R}^{I_1 \times I_2 \times 1}$ be the g-component frame at time $t+1$, and $\mathcal{J}^s_{t+1} \in \mathcal{R}^{I_1 \times I_2 \times 1}$ be the s-component frame at time $t+1$. In this way, we have three 3-order tensors $BM^r_{1:t}$, $BM^g_{1:t}$, and $BM^s_{1:t}$ corresponding to the (r, g, s) components, respectively. For each component, a component-specific tensor-based eigenspace model is learned by *IRTSA*. The learning process of *IRTSA-CBM* is similar to that of *IRTSA-GBM*, and the difference between *IRTSA-GBM* and *IRTSA-CBM* is that *IRTSA-CBM* has three tensor-based eigenspace models corresponding to three color components while *IRTSA-GBM* only has one. Specifically, the tensor-based eigenspace model for $BM^\triangle_{1:t}$ ($\triangle \in \{r, g, s\}$) consists of the maintained eigenspace dimensions $(R^\triangle_1, R^\triangle_2, R^\triangle_3)$ corresponding to three tensor unfolding modes, the mode-n column projection matrices $U^{(n)}_\triangle \in \mathcal{R}^{I_n \times R^\triangle_n}$ for $1 \leq n \leq 2$, the mode-3 row projection matrices $V^{(3)}_\triangle \in \mathcal{R}^{(I_1 I_2) \times R^\triangle_3}$, the column means $\bar{L}^{(1)}_\triangle$ and $\bar{L}^{(2)}_\triangle$ of the mode-$(1, 2)$ unfolding matrices $\mathrm{A}^\triangle_{(1)}$ and $\mathrm{A}^\triangle_{(2)}$, the row means $\bar{L}^{(3)}_\triangle$ of the mode-3 unfolding matrix $\mathrm{A}^\triangle_{(3)}$, and $\triangle \in \{r, g, s\}$. The (r, g, s)-component distance matrices between the new frame and the learned tensor-based eigenspace models are respectively represented as RM^r_{uv}, RM^g_{uv} and RM^s_{uv}, each of which has the same definition as Eq.(2). Given a new frame $\mathcal{J}_{t+1} = \{\mathcal{J}^\triangle_{t+1} \in \mathcal{R}^{I_1 \times I_2 \times 1}\}_{\triangle \in \{r,g,s\}}$, the criterion for foreground segmentation is defined as:

$$p_{uv} \in \begin{cases} \text{background} & \text{if } \mathcal{P}_{uv} > T_{color} \\ \text{foreground} & \text{otherwise}, \end{cases} \tag{6}$$

where $\mathcal{P}_{uv} = \exp\left[-\frac{1}{2}\left(\frac{RM^r_{uv}}{\sigma_r}\right)^2 - \frac{1}{2}\left(\frac{RM^g_{uv}}{\sigma_g}\right)^2 - \frac{1}{2}\left(\frac{RM^s_{uv}}{\sigma_s}\right)^2\right]$, p_{uv} is the u-th and v-th pixel of the scene, σ_r, σ_g and σ_s are three scaling factors, and T_{color} is a threshold. Let $BM^r_{t+1} \in \mathcal{R}^{I_1 \times I_2}$, $BM^g_{t+1} \in \mathcal{R}^{I_1 \times I_2}$, and $BM^s_{t+1} \in \mathcal{R}^{I_1 \times I_2}$ respectively be the (r, g, s)-component background appearance matrices at time $t+1$, whose entry $BM^\triangle_{t+1}(u, v)$ is defined as:

$$BM^\triangle_{t+1}(u, v) = \begin{cases} \mathcal{H}^\triangle_{uv} & \text{if } p_{uv} \in \text{foreground} \\ \mathcal{J}^\triangle_{t+1}(u, v) & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{H}^\triangle_{uv} = (1 - \alpha_\triangle)\mathfrak{BM}^\triangle_t(u, v) + \alpha_\triangle \mathcal{J}^\triangle_{t+1}(u, v)$, $\alpha_\triangle$ is a learning rate factor, and $\mathfrak{BM}^\triangle_t$ is the mean matrix of $BM^\triangle_{1:t}$ at time t, i.e., $\mathfrak{BM}^\triangle_t = \frac{1}{t}\sum_{k=1}^{t} BM^\triangle_k$. Typically, $\mathfrak{BM}^\triangle_t$ is computed recursively as: $\mathfrak{BM}^\triangle_t = \frac{t-1}{t}\mathfrak{BM}^\triangle_{t-1} + \frac{1}{t}BM^\triangle_t$ for $\triangle \in \{r, g, s\}$. Subsequently, *IRTSA* is applied to incrementally update the component-specific tensor-based eigenspace models of the K-neighbor background appearance subtensor $BM^{uv^\triangle}_{1:t}$ (centered at the u-th and v-th pixel p_{uv}) of $BM^\triangle_{1:t}$ as t increases (i.e., each component corresponds to a specific tensor-based eigenspace model learned in the same way of learning the tensor-based eigenspace model for *IRTSA-GBM* in Sec. 3.3). For a better understanding, Fig. 5 is used to illustrate the foreground segmentation process by *IRTSA-CBM*.

