Received 8 September 2018
Revised 7 November 2018
Accepted 17 February 2019

Moving objects classification via category-wise two-dimensional principal component analysis

Falah Alsaqre
Independent Academician, Bedford, UK, and
Osama Almathkour
The Kulliyyah of Engineering, IIUM, Kuala Lumpur, Malaysia

Abstract

Classifying moving objects in video sequences has been extensively studied, yet it is still an ongoing problem. In this paper, we propose to solve moving objects classification problem via an extended version of two-dimensional principal component analysis (2DPCA), named as category-wise 2DPCA (CW2DPCA). A key component of the CW2DPCA is to independently construct optimal projection matrices from object-specific training datasets and produce category-wise feature spaces, wherein each feature space uniquely captures the invariant characteristics of the underlying intra-category samples. Consequently, on one hand, CW2DPCA enables early separation among the different object categories and, on the other hand, extracts effective discriminative features for representing both training datasets and test objects samples in the classification model, which is a nearest neighbor classifier. For ease of exposition, we consider human/vehicle classification, although the proposed CW2DPCA-based classification framework can be easily generalized to handle multiple objects classification. The experimental results prove the effectiveness of CW2DPCA features in discriminating between humans and vehicles in two publicly available video datasets.

Keywords Objects classification, Principal component analysis (PCA), Two-dimensional PCA (2DPCA)
Paper type Original Article

1. Introduction

Moving object classification (MOC) in video sequences is an active research area due to its potential in providing more capabilities to wide range of vision-based systems including video surveillance, traffic monitoring and analysis, and security applications. It aims to correctly assign moving objects in dynamic scenes to their respective categories and in turn extract detailed information that helps in understanding objects' behaviors and assessing events within the observed areas of interest. Despite its significant importance, MOC remains

Applied Computing and Informatics
Vol. 18 No. 1/2, 2022
pp. 136-150
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2210-8327
DOI 10.1016/j.aci.2019.02.001

one of the challenging topics in computer vision, especially for outdoor vision-based systems. The primary obstacles of objects classification lie in different inevitable factors such as uncontrollability of outdoor conditions and complexity of background scenes. Furthermore, moving objects that often appear in the field of view are humans, vehicles, motorcycles, and bicycles. As these objects move, their appearances and motions can vary drastically, bringing further difficulties to the classification task.

There have been various existing methods for classifying moving objects in video sequences. One of the important problems in these methods is the differentiation between dynamic objects and background scenes. There are basically two approaches towards this problem. One is to apply a class-specific detector (e.g., human [1], car [2]) at each frame location. However, besides the lack in their generalizability, the application of such detectors is computationally intensive and inadequate for low resolution video sequences. The other is to perform object segmentation prior to classification process. As opposed to applying class-specific detectors, the vast majority of MOC methods assuming stationary camera, and hence benefiting from background modelling techniques, mainly adaptive background subtraction (BS) [3] and Gaussian mixture model (GMM) [4] to segment multiple objects regardless of their types. Note that, these techniques are still not fully satisfactory in terms of performance and accuracy, but, so far, no appropriate alternatives are available [5]. In this work, we follow the latter approach. More specifically, we segment moving objects by means of adaptive BS.

As common to all prior relevant works, after segmenting moving object, the standard scheme for classification consists of firstly performing features extraction to identify descriptors/signatures that properly characterize both individual object and predefined class to which the object belongs. Secondly, each object can be assigned to its most likely category by applying classification model(s). It is natural that the carried-out works vary greatly in terms of targeted moving objects, types and numbers of exploited features, and employed classification models. Besides, the amounts and distributions of processed video sequences differ considerably. As such, there are substantial differences among the reported classification performances, impeding meaningful quantitative evaluation.

In general, available MOC methods can be grouped according to the different features by which the moving objects are described. Regularly used features, individually or in combination, include shape, motion, and texture features.

