Pattern Recognition 67 (2017) 340–352

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Pedestrian recognition in multi-camera networks using multilevel important salient feature and multicategory incremental learning

Huiyan Wang[a], Yixiang Yan[a], Jing Hua[b], Yutao Yang[a], Xun Wang[a,*], XiaoLan Li[a], John Robert Deller[c], Guofeng Zhang[d], Hujun Bao[d]

[a] School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018, China
[b] Department of Computer Science, Wayne State University, Detroit, 48202, Michigan
[c] Electrical and Computer Engineering Department, Michigan State University, East Lansing, 48824, USA
[d] The State Key Lab of CAD and CG, College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China

ARTICLE INFO

Article history:
Received 10 February 2016
Revised 31 December 2016
Accepted 29 January 2017
Available online 3 February 2017

Keywords:
Video pedestrian recognition
Collaborative multi-camera surveillance
MImSF
ORMIM
Incremental learning
Non-overlapping camera network

ABSTRACT

The ability to recognize pedestrians across multiple camera views is of great importance for many applications in the broad field of video surveillance. Due to the absence of the topology and calibration of distributed cameras, spatio-temporal reasoning becomes unavailable, and therefore only appearance information can be used in real-world scenarios, especially for disjoint camera views. This paper proposes a novel approach based on important salient feature and multi-category transfer incremental learning to recognize pedestrians for long-term tracking in multi-camera networks without space-time cues. An accurate and robust model can be built for pedestrian recognition using few samples. We first propose a novel multi-level important salient feature detection method (MImSF[1]) to formulate the appearance model. Due to environmental changes, the appearances of the pedestrians under the camera can change over time and across space, therefore the classification performance may be impaired. Hence, the appearance models should be continuously updated. We then adopt a novel object recognition multicategory incremental modeling algorithm (ORMIM[2]) to update the appearance model adaptively and recognize the pedestrians based on a classification approach. One of the major advantages of the proposed method is that it can identify new target objects that were never learned in the primary model while improving the matching accuracy of what has been learned. We conduct extensive experiments on CAVIAR, ISCAPS databases and our own databases where the camera views are disjoint and the appearance of objects changes significantly due to variations in the camera viewpoint, illumination, weather and poses. The experiments demonstrate that our proposed model is superior to that of existing classification-based recognition methods in terms of accuracy, robustness and computation efficiency. The developed methodology can be used in retrieval, matching and other real-time video surveillance applications.

1. Introduction

In recent years, multiple camera cooperation surveillance systems have become increasing prevalent in many large public environments, such as banks, stores, airports, stations and squares [1–3]. In video surveillance, it is impossible to monitor a wide area using only a single camera because a single camera cannot provides adequate coverage of a large scale scene due to its finite field of view (FOV). Even a simple surveillance application requires the use of multiple cameras. Therefore, a surveillance system for a wide area must deploy a network of cameras and track objects seen across the different views of those cameras.

Pedestrian recognition and tracking in video streams is a critical problem in many video-based tasks, such as intelligent video surveillance, video retrieval and human-computer interaction. The use of multiple cameras makes it possible for long-distance pedestrian analysis. In practice, in wide-area surveillance, some cameras are configured with overlaps while others are non-overlapping (disjoint) in consideration of cost and computation complexity. The objective of pedestrian recognition is to identify the target ob-

* Corresponding author.
E-mail addresses: cederic@zjgsu.edu.cn (H. Wang), 610605906@qq.com (Y. Yan), jinghua@wayne.edu (J. Hua), 1121082097@163.com (Y. Yang), why_zjgsu@163.com, wx@zjgsu.edu.cn (X. Wang), lixiaolan@mail.zjgsu.edu.cn (X. Li), deller@egr.msu.edu (J.R. Deller), zhangguofeng@cad.zju.edu.cn (G. Zhang), bao@cad.zju.edu.cn (H. Bao).

[1] MImSF is the abbreviation of Multi-level Important Salient Feature detection method

[2] ORMIM is the abbreviation of Object Recognition Multicategory Incremental Modeling algorithm

http://dx.doi.org/10.1016/j.patcog.2017.01.033
0031-3203/

jects at different positions and time instants as well as find out all the tracks of pedestrians seen from different viewpoints. This is also known as re-identification (RE-ID) in multiple camera surveillance. A considerable amount of studies has been carried out on this topic [4–7]. However, our work is quite different from most of these RE-ID methods. In our work, we focus on dealing with the following problems: In a surveillance video, we build a classification model by using some instances of p pedestrians (class). Then, (i) when any of these p pedestrians show up again, we can recognize them with a very high accuracy rate (AR). When a new pedestrian ($p+1$th class) appears, we use several instances of this new class to update the classification model by incremental learning. Then, (ii) the new pedestrian can be subsequently recognized with a very high AR. The pedestrians may undergo significant appearance changes due to different camera parameters, illumination variations, occlusions, camera viewpoints, poses and sensor noises, especially for long distance surveillance. When the recognition performance deteriorates due to the environment changes, (iii) the classification model can maintain a very high AR by incremental learning. In addition, (iv) we want to explore some detailed but important aspects, such as the influence of different quantity of samples for learning on AR, the influence of foreground extraction on AR. From the point of efficiency, (v) we want to achieve a very high AR using only a small amount of instances of each pedestrian for modeling. To our knowledge, up to the present, the state-of-the-art RE-ID methods seldom focus on dealing with these problems. A similar research has been carried out by Teixeira and Luis [8]. However, the problems of (ii), (iv) and (v) are not considered in this work. And our method achieved much higher AR than the method proposed in [8]. We compared it with other methods in our previous work [9].

Many previous methods of pedestrian recognition and tracking across multiple cameras were formulated under the assumption that camera topology or association is known. Therefore, camera calibration is required in these methods. Rahimi et al. [1] described a method to simultaneously recover the trajectory of a target and the external calibration parameters of non-overlapping cameras in a multi-camera system under the assumption that the association was known. Makris et al. [2] derived a model of activity for a multi-camera surveillance network to determine the network's topography. Their method assumed that all departure and arrival pairs within a time window were implicitly associated. In these pioneer studies, the spatio-temporal relationships among multiple cameras were extensively used.

However, in a large scale network of multi-camera system, cameras are difficult to calibrate. Especially for the disjoint cameras, the captures are independent of one another, and the tracking information is discontinuous. The spatio-temporal reasoning requires camera calibration and the knowledge of topology of the camera network. The camera calibration of a large camera network is very difficult and time-consuming. Existing calibration methods have various limitations and may not be efficient or accurate enough in deploying a large number of surveillance cameras [10]. In addition, it is difficult to determine the topology of the camera network structure and infer the distribution of transition times between cameras because the cameras may be widely separated in space and time. Therefore, the traditional trajectory-based pedestrian recognition methods are usually unavailable. These constraints make pedestrian recognition a considerably challenging problem. In this situation, spatio-temporal information is not available and only visual appearance information can be used.

A lot of work has been recently reported in the field of appearance-based object recognition. The features developed to build the appearance model include color, shape and texture. However, these appearance features are sensitive to changes in illumination and deformable geometry of objects. As a result, many research studies have been carried out to address these problems. Prosser et al. [11] proposed a cumulative brightness transfer function for mapping color between cameras located at different physical sites, which makes a better use of the available color information from a very sparse training set. Madden et al. [12] proposed an illumination-tolerant appearance representation base on an online k-means color clustering algorithm which is capable of coping with the typical illumination changes. Recently, the scale invariant feature transform (SIFT) has been used as another feature for object matching [13]. SIFT features are robust to illumination changes but suffer from high-dimensional descriptors. Teixeira and Luis [8] proposed a novel algorithm to identify objects using a vocabulary tree to quantize SIFT features and obtain a multi-dimensional vector. To improve efficiency and accuracy, in our previous work [9], a vocabulary tree vector was combined with color features, and then introduced an approximate kernel-based PCA algorithm (AKPCA) to fuse these features. This method was found to be effective in recognizing objects. However, AKPCA is a nonlinear dimension reduction method and it is memory and time-consuming. Most of these feature-based approaches depended on the spatio-temporal relationship among multiple cameras or were based on a set of assumptions that do not usually apply in a real world scenario. These challenges make the feature-based pedestrian recognition a difficult task.

Generally, pedestrian recognition can be conducted using classifiers or distance metrics, such as Euclidean distance and Mahalanobis distance. For wide-area video surveillance, the appearance of objects captured in different cameras is likely to change over time and new objects may appear. Here, a new object indicates the pedestrian who has no instance in the primary dataset, while having several instances in the new data. To obtain and maintain a high level of recognition accuracy and identify new objects, the appearance model should be continuously updated when new data are available. In this situation, incremental learning classifiers are good choices and superior to traditional classifiers and distance metrics. Incremental learning is a machine learning method in which the learning process is performed whenever a new sample is added to update the model. Compared to traditional machine learning methods, incremental learning does not require a sufficient training set before the training process. Instead, the training examples can be augmented over time.

In this paper, we propose a new methodology for pedestrian recognition across multi-camera networks using only visual information in which spatio-temporal reasoning is unavailable. The major contribution of this paper is three-fold. First, we propose a novel multi-level important salient feature detection (MImSF) method based on data-adapting convolution filters and a data-driven algorithm, and then aggregate our important saliency map with color features to formulate an appearance model. MImSF can extract main representative features with a high discriminative power and the feature aggregation can improve the robustness of the appearance model. Second, we propose a novel support-vector-machine (SVM) based incremental learning method by using modified regularization terms to build and update the appearance model online and recognize the object based on a classification method. We name our method as object recognition multi-category incremental modeling algorithm (ORMIM). Given that the scene undergoes large changes during the period of tracking, for example, when the model was built from a video captured in an evenly illuminated indoor scene, and the target object moved to lighted street at night, the recognition performance will be seriously impaired. Our model solves this problem because it can be built and updated by using or adding few samples. With incremental learning, our model can maintain a very high recognition rate. Third, the proposed approach can effectively discriminate new target objects that were never learned in the primary model and si-

multaneously improve the matching accuracy of old objects. The proposed methodology can achieve a much better performance in recognition accuracy, computation efficiency, robustness, requirement for computing and storage space than existing state-of-the-art classification-based recognition methods.