4. EXPERIMENTS

In order to evaluate the performance of the proposed framework for foreground segmentation, four videos are used in the experiments. The first two videos consist of 8-bit grayscale images while the last two videos are composed of 24-bit color images. In the first video (selected from PETS2001[1]), a person and vehicles enter or leave a bright road scene. In the second video, three persons are walking in a scene containing a building wall, two lightly swaying trees, two cars and so on. The occlusion event, in which these three persons are overlapped, takes place in the middle of the video stream. In the third video, two cars are moving in a dark and blurry traffic scene. In the last video (selected from CAVIAR[2]), several people are walking along a corridor. They come into or leave the corridor from time to time. For the tensor-based eigenspace representation, the settings of the ranks R_1, R_2 and R_3 in *IRTSA* are obtained from the experiments. The tensor-based eigenspace background models (i.e., *IRTSA-GBM* and *IRTSA-CBM*) are updated every three frames.

Four experiments are conducted to demonstrate the claimed contributions of the proposed *IRTSA-GBM* and *IRTSA-CBM*. The first two experiments are performed to evaluate the foreground segmentation performances of the two subspace analysis based foreground segmentation techniques—the one proposed in [4] (referred here as IRSL) and the proposed *IRTSA-GBM* using grayscale videos 1 and 2, respectively. The last two experiments are performed to evaluate the foreground segmentation performances of the proposed *IRTSA-CBM* using color videos 3 and 4, respectively. IRSL [4] is a representative image-as-vector linear subspace learning algorithm which incrementally learns a low dimensional eigenspace representation of a real scene by online PCA. It has been proven in the literature that IRSL is able to obtain a visually feasible foreground segmentation results. Moreover, IRSL is only available for modeling grayscale images. Thus, it is very significant for the proposed *IRTSA-GBM* to make a comparison with IRSL. Furthermore, the parameter settings for the comparing methods are conducted to make them perform best.

In the first experiment, R_1, R_2 and R_3 for *IRTSA* are assigned as 3, 3, and 10, respectively. The scaling factor σ in *IRTSA-GBM* is set as 15. The threshold T_{gray} is chosen as 0.8. The learning rate factor α^* is assigned as 0.08. For IRSL [4], the PCA dimensionality

[1] http://www.cvg.cs.rdg.ac.uk/slides/pets.html

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

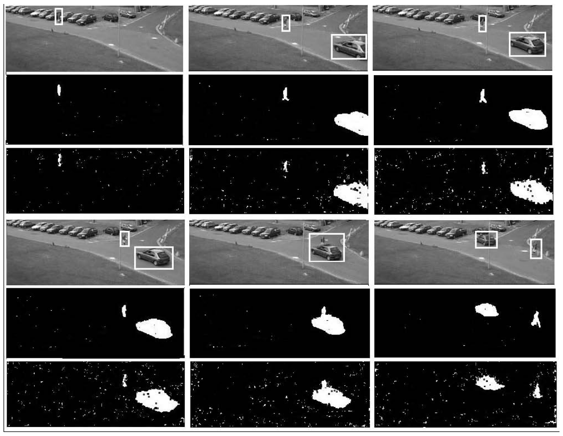

Figure 6: The foreground segmentation results of *IRTSA-GBM* and IRSL using the first video. In rows 1 and 4, the moving regions are highlighted by white boxes. Rows 2 and 5 correspond to *IRTSA-GBM* while rows 3 and 6 are associated with IRSL.

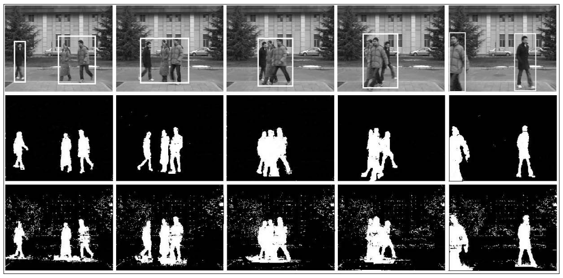

Figure 7: The foreground segmentation results of *IRTSA-GBM* and IRSL using the second video. In row 1, the moving regions are highlighted by white boxes. Rows 2 and 3 correspond to *IRTSA-GBM* and IRSL, respectively.