In shape-based methods, object geometrical properties (dispersedness, silhouette, aspect ratio, area, etc.) are utilized as crucial features for classification. An early example is the work of Lipton et al. [6], which uses dispersedness as a classification metric to discriminate between humans and vehicles. Silhouette-based classification is reported in [7], where a current silhouette is matched with a set of prelabeled silhouettes by distance function. Lin and Wei [8] made use of height/width ratio to specify object category according to predefined thresholds. Often it does not suffice to exploit a single shape feature for classifying different object types. For instance, dispersedness may lead to misclassify human group as a vehicle or vice versa. A straightforward alternative is to use a mixture of shape features. For example, Collins et al. [3] adopted dispersedness, aspect ratio, area, and zoom factor to train a neural network (NN) classifier for categorizing moving objects. The effectiveness of various shape features in conjunction with NN, support vector machine (SVM), and support vector data description (SVDD) are investigated by Hota et al. [9]. However, notwithstanding their simplicity and ease of implementation, shape-based methods unable to accommodate the diverse variabilities in object appearance.

Motion-based methods use temporal information to characterize either entire object or local distinctive patterns. An interesting motion cue is the repetitive movement exhibited by non-rigid articulated objects. In this regard, Lipton [10] considered residual flow as a measure of both rigidity and periodicity of dynamic objects. Cutler and Davis [11] detected and characterized object periodic movement via self-similarity based time-frequency analysis.

In [12], Javed and Shah established recurrent motion image (RMI) to encode the recurrent motion of object parts based on the recovered silhouette changes in consecutive frames. Later, Yogameena et al. [13] and Landabaso et al. [14] followed a similar approach, but replaced silhouette with star skeleton and blob, respectively. Conspicuously, these methods rely heavily upon repetitive motion, and therefore cannot be applied when dynamic objects performing complex and or unconstrained movements. Yet, some methods focused on classifying events into the category of humans or vehicles using classification models (e.g., AdaBoost network [15], Bayesian classifier [16]) built on training data containing labeled trajectories. Definitely, this sort of classification is only applicable to situations where the moving objects tend to generate specific trajectories information.

A group of methods attempts to solve classification problem by benefiting from the functional characteristics of objects' geometrical properties and motion information in a complementary manner. Of these, Zhou and Aggarwal [17] showed that the variances of motion direction yield a good performance in classifying humans and vehicles meanwhile the variances of shape compactness well discriminate human from human group; in doing so, they used K-nearest neighbor (KNN) classifier. Zhaoxiang et al. [18] introduced unsupervised framework to classify human, vehicle, and bicycle, using 5D feature vector formed from shape and motion descriptors (size, compactness, area, velocity, and parameterized angle) coupled with K-means clustering and decision level fusion. Bose and Grimson [19] distinguished between humans and vehicles using a discriminative SVM combined with soft margin and Gaussian kernel. They considered mutual information between candidate features and labeled dataset as scoring criteria to select the informative features and group them into scene-invariant (orientation, variation in area, and percentage occupancy) and scene-specific (image coordinates, motion direction speed, area, and aspect ratio). However, many drawbacks, mainly attributed to instability of objects features among various scenes, limit the application of this group of methods [20].

Texture features have also been exploited for MOC due to their ability to encode various types of visual information within object region. Zhang et al. [21] proposed to apply Adaboost learning algorithm with multi-block local binary pattern and error correcting output code to categorize moving objects. In the work by Liang and Juang [22], local shape features and histograms of orientated gradients (HOGs) are adopted to train hierarchical SVM classifier for differentiating between human, car, motorcycle, and bicycle from their side-view imagery. For more realistic scenarios, texture features are also combined with shape and/or motion features. In this way, Miller et al. [23] classified humans and vehicles with a linear SVM model using 9D feature vector contains eight dimensions of edge histograms and one dimension of aspect ratio. Longbin et al. [24] also tackled human/vehicle classification problem. In their work, object size, location, and velocity are incorporated with the differences between its HOGs calculated in consecutive frames, and the classification task is posed as a Maximum A Posterior problem. Gurwicz et al. [25] considered a broad range of features such as luminance asymmetry, 2D moments, cumulants, and morphological properties. They employed five classification techniques (SVM, NN, Bayesian network, decision tree, and KNN) to classify body organs, human, human group, bag, and clutter. Civelek and Yazici [26] combined speed up robust features (SURF) and shape features (aspect ratio, blob ratio, dispersedness, and compactness) in cascade mode to classify human, human group, and vehicle via KNN classifier. Nevertheless, one of the persistent drawbacks with texture-based methods and their combined use is that they require extensive training datasets, which is impractical. Further, common to all of them is the assumption that both training datasets and testing exemplars are gathered from the same video data, which limits their generalization to new video sequences and classes.