The paper is organized as follows: In Section 2, we propose a novel framework for appearance modeling. In Section 3, we present and discuss the proposed incremental classification algorithm. In Section 4, the architecture of our method is described. Next, Section 5 reports the experimental results and discussion. Finally, the conclusion is drawn in Section 6.

2. Pedestrian recognition via incremental learning

The pedestrian recognition problem can be formulated as learning and updating appearance models and decision functions from query images to successfully determine whether or not the candidates observed in camera networks belong to one of the object targets. It can be defined as follows:

Definition 2.1. Let V denotes a set of pedestrian images observed over a camera network. Given a sequence of training dataset $S_1, S_2, \ldots, S_n$, where $S_i = \{(x_{i_j}, y_{i_j})\} \subset V$, $x_{i_j} \in \mathbb{R}^d$ denotes a pedestrian image, $1 \le i \le n$, $1 \le j \le n_i$. $y_{i_j} \in C_i = \{1, \ldots, K_i\} \subseteq \{1, \ldots, K\}$. C_i indicates the set of class label in S_i, K is the total number of pedestrians of interest in V. Let A_1 denotes the appearance model built on S_1. Based on A_1 a primary decision function vector $F_1 = \{f_{1_1}, \ldots, f_{1_{K_1}}\}$ is trained. The decision function F_1 assigns a candidate $X_i \in V$ to the class label k with $k = \arg_{j=1,\ldots,K_1} \max f_{1_j}(x_i)$. Then the incremental learning procedure L_p of pedestrian recognition can be illustrated as: $L_p(S_i, A_i, F_{i-1}) = F_i, \quad 2 \le i \le n$.

Here the functions f_{i_j} are defined by $f_{i_j}(x) = w_{i_j}^T \phi(x) + b_{i_j}$. For long-distance pedestrian surveillance, new target objects may appear. Some pedestrians may disappear from the camera network and never show up again, and in this case, the class labels are not used again, but the class labels still exist in the set of class labels. It does not affect the recognition of other target objects or new target objects. Therefore, without loss of generality, we only consider the case where the number of class labels do not decrease during the incremental learning procedure, i.e., $C_1 \subseteq C_2 \subseteq \cdots \subseteq C_n$. The procedure of pedestrian recognition is to build a primary recognition model F_1 based on an effective appearance representation A_1 through a sample set S_1, and when the changes of scene degrade the performance of the model F_1 or a new object appears, the model F_1 can be updated to F_2 according to the new sample set S_2 and the new appearance representation A_2, and so forth. The main objective of this paper is to explore an effective pedestrian appearance model A_i and then an accurate recognition model F_i can be built and updated by only a very small number of samples. The model F_i can achieve a very high recognition performance and effectively recognize new objects.

3. A novel framework for appearance modeling based on MImSF

In this section, we present a novel framework for appearance modeling based on a proposed feature extraction method by exploring the most important salient features. We first give the overview of saliency detection methods and then describe the proposed MImSF method and present the construction of the appearance model.

3.1. Overview

Visual saliency measures low-level stimuli to the human brain and visual system. In image/video saliency analysis, salient features are the image regions that most likely attract the human visual attention. Saliency is often attributed to variations in important image features, such as color, gradient, edges and boundaries; thus, extracting salient features is very helpful for object recognition. Detecting the salient regions is now becoming one of the most important tasks in many computer vision and graphics applications. A considerable amount of research has been performed on salient feature detection. The primary salient features are geometry-based features, such as edges, lines, ridges, corners. Recently, many bottom-up models (or systems) were proposed to calculate the saliency maps, such as region-based approaches [14]. Kadir and Brady [15] proposed a scale saliency algorithm to detect the salient points by estimating the information content in circular neighborhoods at different scales in terms of entropy. This algorithm is stable and robust, but the choice of parameters is complicated and the computational cost is high. Several simple and fast algorithms have also been proposed to cope with these disadvantages such as spectral residual [16] and phase spectrum of Fourier transform [17]. These two approaches are independent of parameters and are able to detect salient objects rapidly, but are unable to highlight the uniform salient regions for large objects. Cheng et al. [18] proposed a histogram based contrast and a spatial information-enhanced region based contrast saliency extraction algorithm. This approach is simple and efficient in calculating saliency maps. Nearly all of these methods only consider static images rather than video sequences. To cope with this problem, the concept of saliency has been used for space-time content-based video retrieval and activity recognition, and many researchers developed models to generate spatio-temporal saliency map. The space-time interest points are those points where the local neighborhood has a significant variation in both the spatial and temporal domain. Laptev et al. [19] proposed a well-known space-time interest point detector which uses an extension of the Harris corner detector to detect spatiotemporal events. Willems et al. [20] identified saliency as the determinant of a 3D Hessian matrix, which can be efficiently calculated due to the use of integral videos. However, these methods need spatio-temporal information, which is difficult to acquire. In addition, motion features are required, which increases the computational cost and not suitable for applying to real-time system.

3.2. Proposed MImSF

Suppose $S = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m), \ldots, (X_{M_1}, Y_{M_1})\}$ is the training sample set, $m = 1, \ldots, M_1$, X_m is a training image of size $a_0 \times b_0$, $X_m \in \mathbb{R}^d$ and Y_m is the class label, $Y_m \in \{1, \ldots, K\}$, K is the number of pedestrian of interest. The input training sample set S is used for learning the data-adapting convolution filter bank. For each image X_i, we take a $c_1 \times c_2$ patch around each pixel and compute the patch mean. We collect all of these overlapping patches and denote them as $P = \left[p_{i,1}, p_{i,2}, \ldots, p_{i,ab}\right] \in \mathbb{R}^{c_1 c_2}$, where $p_{i,j}$ is the jth vectorized patch in X_i, $a = a_0 - c_1 + 1$, $b = b_0 - c_2 + 1$. Suppose $\hat{p} = p_{i,j} - \overline{p}_{i,j}$, where $\overline{p}_{i,j}$ is the mean value of $p_{i,j}$, we then get

$$\hat{P}_i = \left[\hat{p}_{i,1}, \hat{p}_{i,2}, \ldots, \hat{p}_{i,ab}\right] \in \mathbb{R}^{c_1 c_2}. \tag{1}$$

Constructing the same matrix for each training image in S and putting them together, we obtain

$$P = \left[\hat{P}_1, \hat{P}_2, \cdots, \hat{p}_{M_1}\right] \in \mathbb{R}^{c_1 c_2 \times M_1 ab}. \tag{2}$$

Our objective is to find a family of orthonormal filters $U = [u_1, u_2, \ldots, u_K]$ to minimize the reconstruction error

$$\varepsilon(U) = \| P - UU^T P \|_F^2, \ s.t.\ UU^T = I_{M_2}, \tag{3}$$

where $\| \ \|_F^2$ denotes the Frobenius norm, M_2 is the number of filters and I_{M_2} is identity matrix of size $M_2 \times M_2$. Eq. (3) can also be expressed as

$$\max_{U \in \mathbb{R}^{c_1 c_2 \times M_2}} \mathrm{tr}\left[U^T P P^T U\right], \ s.t.\ UU^T = I_{M_2}. \tag{4}$$

The M_2 principal eigenvectors of PP^T can be viewed as the solution, then the filters can be represented as

$$f_m = H_{c_1,c_2}(V_m) \in \mathbb{R}^{c_1 \times c_2}, \ m = 1, 2, \ldots, M_2, \tag{5}$$

where H_{c_1,c_2} denotes a mapping function from $U \in \mathbb{R}^{c_1 c_2}$ to $f \in \mathbb{R}^{c_1 \times c_2}$, and V_m is the mth principal eigenvectors of PP^T. We select the leading principal eigenvectors to represent the main characteristics of the training image patches. Therefore, we obtain the feature maps X_i^m for ith training image X_i, we called them multi-level important feature maps, that is

$$X_i^m = X_i * f_m, \ m = 1, 2, \ldots, M_2, \tag{6}$$

where $*$ is 2D convolution. To make X_i^m and X_i have the same size, X_i is zero-padded. Borji et al. [21] proposed a model integration scheme for saliency aggregation. Different from [21], we aggregate the M_2 important feature maps $\left\{X_i^m \mid 1 \le m \le M_2\right\}$ to produce a final important feature map I_i using a different combination function to keep maximum values. Then the aggregated important feature value $I_i(q)$ at pixel q of X_i is modeled as the probability

$$I_i(q) = P\left(y_q = 1 \mid I_i^1(q), I_i^2(q), \ldots, I_i^{M_2}(q)\right) \propto \max_{1 \le m \le M_2} I_i^m(q), \tag{7}$$

where $I_i^m(q)$ represents the feature value of pixel q in the important feature map X_i^m, y_q is a binary random variable that takes the value 1 if q is an important pixel and 0 otherwise.

Considering that the biological vision system is sensitive to contrast in visual signal and the regions that contrast strongly with their surroundings [18], the saliency value at each pixel q is defined as

$$S_i(q) = \sum_{q \in N_q} d(q_k, q_h), \tag{8}$$

where $d(q_k, q_h)$ is the color distance metric between *Luv* pixels q_k and q_h, N_q is the neighborhood of pixel q.

In addition to contrast, spatial relationship is another important piece information for human recognition. Therefore, we use spatial-relationship based contrast to detect saliency. We first segment the video images into regions based on a graph-based method [22], and then build the color histogram for each region. The saliency value of each region is calculated using a global contrast score, measured by the region contrast and spatial distances between regions [18]. For a region r_k, the saliency value is given by

$$B_i(r_k) = \sum_{\substack{r_k, r_h \in X_i \\ r_k \ne r_h}} w(r_h) d(r_k, r_h), \tag{9}$$

where $w(r_h)$ denotes the weight of region r_k and $d(r_k, r_h)$ denotes the color distance metric between these two regions. The weights are set to smaller values for farther regions and larger ones for closer regions. The color distance between two regions r_k and r_h is defined as

$$d(r_k, r_h) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} P\left(c_{k,i}\right) P\left(c_{h,j}\right) d\left(c_{k,i}, c_{h,j}\right), \tag{10}$$

where $P(c_{k,i})$ is the probability of the i-th color $c_{k,i}$ among all n_k colors in the kth region r_k. The probability of a color in the normalized color histogram of a region is used as the weight.