$p = 12$, the update rate $\alpha = 0.96$, and the coefficient $\beta = 11$. The final foreground segmentation results are shown in Fig. 6, where the second and the fifth rows correspond to *IRTSA-GBM* while the third and the sixth ones are associated with the IRSL. For a better visualization, we just show the segmentation results of six representative frames 2, 43, 68, 86, 117, and 154.

In the second experiment, R_1, R_2 and R_3 for *IRTSA* are assigned as 3, 3, and 12, respectively. The scaling factor σ in *IRTSA-GBM* is set as 20. The threshold T_{gray} is chosen as 0.81. The learning rate factor α^* is assigned as 0.09. For IRSL, the PCA dimensionality $p = 13$, the update rate $\alpha = 0.95$, and the coefficient $\beta = 9$. The final foreground segmentation results are shown in Fig. 7, where the second row corresponds to *IRTSA-GBM* while the third one is associated with IRSL. The segmentation results of five representative frames 7, 26, 32, 44, and 72 are displayed.

From the results in the first and the second experiments, we note that *IRTSA* demonstrates a better foreground segmentation result than IRSL. Specifically, *IRTSA-GBM*'s segmentation results are cleaner, more connected, and less noisy, and more shadow-free. This is due to the fact that since the spatial correlation information is ignored in IRSL, the global or local variations of a scene substantially change the vector eigenspace representation of IRSL.

In the third experiment, (R_1^r, R_2^r, R_3^r), (R_1^g, R_2^g, R_3^g), and (R_1^s, R_2^s, R_3^s) for *IRTSA*, corresponding to three components in the (r, g, s) color space, are respectively assigned as (3, 3, 11), (3,3,11) and (3, 3, 10). The learning rate factors α_r, α_g and α_s are all assigned as 0.08. The scaling factors σ_r, σ_g and σ_s in (6) are set as 0.12, 0.13, and 16, respectively. The threshold T_{color} is chosen as 0.79. The final foreground segmentation results are demonstrated in Fig. 8, where row 2 displays the corresponding foreground segmentation results of *IRTSA-CBM*, in which five representative frames (3, 20, 30, 34, and 38) of the video stream are shown.

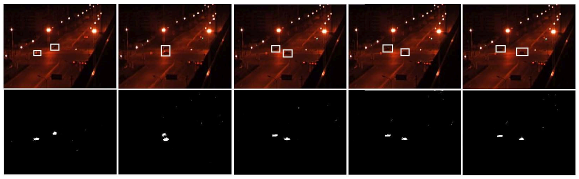

Figure 8: The foreground segmentation results of *IRTSA-CBM* using the third video. In row 1, the moving regions are highlighted by white boxes. Row 2 displays the corresponding foreground segmentation results of *IRTSA-CBM*.

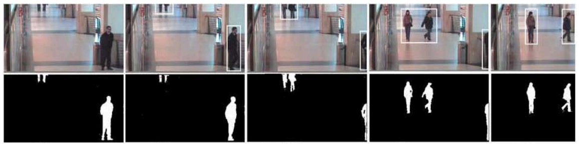

Figure 9: The foreground segmentation results of *IRTSA-CBM* using the fourth video. In row 1, the moving regions are highlighted by white boxes. Row 2 shows the corresponding foreground segmentation results of *IRTSA-CBM*.

In the fourth experiment, (R_1^r, R_2^r, R_3^r), (R_1^g, R_2^g, R_3^g), and (R_1^s, R_2^s, R_3^s) for *IRTSA*, corresponding to the three components in the rgs color space, are respectively assigned as (3,3,9), (3,3,9), and (3,3,11). The learning rate factors α_r, α_g, and α_s are all assigned as 0.08. The scaling factors σ_r, σ_g and σ_s in (6) are set as 0.11, 0.13, and 20, respectively. The threshold T_{color} is chosen as 0.78. The final foreground segmentation results are demonstrated in Fig. 9, where row 2 shows the corresponding foreground segmentation results of *IRTSA-CBM*, in which five representative frames (296, 312, 472, 790, and 814) of the video stream are shown.

From the results in the third and the fourth experiments, we note that *IRTSA-CBM* secures a good foreground segmentation result. *IRTSA-CBM* is able to fully exploit the spatio-temporal redundancies within the image ensembles by tensor-based subspace analysis, resulting in robust foreground segmentation results.