Quite obviously, none of the existing MOC methods can be regarded as a prevailing one or entirely satisfactory. It is therefore necessary to advocate for a different strategy through

which better solution to MOC problems can be attended. Towards this, we propose to leverage the invariant characteristics of object type to remedy as much of the deficiencies as possible. Indeed, an effective approach to capture object invariant characteristics is to use principal component analysis (PCA) [27], which is extensively exploited in face recognition to generate Eigenfaces as compact representations of face images in a lower-dimensional space. In fact, inspired by the Eigenfaces technique, a few attempts have been made to classify one-category object such as pedestrian [28] and vehicle [29]. In this paper, we revisit and extend the application of PCA to multiple objects classification, specifically its 2D version (2DPCA) [30], which, as the name implies, directly deals with 2D images instead of 1D vectorized images. Despite its simplicity, 2DPCA leads to significant improvement in classification performance over traditional PCA, since it, in much lower dimensional representation, effectively preserves structural relation among dataset samples and allows to include more spatial information in produced features.

Most importantly, our application of 2DPCA in MOC differs significantly from its standard application in face recognition, in which 2DPCA is used to map the whole face data (complete set of labeled face images belong to certain individuals/classes) from original space into feature space. In contrast, we utilize 2DPCA in a more generalized manner by applying it independently to object-specific training datasets to generate category-wise feature spaces such that each feature space uniquely captures the invariant characteristics of the underlying object category. That is, by considering each training dataset covers a sufficient range of the object appearance conditions positioned on a uniform background, the retained features convey the most energy of the training samples and some useful local information. Consequently, category-wise 2DPCA (CW2DPCA) not only enables early separation between the different object categories, but also provides effective discriminative representations of both training datasets and test samples to the classification model. In practice, the established classification framework exhibits strong resistant to the variability in objects appearance and, in addition, it is inherently insensitive to objects movement. Note that, the presented framework is explicitly designed to classify moving objects without exploiting tracking information. Also note that, while our MOC method can be used to solve almost any multiclass classification problem, however, the formulations and conducted experiments are confined to human/vehicle classification.

The remainder of this paper is structured as follows. Section 2 reviews the application of PCA and 2DPCA in face recognition. Details of the proposed MOC method are provided in Section 3. Experimental results are presented in Section 4; and finally, Section 5 concludes the paper.

2. Review of PCA and 2DPCA

In this section, we will briefly describe the procedures in PCA-based and 2DPCA-based face recognition methods.

2.1 Principal component analysis

The standard procedure in PCA-based face recognition methods is to represent 2D face images as 1D vectors in image space and project these vectors over a small set of principal components (PCs) to extract the most expressive features while simultaneously prune irrelevant information. And these PCs are indeed the leading eigenvectors of the input images covariance matrix. Let $\boldsymbol{A} = \{\boldsymbol{A}_i\}_{i=1}^{N}$, $\boldsymbol{A}_i \in \mathbb{R}^{p\times q}$, be a training set of N face images exclusively partitioned into classes of individuals, and let $\Lambda = \{\boldsymbol{a}_i\}_{i=1}^{N}$, $\boldsymbol{a}_i \in \mathbb{R}^{pq}$, be a set obtained by vectorizing each image of $\boldsymbol{A}$. Assume $\boldsymbol{P} = \{\boldsymbol{a}_i - \overline{\boldsymbol{a}}\}_{i=1}^{N}$, where $\overline{\boldsymbol{a}}$ denotes the mean vector of Λ, then the covariance matrix of Λ is defined as $\boldsymbol{C} = \boldsymbol{P}\boldsymbol{P}^T \in \mathbb{R}^{pq\times pq}$. Due to the high