To improve the robustness of the appearance model, we propose the multi-level important salient features by aggregating the important feature map I_i and the salient feature map B_i. The aggregated important salient feature values $V_{i,1}(q)$ and $V_{i,2}(q)$ at pixel q of X_i are modeled as

$$V_{i,1}(q) = \max\left(I_i(q), B_i(q)\right) \tag{11}$$

$$V_{i,2}(q) = \frac{1}{2}\left(I_i(q) + B_i(q)\right). \tag{12}$$

Therefore, the total important salient feature vector $\boldsymbol{V_i}$ for X_i can be presented as

$$\boldsymbol{V}_i = \left(\boldsymbol{V}_{i,1} \oplus \boldsymbol{V}_{i,2}\right) \tag{13}$$

in that order, the symbol '$\oplus$' denotes the concatenation of the two sub-vectors.

Color histogram is an effective and common global feature and has high discrimination ability for object recognition due to being robust to the changes in views and poses. Color histogram can be built from various color spaces such as RGB, HSV, and LAB. In our work, the color histogram in the HSV color space is extracted because it gets close to the subjective understanding of color and suitable for addressing the object recognition problem. For a given object, its HSV color histogram feature vector $\boldsymbol{C_i}$ can be represented by

$$\boldsymbol{C}_i = \left\{c_{i_1}, c_{i_2}, \ldots, c_{i_j}, \ldots, c_{i_{256}}\right\}, j = 1, \ldots, 256, \tag{14}$$

where c_{i_j} is the value of each HSV bin.

To obtain the representative features for object recognition, we fuse our important salient feature vector $\boldsymbol{V_i}$ with color feature vector $\boldsymbol{C_i}$ to produce the final appearance model

$$\boldsymbol{A}_i = \left(\boldsymbol{V}_i \oplus \boldsymbol{C}_i\right). \tag{15}$$

The flowchart of our MImSF algorithm is illustrated in Fig. 1. we summarize the main steps as follows. First, we extract multi-level important feature maps X_i^m according to Eq. (6) and aggregate them according to Eq. (7) to produce feature map I_i. Second, salient features S_i are detected based on Eqs. (9) and 10. Third, the important salient features $\boldsymbol{V_i}$ are computed according to Eqs. (11)–(13). Finally, the final appearance model $\boldsymbol{A_i}$ is built by aggregating $\boldsymbol{V_i}$ and color histogram feature vector $\boldsymbol{C_i}$ according to Eq. (15).

4. Pedestrian recognition based on ORMIM

In this section, we describe a new algorithm for learning and updating the appearance model online based on incremental SVM. We first give the overview of incremental methods, and then present the proposed ORMIM method and give the procedure of pedestrian recognition.

4.1. Overview

Generally, pedestrian recognition can be conducted using classifiers or distance metrics, such as Euclidean distance and Mahalanobis distance (M-distance). Recently, distance metric learning has been shown to significantly increase the performance of distance-based classifier. Among the previous work, learning the M-distance for K-nearest neighbor (KNN) classifier has received much attention [23,24]. Mensink et al. [25] explored near-zero shot metric learning (NZ-ML) for KNN, which use large margin nearest neighbor (LMNN) to improve the performance. Extensions

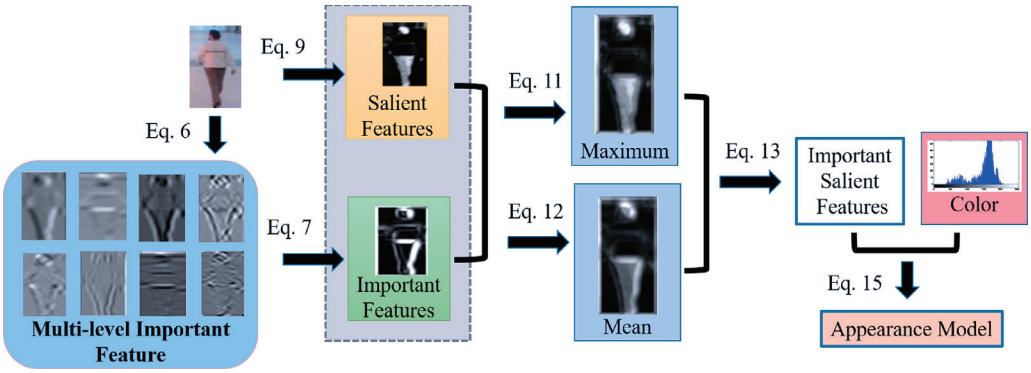


Fig. 1. Processing pipeline of our MImSF algorithm.

of metric learning have also been studied for multiple kernel learning [26,27], semi-supervised learning [28], etc. A novel regularization framework (RF-ML) was developed to learn similarity metrics [29], and the objective function was formulated by incorporating the robustness to the large intra-personal variations. Wang et al. [27] presented a kernel classification framework (KCF) to learn M-distance metric. This paper generalized many metric learning methods into a kernel classification framework, which achieved competitive classification accuracies with state-of-the-art metric learning methods. To show the effectiveness of our proposed method, we compare it with these above mentioned metric learning algorithms.

For wide-area video surveillance, the appearance of pedestrians captured in different cameras is likely to change over time and new objects may appear. To obtain and maintain a high level of recognition, the appearance model should be continuously updated when the scene changes. In this situation, incremental learning classifiers are good choices and superior to traditional classifiers and distance metrics. Compared to traditional machine learning methods, incremental learning does not require a sufficient training set before the training process. Instead, the training examples can be augmented over time.

Significant research has been performed on incremental learning resulting in the development of algorithms such as incremental Bayesian approach [30], Learn ++ [31] and an improved method called Learn ++. MF [32], incremental KNN model [33], etc. Recently, incremental learning methods have been applied to visual tracking. Ross et al. [34] proposed a tracking method which can incrementally learn a low dimensional subspace representation and adapt to appearance changes. Jiang et al. [35] proposed an incremental label consistent K-SVD (LC-KSVD) dictionary learning algorithm for recognition, which combines a new label consistency constraint with the reconstruction error and classification error to construct the object function and the experimental results demonstrate this algorithm outperforms many recently proposed sparse-coding techniques. However, many parameters are required to be regulated and adjusted when new data are available. The regulation of parameters is a difficult task and the recognition results largely depend on these parameters.

In general, a common trend in pedestrian recognition is to detect sparse, informative feature points, which increases robustness to noise, illumination changes and pose variation. However, most of the existing methods rely on geometry models, trajectory association, spatio-temporal relationships and topology of camera networks, which are often unavailable for long distance surveillance. Moreover, most methods require a large set of samples to build a recognition model. Our proposed method can be built based on a very small quantity of samples and does not require any prior information.

4.2. Proposed ORMIM

In long distance video surveillance, incremental learning is an effective tool to construct and update the recognition model, so that history data does not have to be stored and the model can be continuously adapted to changes in appearance. In this study, we proposed a novel SVM based incremental learning algorithm that adds modified regularization terms, and adapts them to variations in regulation and kernel parameters.

Suppose $\hat{S} = \{(\hat{A}_1, \hat{Y}_1), (\hat{A}_2, \hat{Y}_2), \ldots, (\hat{A}_m, \hat{Y}_m), \ldots, (\hat{A}_{M_2}, \hat{Y}_{M_1})\}$ is the primary training sample set, $m = 1, \ldots, M_1$, $\hat{A}_m \in \mathbb{R}^d$ is a feature vector and $\hat{Y}_m$ is its corresponding label, $\hat{Y}_m \in \{1, \ldots, K\}$. We used the formalism of linear classifier for clarity. The primary multicategory classifier is denoted as a matrix $\hat{W} = \{\hat{w}_1, \ldots, \hat{w}_K\}$, where $\hat{w}_i$ is the hyperplane separating one of the K categories from the others. When the new data are available, we use $S = \{(A_1, Y_1), (A_2, Y_2), \ldots, (A_m, Y_m), \ldots, (A_{M_1}, Y_{M_2})\}$ to denote the new sample set, $m = 1, \ldots, M_2$. Suppose that a new object (not belonging to the former K categories, and therefore, assigning a new class label to $K+1$) appears in S, then the new model can be encoded as a new set of $K+1$ hyperplanes, which is described in matrix form as $W = \{w_1, \ldots, w_K, w_{K+1}\}$. Therefore, the label of a given sample image X is predicted as $f_W(X) := \underset{i=1,\ldots,K,K+1}{\operatorname{argmax}} w_i^T X + b_i$. It can be extended to nonlinear domain based on kernels.

A general method to find the set of hyperplanes W is to solve the minimum of a regularized least-squares loss function [36,37]

$$\min_{W,\varepsilon,b} \frac{1}{2}\|W\|^2 + \frac{\beta}{2}\sum_{m=1}^{M}\varepsilon_m^2 \quad . \tag{16}$$
$$s.t. \quad Y_m = wX_m + b + \varepsilon_m, \quad m = 1, \ldots, M.$$

Define $Y_{mi} = \begin{cases} 1, & Y_m = i \\ 0, & otherwise \end{cases}$, then the $K+1$ multicategory SVM objective function is obtained

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{K}\|w_i\|^2 + \frac{1}{2}\|w_{K+1}\|^2 + \frac{\beta}{2}\sum_{m=1}^{M}\sum_{i=1}^{K+1}(w_i^T X_m + b_i - Y_{mi})^2, \tag{17}$$

where $\|w_i\|$ is a regularization term that inversely relates to margin between training images of two categories. In SVM-based multicategory classifier, the margin between two category i and j can be denoted as $\frac{2}{\|w_i - w_j\|}$ [38]. Therefore, to find a new set of hyperplanes and achieve the goal of recognizing new objects, we will implement the transfer learning problem [39] by augmenting regularization terms to minimize the sum of the square of $\|w_i - \beta_j w_j\|$, $i, j = 1, \ldots, K+1$, $i \neq j$. By experiments, we found that if the parameters β_j are set to 1, the computation is much

faster while the performance of our system is not adversely affected. Therefore, we obtain