In summary, we observe that *IRTSA-GBM* and *IRTSA-CBM* perform well in complex scenarios. Consequently, *IRTSA-GBM* and *IRTSA-CBM* are two effective models for foreground segmentation.

5. CONCLUSION

In this paper, we have developed an effective framework for foreground segmentation. In the framework, two novel background models (i.e., *IRTSA-GBM* and *IRTSA-CBM*) have been proposed for robust foreground segmentation. These two background models are based on *IRTSA* [1], which incrementally learns a low-order tensor-based eigenspace representation through adaptively updating the sample mean and eigenbasis. Compared with existing background models, the proposed *IRTSA-GBM* or *IRTSA-CBM* better captures the intrinsic spatio-temporal characteristics of a scene, leading to robust foreground segmentation results. Experimental results have demonstrated the robustness and promise of the proposed *IRTSA-GBM* and *IRTSA-CBM*.

6. 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). Z.Z. is supported in part by NSF (IIS-0535162).

7. REFERENCES

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

[2] C. Stauffer, and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in *Proc. CVPR'99,* Vol. 2, 1999.

[3] I. Haritaoglu, D. Harwood, and L.S. Davis, "W^4: Real-Time Surveillance of People and Their Activities," *IEEE Trans. PAMI.,* Vol. 22, Iss. 8, pp.809-830, 2000.

[4] Y. Li, "On Incremental and Robust Subspace Learning," *Pattern Recognition,* Vol. 37, Iss. 7, pp.1509-1518, 2004.

[5] Y. Sheikh, and M. Shah, "Bayesian Object Detection in Dynamic Scenes," in *Proc. CVPR'05,* Vol. 1, pp.74-79, 2005.

[6] J. Cezar Silveira Jacques, C. Rosito Jung, and S.R. Musse, "A Background Subtraction Model Adapted to Illumination Changes," in *Proc. ICIP'06,* pp.1817-1820, 2006.

[7] Y. Wang, T. Tan, K.F. Loe, and J.K. Wu, "A Probabilistic Approach for Foreground and Shadow Segmentation in Monocular Image Sequences," *Pattern Recognition,* Vol. 38, Iss. 11, pp.1937-1946, Nov. 2005.

[8] Y. Wang, K. Loe, and J. Wu, "A Dynamic Conditional Random Field Model for Foreground and Shadow Segmentation ," *IEEE Trans. PAMI.,* Vol. 28, Iss. 2, pp.279-289, 2006.

[9] Y. Tian, M. Lu, and A. Hampapur, "Robust and Efficient Foreground Analysis for Real-Time Video Surveillance," in *Proc. CVPR'05,* Vol. 1, pp.1182-1187, 2005.

[10] H. Wang and N. Ahuja, "A Background Subtraction Model Adapted to Illumination Changes," in *Proc. CVPR'05,* Vol. 2, pp.346-353, 2005.

[11] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang and H. Zhang, "Discriminant analysis with tensor representation," in *Proc. CVPR'05,* Vol. 1, pp.526-532, June 2005.

[12] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Subspace Analysis of Image Ensembles," in *Proc. CVPR'03,* Vol. 2, pp.93-99, June 2003.

[13] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Subspace Analysis of Image Ensembles: TensorFaces," in *Proc. ECCV'02,* pp.447-460, May 2002.

[14] D. Tao, X. Li, W. Hu, S. Maybank, and X. Wu, "Supervised Tensor Learning," in *Proc. ICDM'05,* Nov. 2003.

[15] X. He, D. Cai and P. Niyogi, "Tensor Subspace Analysis," *NIPS'05,* Dec. 2005.

[16] H. Wang, S. Yan, T. Huang and X. Tang, "A Convergent Solution to Tensor Subspace Learning," in *Proc. IJCAI'07,* 2007.

[17] J. Sun, D. Tao and C. Faloutsos, "Beyond Streams and Graphs: Dynamic Tensor Analysis," *ACM KDD'06,* Aug. 2006.

[18] J. Sun, S. Papadimitriou and P. S. Yu, "Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams," in *Proc. ICDM'06,* Dec. 2006.

[19] 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.

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

[21] L. D. Lathauwer, B.D. Moor and J. Vandewalle, "On the Best Rank-1 and Rank-$(R_1, R_2, \ldots, R_n)$ Approximation of Higher-order Tensors," *SIAM Journal of Matrix Analysis and Applications,* Vol. 21, Iss. 4, pp.1324-1342, 2000.

[22] K. Patwardhan, V. Morellas, and G. Sapiro, "Robust Foreground Detection In Video Using Pixel Layers," *IEEE Trans. PAMI.,* Vol. 30 , Iss. 4, pp.746-751, 2008.