dimensionality of image vector space and hence the extreme difficulty in computing $\boldsymbol{C}$, PCA is very often solved via the eigen-decomposition of the matrix $\boldsymbol{L} = \boldsymbol{P}^T\boldsymbol{P} \in \mathbb{R}^{N\times N}$ [31]. Suppose that $\boldsymbol{E} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d] \in \mathbb{R}^{N\times d}$ computed as the eigenvectors of $\boldsymbol{L}$ corresponding to the first d biggest eigenvalues, then $\boldsymbol{X} = \boldsymbol{PE} \in \mathbb{R}^{pq\times d}$ gives the PCs of Λ. In other words, columns of $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{d}$ are the eigenfaces spanning d-dimensional subspace (facespace) of image vector space. Once the facespace is established, the training images and a given test image, $\boldsymbol{a} \in \mathbb{R}^{pq}$, are then projected onto this subspace to produce the weight vectors $\boldsymbol{Z}_\Lambda = \boldsymbol{X}^T\boldsymbol{P} \in \mathbb{R}^{d\times N}$ and $\boldsymbol{Z}_a = \boldsymbol{X}^T(\boldsymbol{a} - \overline{\boldsymbol{a}}) \in \mathbb{R}^d$, respectively. By measuring the similarity between $\boldsymbol{Z}_a$ and each column of $\boldsymbol{Z}_\Lambda$, $\boldsymbol{a}$ can be assigned to its relevant class.

2.2 2D principal component analysis

As opposed to PCA, 2DPCA directly transforms a training set of 2D face images into a set of training feature matrices without vectorization process. Essentially, 2DPCA seeks to construct an optimal projection matrix whose column vectors are the optimal projection axes that maximize total scatter of projected images. In fact, it is proven that these axes are basically the principal eigenvectors of image scatter matrix corresponding to the largest eigenvalues [30]. In particular, consider the set $\boldsymbol{A}$, and let $\overline{\boldsymbol{A}}$ be the mean image of all training samples. The total scatter matrix of $\boldsymbol{A}$ is defined as $\boldsymbol{S} = 1/N\sum_{i=1}^{N}(\boldsymbol{A}_i - \overline{\boldsymbol{A}})^T(\boldsymbol{A}_i - \overline{\boldsymbol{A}}),\ \boldsymbol{S} \in \mathbb{R}^{q\times q}$. It then follows, by application of eigen-decomposition to $\boldsymbol{S}$ matrix, that the eigenvectors associated with the first k eigenvalues of $\boldsymbol{S}$ form the optimal projection matrix $\boldsymbol{U}_{opt} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{q\times k}$. The 2DPCA transformation is then applied to each training image, resulting in a set of training feature matrices $\boldsymbol{Y} = \{\boldsymbol{Y}_i\}_{i=1}^{N}$, where $\boldsymbol{Y}_i = \boldsymbol{A}_i\boldsymbol{U}_{opt} \in \mathbb{R}^{p\times k}$. For a given test image $\boldsymbol{T} \in \mathbb{R}^{p\times q}$, its feature matrix $\boldsymbol{Y}_T = \boldsymbol{TU}_{opt} \in \mathbb{R}^{p\times k}$ is compared to each training feature matrix and is ascribed to the class whose training samples yield highest similarity measure.

3. The proposed method

The framework of the proposed MOC method is illustrated in Figure 1. Given a video sequence, the objects segmentation is first conducted to segregate foreground objects from the background of each observed frame using adaptive BS. As next step, feature extraction is carried out using CW2DPCA. In the training phase, since our focus here is on classifying two categories of objects (humans and vehicles), two disjoint training datasets are used, each comprises a relatively small number of object images per category with diverse object appearance conditions and uniform background. 2DPCA is then applied to each training dataset, leading to construct two category-wise optimal projection matrices. It follows that two sets of training feature matrices are derived, each capturing the underlying invariant characteristics of its respective object. During the testing phase, for each test object segment, two test feature matrices are generated, each by one of the optimal projection matrices. Such representation allows for facilitating the subsequent classification by only returning the minimum distances between each of the test feature matrices and its relevant set of training feature matrices (i.e., obtained by the same optimal projection matrix), thereby the test object image is assigned to its category. To this end, a nearest neighbor classifier based on Euclidean distance is employed.