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{K}\|w_i\|^2 + \frac{1}{2}\|w_{K+1}\|^2 + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=i+1}^{K}\|w_i - w_j\|^2 + \frac{1}{2}\sum_{i=1}^{K}\|w_{K+1} - w_i\|^2 + \frac{\beta}{2}\sum_{m=1}^{M}\sum_{i=1}^{K+1}(w_i^T X_m + b_i - Y_{mi})^2, \quad (18)$$

where w_{K+1} denotes the hyperplane separating the new object from the others. Furthermore, we extend Eq. (18) by aggregating the knowledge of the primary data $\hat{S}$ with the new data S. By treating the $\hat{W}$ as the prior classifier, we add two modified regularization terms to penalize the discrepancy between the new hyperplanes W and the primary hyperplanes $\hat{W}$ and enforce W to remain close to $\hat{W}$. Thus, we get

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{K}\|w_i\|^2 + \frac{1}{2}\|w_{K+1}\|^2 + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=i+1}^{K}\|w_i - w_j\|^2 + \frac{1}{2}\sum_{i=1}^{K}\|w_{K+1} - w_i\|^2 + \frac{\beta}{2}\sum_{m=1}^{M}\sum_{i=1}^{K+1}(w_i^T X_m + b_i - Y_{mi})^2 + \frac{1}{2}\sum_{i=1}^{K}\|w_i - \hat{w}_i\|^2 + \frac{1}{2}\|w_{K+1} - \sum_{i=1}^{K}\hat{w}_i\|^2. \quad (19)$$

Note that, the above optimization is subject to the constrains $Y_m = wX_m + b + \varepsilon_m$. Therefore we can define the objective function in Eq. (19) using Lagrangian

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{K}\|w_i\|^2 + \frac{1}{2}\|w_{K+1}\|^2 + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=i+1}^{K}\|w_i - w_j\|^2 + \frac{1}{2}\sum_{i=1}^{K}\|w_i - \hat{w}_i\|^2 + \frac{\beta}{2}\sum_{m=1}^{M}\sum_{i=1}^{K+1}\varepsilon_{mi}^2 + \frac{1}{2}\|w_{K+1} - \sum_{i=1}^{K}\hat{w}_i\|^2 + \frac{1}{2}\sum_{i=1}^{K}\|w_{K+1} - w_i\|^2 - \sum_{m=1}^{M}\alpha_{mi}\{w_i^T X_m + b_i + \varepsilon_{mi} - Y_{mi}\}, \quad (20)$$

where $\alpha = \{\alpha_{mi}\} \in \mathbb{R}^{M\times(K+1)}$, $i = 1, \ldots, K+1$, $m = 1, \ldots, M$ are Lagrange multipliers. To compute the model parameters (α, b), the optimality conditions for this Eq. (20) can be written as follows

$$\frac{\partial \mathcal{L}}{\partial w_i} = (K+3)w_i - \hat{w}_i - \sum_{m=1}^{M}\alpha_{mi}X_m - w_{K+1} = 0 \quad (21)$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = -\sum_{m=1}^{M}\alpha_{mi} = 0 \quad (22)$$

$$\frac{\partial \mathcal{L}}{\partial w_{K+1}} = (K+2)w_{K+1} - \sum_{i=1}^{K}\hat{w}_i - \sum_{m=1}^{M}\alpha_{m(K+1)}X_m - \sum_{i=1}^{K}w_i = 0 \quad (23)$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon_{mi}} = \beta\varepsilon_{mi} - \alpha_{mi} = 0, \ \ i = 1, \ldots, K+1 \quad (24)$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_{mi}} = w_i^T X_m + b_i + \varepsilon_{mi} - Y_{mi} = 0 \quad (25)$$

From Eq. (21), we get

$$(K+3)w_i - w_{K+1} = \hat{w}_i + \sum_{m=1}^{M}\alpha_{mi}X_m, i = 1, \ldots, K. \quad (26)$$

From Eq. (23), we get

$$-\sum_{i=1}^{K}w_i + (K+2)w_{K+1} = \sum_{i=1}^{K}\hat{w}_i + \sum_{m=1}^{M}\alpha_{m(K+1)}X_m. \quad (27)$$

Expressed in the matrix form, we have shown that

$$PW = \hat{W} + \alpha X, \quad (28)$$

where $P = \begin{bmatrix} K+3 & 0 & \ldots & 0 & -1 \\ 0 & K+3 & \ddots & \vdots & -1 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & K+3 & -1 \\ -1 & -1 & \ldots & -1 & K+2 \end{bmatrix}$, $W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \\ w_{K+1} \end{bmatrix}$, $\hat{W} = \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \vdots \\ \hat{w}_K \\ \sum_{i=1}^{K}\hat{w}_i \end{bmatrix}$, $\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1(K+1)} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2(K+1)} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{M1} & \alpha_{M2} & \ldots & \alpha_{M(K+1)} \end{bmatrix}$, and $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix}$.

Based on the constrain equations (Eqs. (24) and (25)), we reformulate the solution of (α, b) in matrix form

$$\begin{bmatrix} X^T X P^{-1} + \frac{1}{\beta}I & \mathbf{1}_{K,1} \\ \mathbf{1}_{1,K} & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} Y - X^T P^{-1}\hat{W} \\ 0 \end{bmatrix}, \quad (29)$$

where, according to the block matrix inversion lemma, the inverse of P can be expressed as

$$P^{-1} = \frac{1}{\rho}\begin{bmatrix} \rho I + \frac{1}{\delta}\mathbf{1}_K & \mathbf{1}_{K,1} \\ \mathbf{1}_{1,K} & \delta \end{bmatrix}, \quad (30)$$

where $\mathbf{1}_K$, $\mathbf{1}_{K,1}$ and $\mathbf{1}_{1,K}$ are all-ones matrix, the coefficients $\rho = (K^2 + 4K + 6)$, $\delta = K + 3$.

The ORMIM based incremental learning is outlined in Algorithm 1.

Algorithm 1 ORMIM based incremental learning.

Require: the new training dataset $S = \{(X_i, Y_i)\}$, the feature set X_i, the class label Y_i, $Y_i \in \{1, \ldots, K+1\}$, $i = 1, 2, \ldots, M$, the primary multiclass classifier $\hat{W} = \{\hat{w}_1, \ldots, \hat{w}_K\}$

Ensure: $W = \{w_1, \ldots, w_K, w_{K+1}\}$

1. Initialize $\mathbf{1}_K$, $\mathbf{1}_{K,1}$ and $\mathbf{1}_{1,K}$ are all-ones matrix, $\rho = (K^2 + 4K + 6)$, $\delta = K + 3$, compute P^{-1} according to Eq. (30);
2. Compute (α, b) according to Eq. (29);
3. Use Eqs. (26) and (27) to update the primary K hyperplanes $\hat{W}$ and obtain W.

5. Experiment results and analysis

Experiments were implemented with four objectives:

1. To assess the performance of the new appearance model in effectively representing the objects and having a high discriminative power in the pedestrian recognition.
2. To assess the performance of the new incremental modeling algorithm in discerning targets including the new objects that were not learned in the primary model.
3. To assess the performance of the new algorithm in recognition accuracy and computation efficiency.

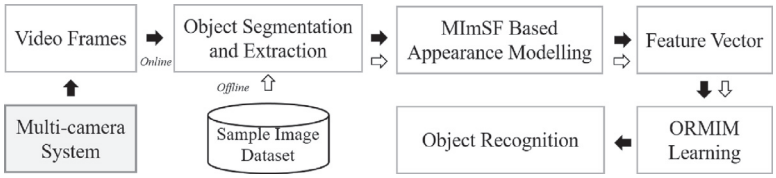
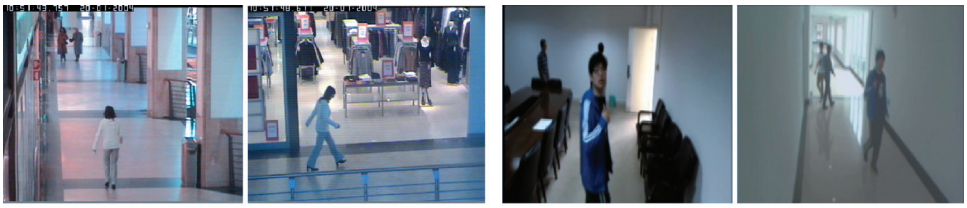


Fig. 2. The architecture of our method, where the solid arrows represent the training process while the hollow arrows represent the offline processing.



(a) CAVIAR (b) Our indoor dataset

Fig. 3. One object captured from two surveillance cameras having different field of view. In (a), the images are captured by two overlapping cameras. In (b), the images are captured by two non-overlapping cameras, the left one is captured in one meeting room and the right one is captured in one corridor.

4. To evaluate the performance of the proposed method in discriminating the target objects when few samples are available.

The main steps of our method are summarized in Fig. 2. A block-based foreground detection (BFD) method [40] was employed to segment and extract the objects in the first step.

5.1. Dataset and experimental environment

To validate the efficiency of the proposed methodology, we use two standard datasets and our own datasets. We adopt the CAVIAR (Context Aware Vision using Image-based Active Recognition) dataset [41] that is widely applied in evaluating the performance of video tracking methods. The dataset was gathered from two cameras having different field of views, as shown in Fig. 3(a). It gives XML-based benchmark information including the segmented object and the bounding box. We extract 15 visual objects based on the XML-based information, and used 5472 object images to train and test our method. Thus, each object has an average of 364 frame images as samples. The ISCAPS (Integrated Surveillance of Crowded Areas for Public Security) [42] is also used to testify our algorithm, in which 2093 object images and 15 visual objects are used.

In addition, to further evaluate the performance of our method in different non-overlapping camera views that undergo significant appearance changes, we built our own datasets, one is an indoor dataset and the other is an outdoor dataset. The indoor dataset is captured in one meeting room and one corridor. As shown in Fig. 3(b), objects were partially occluded by other persons, table and chairs in the meeting room. The lighting condition in the meeting room is relatively uniformly distributed, while the illumination in the corridor is very complicated due to dim light and the light reflection and refraction from window glass and floor. The outdoor dataset includes 10 objects. It is captured in the university campus by four non-overlapping surveillance cameras. The four cameras are far away from another. The outdoor scenes are more complex than indoor scenes, including some walking passersby, riding passersby and cars.