3.1 Dynamic objects segmentation

Despite the existence of many sophisticated segmentation techniques [32,33], BS still, by far, occupies a prime position in this context because of its algorithmic simplicity and low computational expense. But on the other hand, BS is susceptible to environmental conditions

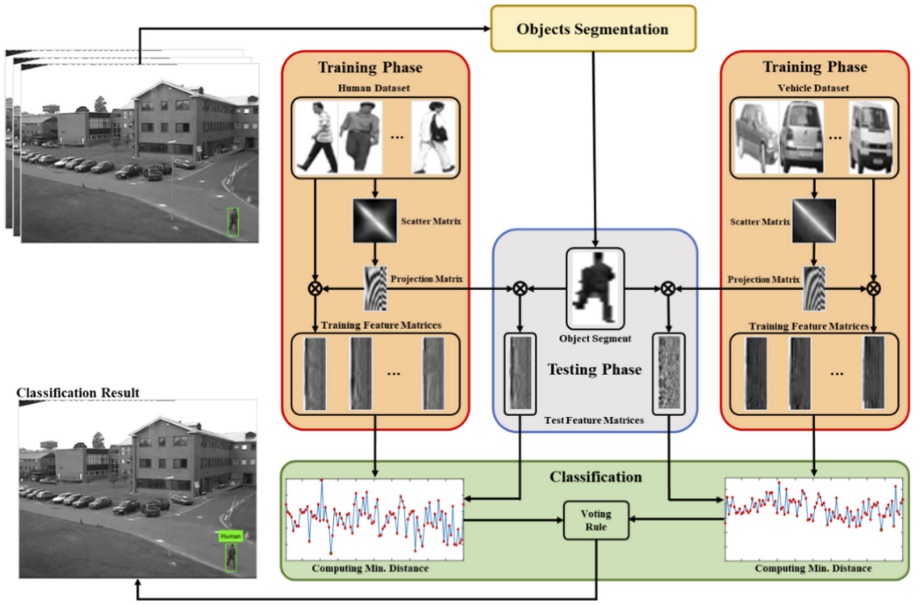

Figure 1.
The framework of proposed MOC method.

and subtle variations in background scene. It also fails when background geometry is modified, for example, a moving object becomes static or vice versa. To circumvent these issues, a common practice involves the employment of adaptation mechanism to learn background scene over time [34]. Thusly, the segmentation procedure entails background scene modelling and maintaining as well as foreground objects detection. With this in mind, we tackle the segmentation problem by means of adaptive BS.

To do so, we first construct a reference background model by averaging a set of successive background frames void of dynamic objects. We then exploit the concept of exponential forgetting [35] to recursively update the background scene as follows:

$$\boldsymbol{B}_n = \alpha \boldsymbol{F}_n + (1-\alpha)\boldsymbol{B}_{n-1}, \tag{1}$$

where $\alpha \in [0, 1]$ is a learning constant, typically set to 0.05, and both $\boldsymbol{B}_n$ and $\boldsymbol{F}_n$ are the background model and observed frame at time instant n, respectively. In order to reveal object-like regions, the difference map between $\boldsymbol{F}_n$ and $\boldsymbol{B}_n$ is calculated and thresholded, resulting in a binary image $\boldsymbol{BI}_n$ as

$$\boldsymbol{BI}_n = \begin{cases} 1 & if\,|\boldsymbol{F}_n - \boldsymbol{B}_n| > \Theta_n \\ 0 & otherwise. \end{cases} \tag{2}$$

Here, Θ_n is a dynamic threshold value, empirically defined as $\Theta_n = 20 + 2.5\sigma(|\boldsymbol{F}_n - \boldsymbol{B}_n|)$, where σ stands for standard deviation. The $\boldsymbol{BI}_n$ image is further processed by standardized binary morphological operations to filter out small noises and to alleviate possible false detections, yielding final mask of moving objects $\boldsymbol{BF}_n$; examples of $\boldsymbol{F}_n$ and $\boldsymbol{BF}_n$ images are given in Figure 2. After that, the foreground objects are segregated by performing pixel-wise multiplication between $\boldsymbol{F}_n$ and $\boldsymbol{BF}_n$. As with most segmentation techniques, we extract a set of segments corresponding to the bounding boxes of moving objects. It is typical that different objects return different segments sizes, so these segments are normalized to fit the