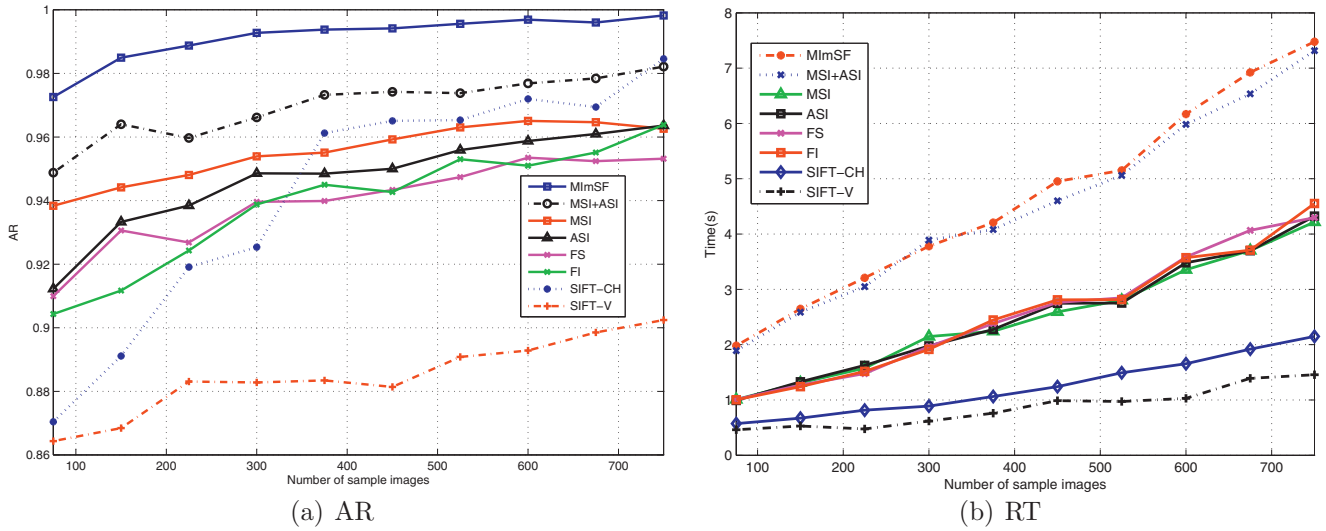
We use 10-fold cross validation to evaluate the models. All algorithm results were achieved by MATLAB using an Intel Xeon E3-1230 V2 processor with a 3.3 GHz frequency and a 4 GB memory.

5.2. Evaluation of the proposed appearance model

In this section, we evaluate the efficiency of the proposed appearance model by comparing it with the salient features, important features and their aggregations. We also compare it with previous appearance models. The experiments were conducted on the CAVIAR dataset, and the 15 persons are the target objects. The results are shown in Fig. 4. In Fig. 4, FI denotes the aggregated important feature vector defined in Eq. (7), FS is the salient feature vector defined in Eq. (9), MSI and ASI are the aggregated important salient feature vectors modeled in Eqs. (11) and (12), respectively, and CH is the color histogram feature vector represented by Eq. (14). The symbol ' + ' denotes the concatenation of two feature vectors. The recognition results of objects are illustrated in Fig. 4(a). The abscissa is the quantity of sample images, the ordinate is the accuracy rate (AR). We randomly selected 75–750 samples (5–50 for each object) to train the models and used the remaining samples for testing. The quantity of sample images for each object is denoted as NE in the experiments. When only 75 samples (NE = 5) were used for building the model, FI and FS achieve accuracy rates (ARs) of 90.43% and 91%, respectively, MSI and ASI achieve ARs of 93.84% and 91.23%, respectively, the concatenation of MSI and ASI achieves an AR of 94.88%, and MImSF achieves an AR of 97.25%. Our appearance model achieves the highest AR based on a small quantity of sample images. When the quantity of sample images increases to 375 (NE = 25), the performance improves for all models. The AR of MImSF increases to 99.47%, FI and FS achieve accuracy rates (ARs) of 94.5% and 93.99%, respectively, MSI and ASI achieve ARs of 95.57% and 94.85%, respectively, and the concatenation of MSI and ASI achieves an AR of 97.33%. Generally, from the results of object recognition, important features are slightly better than salient features, the aggregation of important salient features offers a more robust and effective appearance model than either important or salient features alone. The proposed model achieves significantly better results. The confusion matrices of 15 target objects are sketched in Fig. 5.

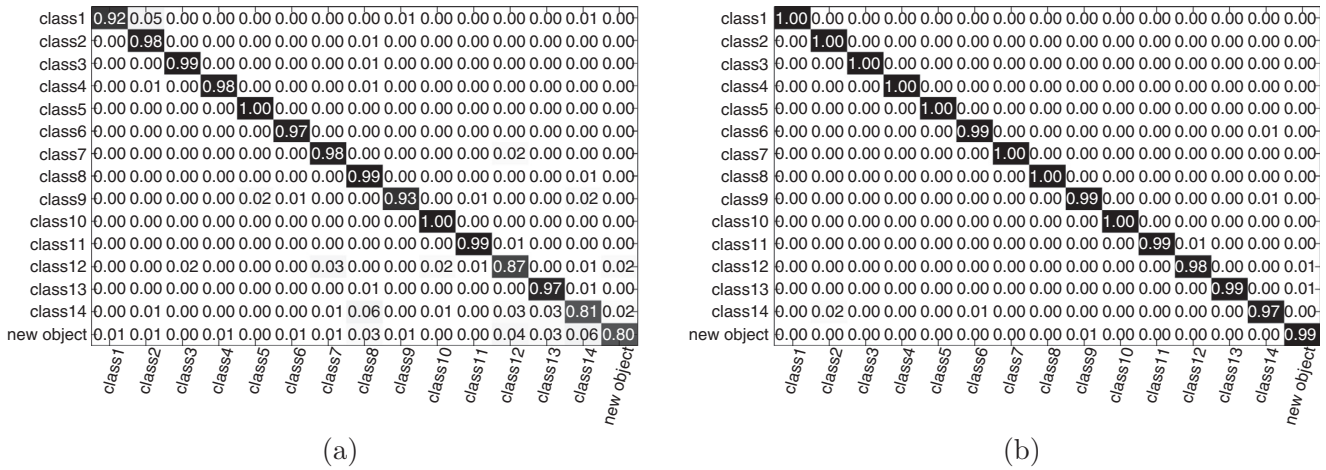
To show the efficiency of our model, we also compare runtime (RT) for appearance modeling using different methods, as shown in Fig. 4(b). The RTs of all of the models increase slowly with the augmentation of the amount of sample images. The RTs of MImSF and MSI + ASI are roughly the same, and the RTs of MSI, ASI, FI, FS are almost the same. When NE = 5, 25, 50, the RTs of MImSF are 1.98 s, 4.01 s and 7.32 s, respectively, while the RTs of FS are 1 s, 2.45 s and 4.56 s, respectively. Our model consumes slightly more time (approximately 1-2 s) than other models, however, it significantly improves the recognition performance.

The results illustrated in Fig. 4 indicate that the performance of our methodology is not significantly affected by the quantity of sample images. As shown in Fig. 4(a), when NE = 20, MImSF has



(a) AR (b) RT

Fig. 4. Comparison of accuracy rate and runtime by using different appearance models.



(a) (b)

Fig. 5. Confusion matrix for recognition of 15 target objects by using different appearance models, 25 sample images for each object: (a) MSI, mean ARs of primary and new objects are 95.57% and 80%, respectively; (b)MImSF, mean ARs of primary and new objects are 99.47% and 99.46%, respectively.

achieved an AR of over 99%. It validates that an accurate and efficient recognition model can be built based on a small set of sample images using our method.

The SIFT features and color features are the most commonly used and useful features. To show the performance of the our method, we compare it with the SIFT-based vocabulary vector model (SIFT-V) [8], and the fusion model of SIFT-V and color histogram feature vector (SIFT-CH) [9]. As illustrated in Fig. 4(a), MImSF achieves significantly better results than SIFT-CH and SIFT-V, especially when the quantity of samples is small.

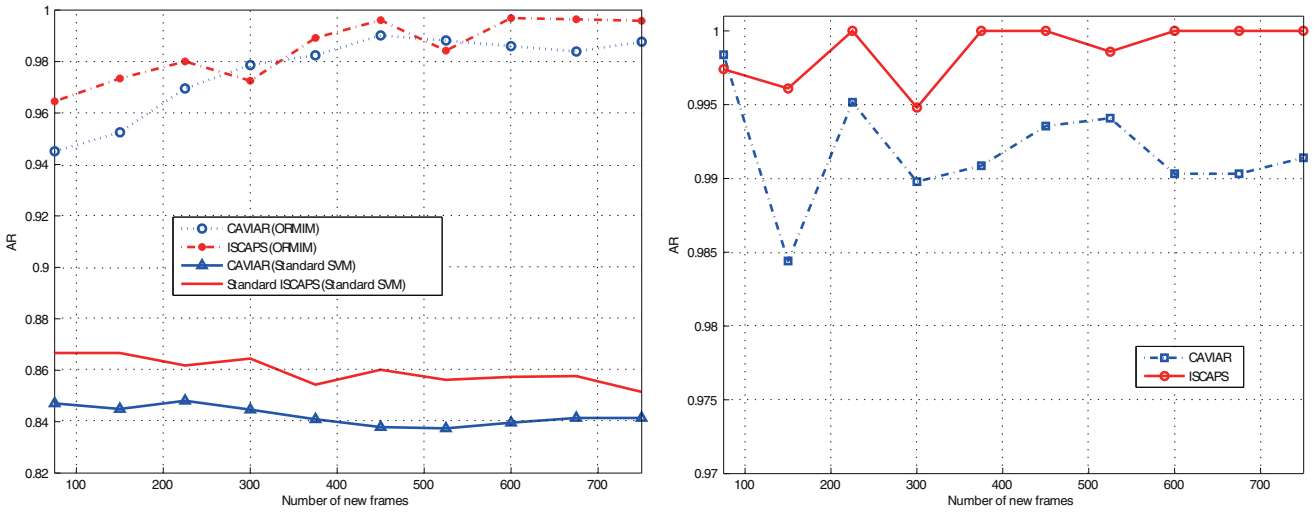
5.3. Performance analysis of the ORMIM

In long distance tracking, when the appearance of objects changes, the recognition performance decreases. To maintain a satisfactory performance, the models need to be updated continuously. When new data are available, the traditional methods accommodate changes using both the new data and previous data to retain and recreate a new model. The previous models are completely discarded. Although the retaining based methods are simple, their time and memory consumption are huge; sometimes, their performances are not very satisfactory.