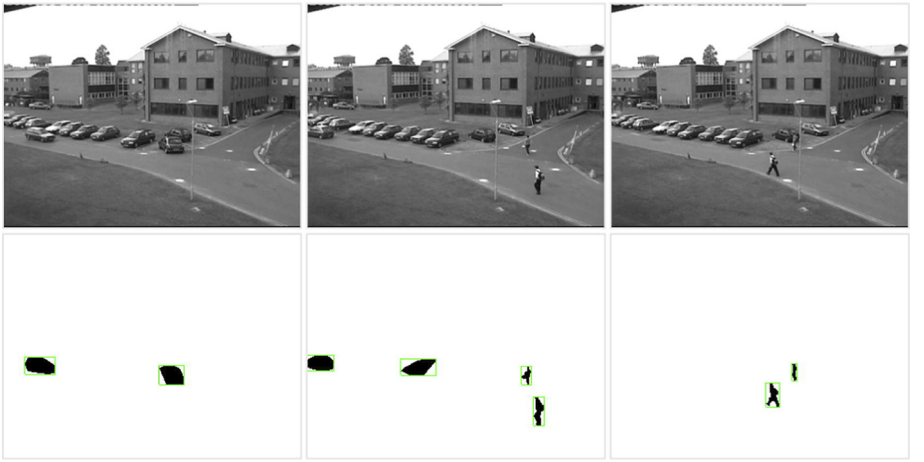

Figure 2.
(Top) Three frames from PETS2001 video sequence [36]. (Bottom) Resulted moving objects masks after performing adaptive BS and morphological operations.

height and width of training samples. Therefore, at each n, the segmentation result is a set, denoted as $\boldsymbol{O}_n = \{\boldsymbol{Q}_{nr}\}_{r=1}^{w}$, $\boldsymbol{Q}_{nr} \in \mathbb{R}^{p \times q}$, composed of w equal size objects segments have uniform background. And of course, when no object exists in the scene, $\boldsymbol{O}_n = \{\phi\}$.

3.2 Category-Wise 2DPCA

Classical 2DPCA, by nature, deals with a single-object classification, where the training samples are partitioned into a number of pattern classes, so that all classes are equally contributed to the computation of total scatter matrix. Unfortunately, this is not the case in multiple objects scenario, because directly applying 2DPCA to training dataset containing samples from more than one object type leads to potential ambiguity in the definition of total scatter, since different object types have different structural and spatial properties. Alternatively, we introduce CW2DPCA for multiple objects' classification, in which the scatter matrices are defined from the category perspective to clearly quantify the typical correlation between intra-category samples and, in turn, ensure that each object category is uniquely characterized by its own invariant features. Therefore, the main purpose of CW2DPCA is to transform each object-specific dataset from image space to feature space, through which the preserved information in transformed space is well discriminative while being separable prior to the classification task. Without loss of generality, we address a simplified two-object classification problem. Specifically, given a segmented object in a video sequence, the goal is to coarsely categorize it into either human or vehicle. We hence use two disjoint training datasets, one constructed from human images and the other from vehicle images, to learn optimal projection for each object separately.

Let $\boldsymbol{H} = \{\boldsymbol{H}_i\}_{i=1}^{N}$, $\boldsymbol{H}_i \in \mathbb{R}^{p \times q}$, and $\boldsymbol{V} = \{\boldsymbol{V}_i\}_{i=1}^{N}$, $\boldsymbol{V}_i \in \mathbb{R}^{p \times q}$, be the human and vehicle training datasets, respectively, where N is the number of samples in each dataset. Since, here, the objective is to obtain two category-wise optimal projection matrices, each of which maximizes the scatter within its respective intra-category samples, the scatter matrices $\boldsymbol{S}^H$ and $\boldsymbol{S}^V$ of $\boldsymbol{H}$ and $\boldsymbol{V}$, respectively, are individually computed as

$$\boldsymbol{S}^H = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{H}_i - \overline{\boldsymbol{H}})^T(\boldsymbol{H}_i - \overline{\boldsymbol{H}}), \tag{3}$$

$$\boldsymbol{S}^V = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{V}_i - \overline{\boldsymbol{V}})^T(\boldsymbol{V}_i - \overline{\boldsymbol{V}}) \tag{4}$$