To validate the performance of the ORMIM and assess the second and third objectives, the following experiments were conducted. First, we validated the proposed model on the two standard datasets. We randomly selected 75-750 samples (NE = 5, 10, ⋯, 50) to train the models. Second, our model was validated using our own datasets. We randomly selected 20-200 samples (NE = 5, 10, ⋯, 50) to train the models. To illustrate the effectiveness of our model, we compare our method with the standard SVM based retraining.

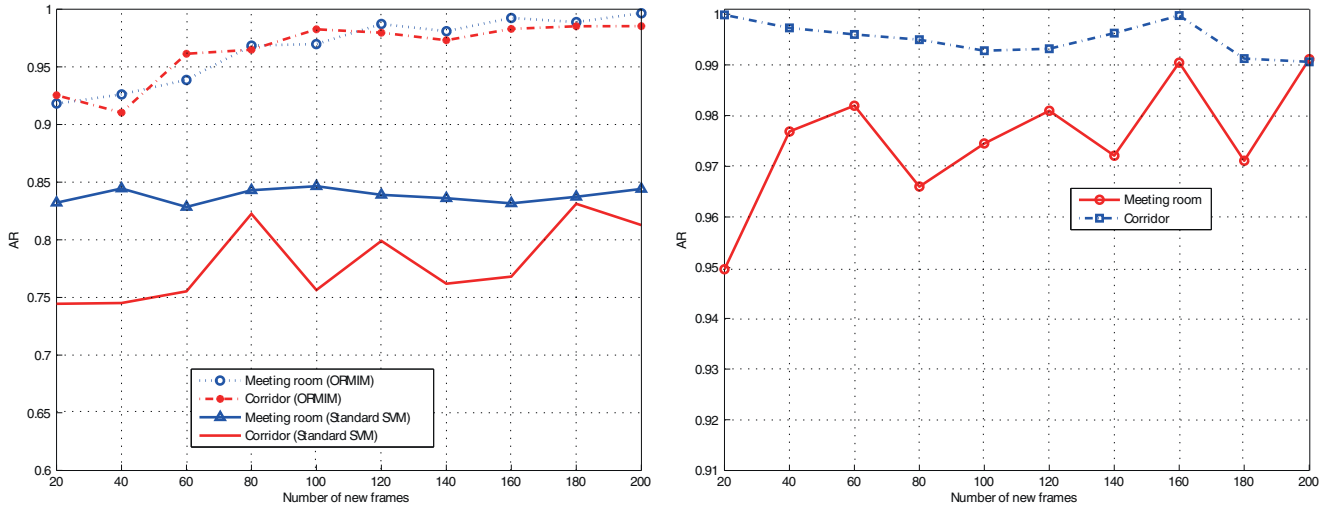
The comparisons of ARs and RTs between ORMIM and standard SVM using standard datasets are shown in Figs. 6 and 8(a). In the CAVIAR dataset, when 450 new frames are available and added to update the model,the AR and RT of ORMIM are 99.01% and 4.98 s, respectively. For standard SVM, the AR and RT are 83.79% and 51.12 s, respectively. In the ISCAPS dataset, when 450 new frames are available and added to update the model,the AR and RT of ORMIM are 99.61% and 5.03 s, respectively. For standard SVM the AR and RT are 86.02% and 53.97 s, respectively. We record the standard deviation (SDs) of our method. In the CAVIAR dataset, when NE = 5, 15, 25, 35 and 50, the SDs are 0.0078, 0.0017, 0.003, 0.0043 and 0.0027, respectively. In our own indoor dataset, when NE = 5, 15, 25, 35 and 50, the SDs are 0.0096, 0.0062, 0.0039, 0.0036 and 0.0029, respectively.

For the recognition of new objects based on ORMIM is shown in Fig. 6(b). Note that the standard SVM cannot recognize the new objects. In the CAVIAR dataset, when only 5 new frames for each



(a) Primary objects (b) New objects

Fig. 6. Comparison of ARs by using different updation models in standard datasets. It demonstrates that the proposed ORMIM is significantly better than the standard SVM in standard datasets.



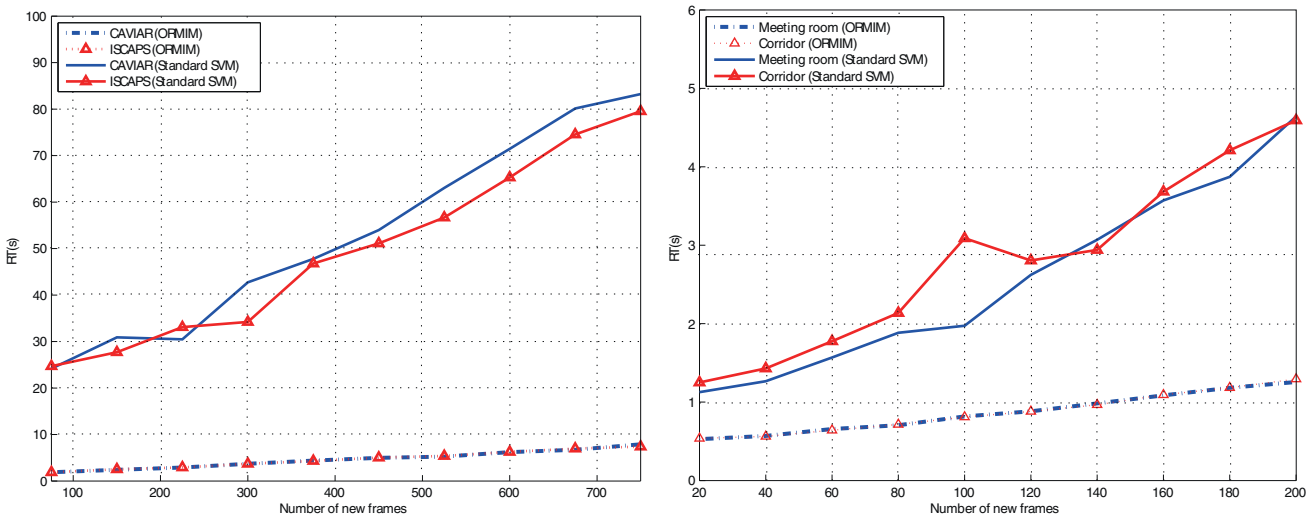
(a) Primary objects (b) New objects

Fig. 7. Comparison of ARs by using different updation models in our own indoor dataset. It demonstrates that the proposed ORMIM is significantly better than the standard SVM in our own dataset.

object are used to update the model, the ORMIM of new object (ORMIM-N) achieves an AR of 99.84%. When NE increases to 15, ORMIM-N achieves an AR of 99.51%. On the whole, the AR of ORMIM-N is slightly changed approximately 99%. In the ISCAPS dataset, when only 5 new frames for each object are used to update the model, ORMIM-N achieves an AR of 99.74%. When NE = 15, ORMIM-N achieves an AR of 99.9958%, which is round up to 100%. To further verify the model performance, we also conduct another experiment on PRID (Person Re-ID) dataset [43]. We selected 20 objects whose NEs are larger than 35. When NE = 5, 15, 25 and 35, the ARs of ORMIN are 97.84%, 98.45%, 98.23% and 98.39%, respectively. When NE = 5, 15, 25 and 35, the ARs of ORMIN-N are 98.13%, 97.85%, 98.23% and 96.70%, respectively.

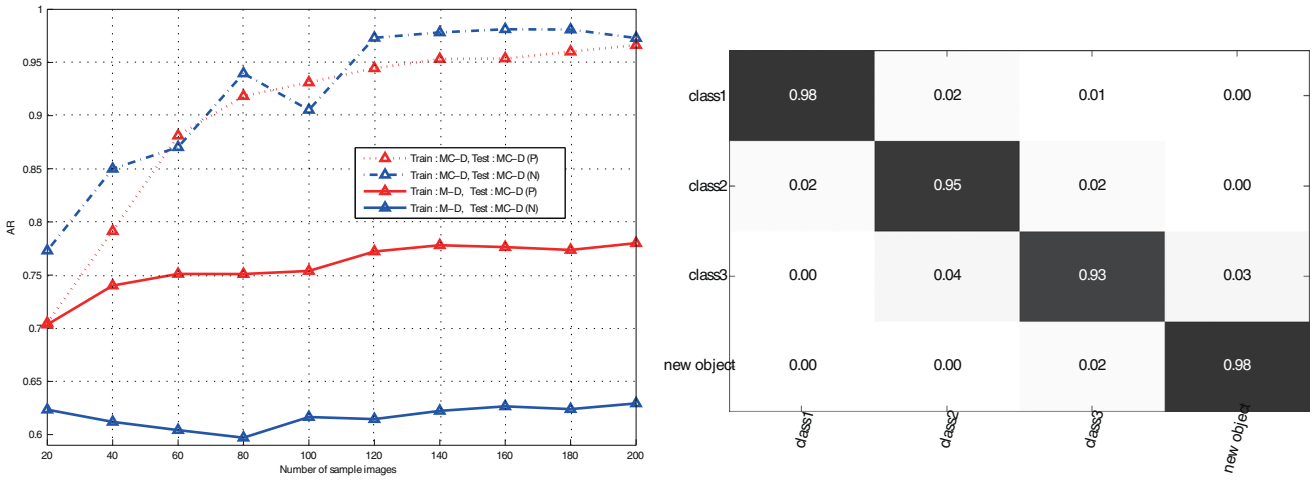
To further validate the efficiency and robustness of the proposed algorithm, we also tested it using our own dataset. The comparisons of ARs and RTs between ORMIM and standard SVM in our indoor dataset are shown in Figs. 7 and 8(b). In the meeting room dataset, when 120 new frames are available and added to update the model, the AR and RT of ORMIM are 98.71% and 0.88 s, respectively. The AR and RT for standard SVM are 83.90% and 2.62 s, respectively. When the number of new frames increases to 200, ORMIM achieves an AR of 99.65%. In the corridor dataset, when 120 new frames are available and added to update the model,the AR and RT of ORMIM are 97.96% and 0.88 s, respectively. The AR and RT for standard SVM are 79.91% and 2.81 s, respectively. When the number of new frames increases to 200, ORMIM achieves an AR of 98.53%. The AR obtained using our own dataset is slightly lower than the standard dataset. Although there are serious occlusion and complicated illumination in our own dataset, our algorithm still achieves a very high AR. Therefore, we also conclude that the proposed method achieves much higher AR and consumes less RT than the standard SVM. In all of the datasets we used, our method achieves significantly high accuracy.

For the recognition of new objects, in the meeting room dataset, when only 5 new frames for each object are used, ORMIM-N achieves an AR of 94.97%, and when NE = 50, ORMIM-N achieves an AR of 99.12%. On the whole, the AR of ORMIM-N is changed approximately 98%. In the ISCAPS dataset, when only 5 new frames



(a) Standard datasets (b) Our own indoor dataset

Fig. 8. Comparison of RTs by using different updation models in different dataset.



(a) AR (b) Confusion matrix, NE=40

Fig. 9. The performance comparison before and after incremental learning when new data is available by using our own indoor dataset.

for each object are used, ORMIM-N achieves an AR of 99.98% and all of the ARs are over 99%.

As illustrated in Figs. 6(b) and 7(b), we also found that the AR of ORMIM-N does not increase with the augmentation of new frames and fluctuates more than that of primary objects. It seems that the recognition results of new objects are more often affected by the sample selection. It also validates that the new target objects can be effectively recognized based on few samples and collecting a large quantity of samples is not necessary for our model.

To show the necessity of incremental learning and further validate the performance of our method, we design and perform another experiments using our own indoor dataset. First, we build a primary model using the meeting room dataset, and then, the new frames including both meeting room data (M-D) and corridor data (C-D) are added. We denote the aggregation of M-D and C-D as MC-D in the following experiments. As a comparison, we also perform the experiments by adding only M-D. The remaining MC-D is used for testing. As shown in Fig. 9, when only M-D is added to update the model, ORMIM achieves an average AR of approximately 75%. When NE = 5, the AR of ORMIM and ORMIM-N are 70.34% and 62.34%, respectively. When NE = 50, ORMIM and ORMIM-N achieve ARs of 78% and 62.93%, respectively. However, when corridor data are also added, i.e., MC-D is added to update the model, the performance of the model significantly improves. When NE = 20, the AR of ORMIM and ORMIM-N are 91.84% and 93.97%, respectively, while NE = 50, ORMIM and ORMIM-N achieve ARs of 96.64% and 97.29%, respectively. It can be seen that the AR of our model becomes significantly higher after new frames (C-D) are used to update the model by using incremental learning. The confusion matrix of three primary objects and one new object is shown in Fig. 9(b).

In addition, to further validate the fourth experimental objective, we also conduct the following experiments using only 1-10 sample images for each object to update the model. The two standard datasets are used. The results are illustrated in Fig. 10. We compare our method with the standard SVM in the experiment. In the CAVIAR dataset, when NE = 1, the AR of ORMIM is 64.12%, and the AR for standard SVM is only 39.98%. When NE = 3, 6, 10, the ARs of ORMIM are 85.33%, 90.12% and 94.81%, respectively, and the ARs for standard SVM are 54.13%, 67.87% and 73.17%. In the IS-

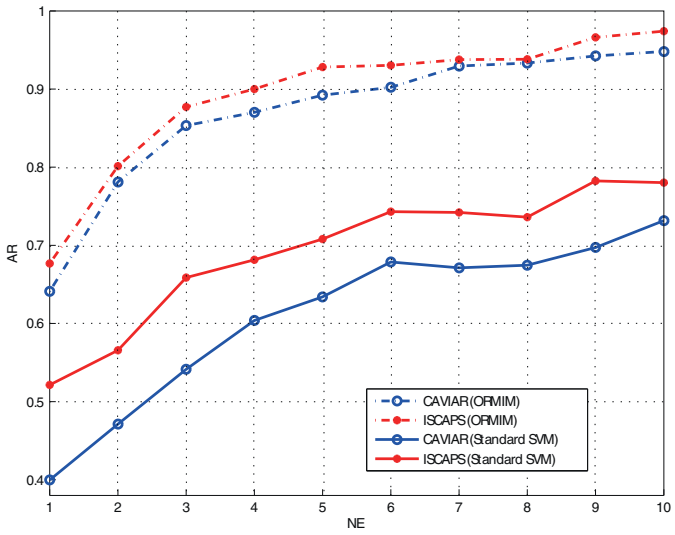


Fig. 10. Comparison of ARs using different updation models in CAVIAR dataset and ISCAPS dataset with only 1–10 sample images for each objects.

Table 1
Comparison of AR by using different recognition models (the CAVIAR dataset).

Recognition model	AR (%)				
	$NE = 5$	$NE = 15$	$NE = 25$	$NE = 35$	$NE = 50$
KCF	64.08	82.47	87.75	90.7	93.02
RF-ML	59.93	46.99	31.51	49.25	82.76
NZ-ML-1(Top-1)	58.1	72.99	77.84	81.19	84.87
NZ-ML-2 (Top-1)	59.81	79.38	85.22	87.19	89.15
NZ-ML-1 (Top-5)	83.20	95.10	97.07	97.39	98.07
NZ-ML-2 (Top-5)	87.48	96.22	97.06	97.71	98.38
LC-KSVD	79.2	83.73	84.73	84.85	98.38
Our model	**97.26**	**98.87**	**99.37**	**99.56**	**99.82**

CAPS dataset, when NE = 1, the AR of ORMIM is 67.68%, and the AR for standard SVM is only 52.13%. When NE = 3, 6, 10, the ARs of ORMIM are 87.71%, 93.04% and 97.41%, respectively, and the ARs for standard SVM are 65.88%, 74.31% and 78.02%. The results indicate that our method can build an accurate and robust model using only a very small quantity of sample images, and achieves a significantly higher AR than the standard SVM.

We also compare several different recognition models in our indoor dataset and outdoor dataset. To compare the proposed method with the RF-ML algorithm [29] and KCF classifier [27], we used MImSF as the input for all these three methods. To compare the proposed method with NZ-ML algorithm [25], we build two models by using SIFT-CH and MImSF as the inputs and denote them as NZ-ML-1 and NZ-ML-2, respectively. We report both the top-1 and top-5 flat error. The flat error is zero if the ground-truth label corresponds to the top-1 label with the highest score (or any of the top-5 labels). The comparison results in our indoor dataset are shown in Table 1, from which it illustrates that the proposed method obtained significantly better results than other methods, especially when NE is small. When only 5 samples for each object is used (NE=5), our method achieves over 18% higher AR than other top-1 methods and over 10% higher AR than the top-5 NZ-ML. When NE is 25, our method still achieves over 10% higher AR than other top-1 methods. The NZ-ML-2 achieves better results than NZ-ML-1, which demonstrates the effectiveness of our proposed appearance model. We found an interesting phenomenon that the performance of our method is slightly better than NZ-ML-2 when NE is larger than 15. However, our method obtain an unique label while NZ-ML-2 gives 5 alternative labels. We also compare the proposed method with LC-KSVD algorithm [35]. For

Table 2
Comparison of AR and SD by using different recognition models (our own outdoor dataset).

Recognition model		LC-KSVD	NZ-ML	KCF	Our model
AR(%)	$NE = 10$	81.85	71.78	38.14	**92.04**
	$NE = 15$	87.45	76.08	61.93	**95.19**
	$NE = 20$	89.61	79.53	81.67	**96.2**
	$NE = 30$	93.19	83.03	93.05	**98.03**
SD(%)	$NE = 10$	2.15	1.7	3.45	**1.37**
	$NE = 15$	1.01	1.02	11.52	**0.49**
	$NE = 20$	1.3	1.1	3.48	**0.82**
	$NE = 30$	0.88	0.77	1.47	**0.55**

LC-KSVD, we perform 25 iterations and select the highest ARs as the final results. From Table 1, we found that our method is superior to LC-KSVD. The comparison results in our outdoor dataset are shown in Table 2. We record the ARs and SDs when NE = 10, 15, 20, 30 in each scene. It shows that our method obtained much higher ARs and smaller SDs than other methods, which demonstrates the effectiveness and robustness of our method. The results further validate the first and third experimental objectives.

6. Conclusion

In this paper, we have presented a new framework based on novel multilevel important salient feature and multicategory incremental learning for object recognition across non-overlapping multicamera views. Our method is built using only appearance information without spatio-temporal reasoning. We have proposed a novel algorithm for appearance modeling called MImSF. Our method uses data-adapting convolution filters to obtain the important feature maps. To improve the robustness of the appearance model, the important feature maps are aggregated with the salient feature maps to produce the multi-level important salient features, and then the important salient features are fused with color feature vector to obtain the final appearance model. To accommodate the appearance change of objects, we have proposed a novel multicategory incremental learning algorithm, ORMIM. The experiments have been conducted on two standard datasets and our own datasets. The experimental results have demonstrated that our method significantly improves the accuracy rate of object recognition, and reduces the time and memory consumption at the same time. In summary, compared with other state-of-the-art classification-based recognition algorithms, the proposed methodology achieves higher ARs and lower RTs when using both standard datasets and our own datasets. Note that, our model can identify new target objects that were never learned in the primary model. In addition, the proposed model can be built and updated using only a small number of sample images. Therefore, our method is very suitable for object recognition across cameras with disjoint views, especially for real-time long distance object tracking. The proposed feature extraction method and incremental learning method can also be directly used in dealing with other related image/video processing and recognition problems.

Acknowledgements

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 61472362, 61379075), National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2014BAK14B01), Zhejiang Provincial Natural Seience Foundation of China (Grant No. LZ16F020002), Zhejiang Provincial public wlfare technology research on society development (2015C33081).

References

[1] A. Rahimi, B. Dunagan, T. Darrell, Simultaneous calibration and tracking with a network of non-overlapping sensors, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, IEEE, 2004, pp. 1–187.

[2] D. Makris, T. Ellis, J. Black, Bridging the gaps between cameras, in: CVPR, 2, IEEE, 2004, pp. 2–205.

[3] J. Omar, S. Khurram, S. Mubarak, Appearance modeling for tracking in multiple non-overlapping cameras, in: CVPR, 2, IEEE, 2005, pp. 26–33.

[4] L. Ma, Y. Xiaokang, T. Dacheng, Person re-identification over camera networks using multi-task distance metric learning, IEEE Trans. Image Process. 23 (8) (2014) 3656–3670, doi:10.1109/TIP.2014.2331755.