Here, $\overline{\boldsymbol{H}}$ and $\overline{\boldsymbol{V}}$ denote the mean images of $\boldsymbol{H}$ and $\boldsymbol{V}$, respectively, and both $\boldsymbol{S}_H$ and $\boldsymbol{S}_V \in \mathbb{R}^{q\times q}$. By computing the eigen-decomposition of each scatter matrix, we can form two optimal projection matrices $\boldsymbol{U}^H_{opt}$ and $\boldsymbol{U}^V_{opt}$ such that $\boldsymbol{U}^H_{opt} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k] \in \mathbb{R}^{q\times k}$ and $\boldsymbol{U}^V_{opt} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{q\times k}$, where $\{\boldsymbol{h}_i | i = 1, \ldots, k\}$ and $\{\boldsymbol{v}_i | i = 1, \ldots, k\}$ are, respectively, sets of the first k dominant eigenvectors of $\boldsymbol{S}^H$ and $\boldsymbol{S}^V$. So, accordingly, we derive two distinct sets of feature matrices denoted by $\boldsymbol{Y}^H$ and $\boldsymbol{Y}^V$ for the training images of $\boldsymbol{H}$ and $\boldsymbol{V}$, respectively, as follows:

$$\boldsymbol{Y}^H = \left\{\boldsymbol{Y}^H_i\right\}_{i=1} \text{ with } \boldsymbol{Y}^H_i = \boldsymbol{H}_i\boldsymbol{U}^H_{opt} \in \mathbb{R}^{p\times k}, \tag{5}$$

$$\boldsymbol{Y}^V = \left\{\boldsymbol{Y}^V_i\right\}_{i=1} \text{ with } \boldsymbol{Y}^V_i = \boldsymbol{V}_i\boldsymbol{U}^V_{opt} \in \mathbb{R}^{p\times k} \tag{6}$$

As we have already defined two feature spaces (one per category), it then becomes natural to base classification task on the fact that feature representation of a test sample from a certain category lies close only to that of training data from the same category. To proceed with this, we first need to map the test object image onto each feature space, and then by using NN classifier, the object membership can be specified, as will be detailed in subsequent section.

3.3 Dynamic objects classification

Assume that $\boldsymbol{O}_{nr}$ is a test object image segmented at time instant n, and that $\boldsymbol{O}^H_{nr} \in \mathbb{R}^{p\times k}$ and $\boldsymbol{O}^V_{nr} \in \mathbb{R}^{p\times k}$ are two feature matrices for $\boldsymbol{O}_{nr}$, where $\boldsymbol{O}^H_{nr} = \boldsymbol{O}_{nr}\boldsymbol{U}^H_{opt}$ and $\boldsymbol{O}^V_{nr} = \boldsymbol{O}_{nr}\boldsymbol{U}^V_{opt}$. On the other words, $\boldsymbol{O}^H_{nr}$ and $\boldsymbol{O}^V_{nr}$ are the projections of $\boldsymbol{O}_{nr}$ into the feature spaces defined by $\boldsymbol{U}^H_{opt}$ and $\boldsymbol{U}^V_{opt}$, respectively. Thus, the evidence for $\boldsymbol{O}_{nr}$ being belonged to either of the two categories is simply the highest similarity, within each feature space, between $\boldsymbol{O}_{nr}$ feature matrix and the set of training feature matrices. Towards this end, we make use of NN classifier based on Euclidean distance in feature space.

More formally, we first compute the minimum distance between $\boldsymbol{O}^H_{nr}$ and $\boldsymbol{Y}^H$ as well as between $\boldsymbol{O}^V_{nr}$ and $\boldsymbol{Y}^V$ as below:

$$D^H_{nr} = \min_i \left\| \boldsymbol{O}^H_{nr} - \boldsymbol{Y}^H_i \right\|, \quad i = 1, \cdots, N \tag{7}$$

$$D^V_{nr} = \min_i \left\| \boldsymbol{O}^V_{nr} - \boldsymbol{Y}^V_i \right\|, \quad i = 1, \cdots, N \tag{8}$$

where $\|\cdot\|$ refers to the Euclidean norm. It follows by returning $D_{nr} = \min(D^H_{nr}, D^V_{nr})$ that the category index associated with the assigned minimum distance to D_{nr} is the category membership of $\boldsymbol{O}_{nr}$.