[5] A. Bedagkar-Gala, S.K. Shah, A. Bedagkar-Gala, S. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (4) (2014) 270–286, doi:10.1016/j.imavis.2014.02.001.

[6] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identfiation, CVPR, Columbus, USA, 2014.

[7] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, CVPR, Boston, MA, USA, 2015.

[8] L.F. Teixeira, C.-R. Luis, Video object matching across multiple independent views using local descriptors and adaptive learning, Pattern Recognit. Lett. 30 (2) (2009) 157–167.

[9] H.Y. Wang, X. Wang, J. Zheng, J.R. Deller, H.Y. Peng, L.Q. Zhu, W.G. Chen, X.L. Li, R.J. Liu, H.J. Bao, Video object matching across multiple non-overlapping camera views based on multi-feature fusion and incremental learning, Pattern Recognit. 47 (12) (2014) 3841–3851.

[10] X. Wang, Intelligent multi-camera video surveillance: a review, Pattern Recogn. Lett. 34 (1) (2013) 3–19, doi:10.1016/j.patrec.2012.07.005.

[11] B. Prosser, S.G. Gong, X. Tao, Multi-camera matching using bi-directional cumulative brightness transfer functions, in: Proceedings of the British Machine Vision Conference (BMVC), 8, BMVC Press, 2008. 164–1.

[12] C. Madden, E.D. Cheng, M. Piccardi, Tracking people across disjoint camera views by an illumination-tolerant appearance representation, Mach. Vis. Appl. 18 (3-4) (2007) 233–247.

[13] N.S. Sudipta, J.-M. Frahm, M. Pollefeys, Y. Genc, Feature tracking and matching in video using programmable graphics hardware, Mach. Vis. Appl. 22 (1) (2011) 207–217.

[14] M. Aziz, B. Mertsching, Fast and robust generation of feature maps for region-based visual attention, IEEE Trans. Image Process. 17 (5) (2008) 633–644.

[15] T. Kadir, M. Brady, Scale saliency: a novel approach to salient feature and scale selection, in: International Conference Visual Information Engineering, IET, 2003, pp. 25–28.

[16] X.D. Hou, L.Q. Zhang, Saliency detection: a spectral residual approach, in: CVPR, IEEE, 2007, pp. 1–8.

[17] C.L. Guo, Q. Ma, M. Zhang Li, Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform, in: CVPR, IEEE, 2008, pp. 1–8.

[18] M.-M. Cheng, G.-X. Zhang, N.J. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: CVPR, 37, IEEE, 2011, pp. 569–582.

[19] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2–3) (2005) 107–123.

[20] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: European Conference on Computer Vision (ECCV), Springer, 2008, pp. 650–663.

[21] A. Borji, M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, CoRR (2015). abs/1501.02741. URL http://arxiv.org/abs/1501.02741.

[22] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.

[23] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[24] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? metric learning approaches for face identification, in: ICCV, IEEE, 2009, pp. 498–505.

[25] T. Mensink, J. Verbeek, F. Perronnin, G. Csurka, Metric learning for large scale image classification: generalizing to new classes at near-zero cost, in: ECCV, Springer, 2012, pp. 488–501.

[26] P. Jain, B. Kulis, J.V. Davis, I.S. Dhillon, Metric and kernel learning using a linear transformation, J. Mach. Learn. Res. 13 (1) (2012) 519–547.

[27] F. Wang, W. Zuo, L.Q. Zhang, D. Meng, D.J. Zhang, A kernel classification framework for metric learning, 2014, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6932476. [Online; posted 21-October-2014].

[28] Q.Y. Wang, P.C. Yuen, G.C. Feng, Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions, Pattern Recognit. 46 (9) (2013) 2576–2587.

[29] Q. Cao, Y. Ying, P. Li, Similarity metric learning for face recognition, in: ICCV, IEEE, 2013, pp. 2408–2415.

[30] F.F. Li, F. Rob, P. Pietro, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, Comput. Vis. Image Understand. 106 (1) (2007) 59–70.

[31] R. Polikar, L. Upda, S.S. Upda, V. Honavar, Learn++: an incremental learning algorithm for supervised neural networks, IEEE Trans. Syst. Man Cybern. Part C 31 (4) (2001) 497–508.

[32] R. Polikar, J. DePasquale, H.S. Mohammed, G. Brown, L.I. Kuncheva, Learn++.mf: a random subspace approach for the missing feature problem, Pattern Recognit. 43 (11) (2010) 3817–3832.

[33] D.D. Guo, J. Huang, L.F. Chen, Knn model based incre-mental learning algorithm, Pattern Recognit. Artif. Intell. 23 (5) (2010) 701–707.

[34] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141.

[35] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.

[36] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.

[37] F. Orabona, C. Castellini, B. Caputo, L. Jie, G. Sandini, On-line independent support vector machines, Pattern Recognit. 43 (4) (2010) 1402–1412.

[38] K. Boukharouba, L. Bako, S. Lecoeuche, Incremental and decremental multi–category classification by support vector machines, in: Machine Learning and Applications, International Conference on, IEEE, 2009, pp. 294–300.

[39] T. Tommasi, F. Orabona, B. Caputo, Safety in numbers: Learning categories from few examples with multi model knowledge transfer, in: CVPR, IEEE, 2010, pp. 3081–3088.

[40] V. Reddy, C. Sanderson, B.C. Lovell, Improved foreground detection via block--based classifier cascade with probabilistic decision integration, IEEE Trans. Circuits Syst. Video Technol. 23 (1) (2013) 83–93.

[41] CAVIAR: context aware vision using image-based active recognition, 2007, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[42] PETS 2006, 2006, http://www.cvg.rdg.ac.uk/PETS2006/data.html.

[43] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Proceedings Scandinavian Conference on Image Analysis (SCIA), 2011.

Huiyan Wang was born in Yantai, China. She received the M.S. degree in power engineering from Shandong University, Jinan, China and the Ph.D. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 1999 and 2003, respectively. Then she conducted as a postdoctoral research fellow (2003-2005) in clinical medicine from pharmaceutical informatics institute, Zhejiang University, Hangzhou, China. She is currently a professor of Computer Science and Technology in the school of Computer Science and Information Engineering, Zhejiang Gongshang University, China. Her current interests are biomedical signal processing and pattern recognition. She also works on image and video processing.

Yixiang Yan was born in 1990. He is pursuing his Master's degree in computer science and technology in Zhejiang Gongshang University. His interests include video/image processing and pattern recognition.

Jing Hua is a Professor of Computer Science and the founding director of Computer Graphics and Imaging Lab (GIL) and Visualization Lab (VIS) at Computer Science at Wayne State University (WSU). Dr. Hua received his Ph.D. degree (2004) in Computer Science from the State University of New York at Stony Brook. He also received his M.S. degree (1999) in Pattern Recognition and Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences in Beijing, China and his B.S. degree (1996) in Electrical Engineering from the Huazhong University of Science & Technology in Wuhan, China. His research interests include Computer Graphics, Visualization, Image Analysis and Informatics, Computer Vision, etc. He has published over 100 peer-reviewed papers in the above research fields at top journals and conferences, such as IEEE Transactions on Visualization and Computer Graphics, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Visualization, MICCAI, CVPR, ICDM, etc.

Yutao Yang was born in 1990. He is pursuing his Masters degree in computer science and technology in Zhejiang Gongshang University. His interests include video/image processing and pattern recognition.

Xun Wang is currently a professor at the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. He received his BSc in mechanics, MSc and Ph.D. degrees in computer science, all from Zhejiang University, Hangzhou, China, in 1990, 1999 and 2006, respectively. His current research interests include mobile graphics computing, image/video processing, pattern recognition and intelligent information processing. In recent years, He has published over 80 papers in high-quality journals and conferences. He holds 9 authorized invention patents and 5 provincial and ministerial level scientific and technological progress awards. He is a member of the IEEE and ACM, and a senior member of CCF.

Xiaolan Li received her bachelor and master degree in computational mathematics from Xiangtan University in 1998 and 2001, respectively, and Ph.D. degree in signal processing from Peking University in 2006. Her current interests include pattern recognition, image understanding, image processing, 3D modeling, etc.

John Robert Deller is a Fellow of the IEEE and a Professor of Electrical and Computer Engineering at Michigan State University where he directs the Speech Processing Laboratory. Deller holds the Ph.D. (Biomedical Engineering, 1979), M.S. (Electrical and Computer Engineering, 1976), and M.S. (Biomedical Engineering, 1975) degrees from the University of Michigan and the B.S. (Electrical Engineering, Summa Cum Laude, 1974) from the Ohio State University. His research interests include statistical signal processing with applications to bioinformatics, medical diagnostics, speech processing, and communications technologies. He has co-authored two textbooks, is completing a third two-volume text on deterministic and stochastic signal processing. Deller is a recipient of the IEEE Millennium Medal for contributions in signal processing research and education, IEEE Signal Processing Magazine Best Paper Award in 1998, and the IEEE Signal Processing Society's 1997 Meritorious Service Award for his six-year service as Editor-in-Chief of IEEE Signal Processing Magazine.

Guofeng Zhang received the BS and PhD degrees in Computer Science from Zhejiang University, in 2003 and 2009, respectively. He was a recipient of the National Excellent Doctoral Dissertation Award and the Excellent Doctoral Dissertation Award of the China Computer Federation. He is currently an associate professor at State Key Laboratory of CAD&CG, Zhejiang University. His research interests include structure-from-motion, 3D reconstruction, augmented reality, video segmentation and editing. He is a member of IEEE.

Hujun Bao is a professor in the Computer Science Department of Zhejiang University, and the director of the state key laboratory of Computer Aided Design and Computer Graphics. He graduated from Zhejiang University in 1987 with a B.Sc. degree in mathematics, and obtained his Ph.D. degrees in applied mathematics from the same university in 1993.In August 1993, he joined the laboratory. He leads the virtual reality and visual analysis center in the lab, which mainly makes researches on geometry computing, 3D visual computing, real-time rendering, virtual reality and visual analysis. He has published a number of papers over the past few years. These techniques have been successfully integrated into our virtual reality system VisioniX, 3D structure recovery system from videos ACTS, the spatio-temporal information system uniVizal, and the 2D-to-3D video conversion system. His researches are supported by National Natural Science Foundation, the 973 programand the 863 program of China.