4. Experimental results

To evaluate the effectiveness of the proposed MOC method, we conducted experiments on two publicly available video datasets: PETS2000 and PETS2001 [36]. It is important to point out that we can only compare the results of the presented CW2DPCA method against those provided by category-wise PCA (CWPCA) methods, which, however, not exist as yet in literature. For fairness, we have compared our method with an extended version of the PCA-based vehicle classification framework introduced in [29], which initially structured to

classify vehicles at finer-level, and only a set of training vehicle images was employed to generate the PCs (eigenvehicles). We extended it to the CWPCA by following the similar procedure as in CW2DPCA with exception that the formulation is made on vectorized version of input images.

Before giving the classification results, we will first introduce the used training datasets for constructing category-wise feature spaces, then we will illustrate the role of feature extraction within both CW2DPCA and CWPCA.

4.1 Training datasets

In order to define the bases of the category-wise feature spaces, we constructed two separate training datasets, one comprising human samples and the other of vehicle samples, each one 200 samples long, spanning a sufficient range of object appearance conditions. The human and vehicle samples are manually segmented from the images of Penn-Fudan database [37] and Graz-02 dataset [38], respectively. The background intensity and size of each sample are respectively set to 255 and 50×30 pixels. Figure 3 shows five samples from each dataset.

4.2 The role of feature extraction

We applied CW2DPCA to transform input images into feature matrices. Here, according to the CW2DPCA formulation, the size of scatter matrices $\boldsymbol{S}^H$ and $\boldsymbol{S}^V$ is 30×30, and consequently it is quite easy to form the optimal projection matrices $\boldsymbol{U}^H_{opt}$ and $\boldsymbol{U}^V_{opt}$ (and hence the feature matrices). For instance, when considering $k=15$, both $\boldsymbol{U}^H_{opt}$ and $\boldsymbol{U}^V_{opt} \in \mathbb{R}^{30\times15}$, and the two sets of training feature matrices receive the forms $\boldsymbol{Y}^H = \{\boldsymbol{Y}^H_i\}^{200}_{i=1}$ and $\boldsymbol{Y}^V = \{\boldsymbol{Y}^V_i\}^{200}_{i=1}$, where $\boldsymbol{Y}^H_i$and $\boldsymbol{Y}^V_i \in \mathbb{R}^{50\times15}$, whereas the projected test feature matrices $\boldsymbol{O}^H_{nr}$ and $\boldsymbol{O}^V_{nr} \in \mathbb{R}^{50\times15}$. Although the feature matrices are relatively compact, they convey the most energy of the original images [30] and preserve some local details which may useful in distinguishing between different objects. That is to say, these feature matrices provide

Figure 3.
First row: Five samples from human dataset. Second row: Five samples from vehicle dataset.

compact and meaningful descriptions to the content of input images while performing classification. As evidence of this, Figure 4 depicts the reconstructed images of the first sample in each row of Figure 3, when $k=1, 3, 6, 9, 12$, and 15. One can observe that the first few principal eigenvectors are sufficient enough to produce a good approximation to the original samples. For comparison, Figure 4 also depicts the reconstructed images by CWPCA (eigenhumans and eigenvehicles) as the number of PCs d set to 10, 20, 30, 40, 50, and 60. As expected, CWPCA yields much lower reconstruction quality compared to CW2DPCA.

4.3 Experiments on PETS2000 and PETS2001 video datasets

PETS2000 and PETS2001 are surveillance type of video sequences containing video objects with diverse appearance conditions and motion patterns. These sequences also have some challenging factors such as illumination variations, complex background, and background modifications. PETS2000 consists of 1452 frames with resolution of 480×640 pixels (height $\times$ width), containing only humans and vehicles. Although PETS2001 composed of 3064 frames with size 576×768 pixels, we considered only the first 2550 frames in which the moving objects are solely humans and vehicles. In our experiments, the original frames of PETS2000 and PETS2001 are converted to gray scale and normalized to 240×320 pixels.

Again, in this paper, CW2DPCA and CWPCA methods are used for feature extraction to distinguish between humans and vehicles. Note that, since in general the number of principal eigenvectors/components to be retained is user-defined, we selected values ranging from 1 to 15 in an incremental manner, so that for each method, 15 test runs have been performed on each video sequence.

In the segmentation procedure, we picked up the first 30 and 170 frames to infer the background scene of PETS2000 and PETS2001, respectively, and then updated using (1). It is worthwhile mentioning that invalid objects segments are excluded from the subsequent classifications. These segments mostly correspond to poorly/erroneously segmented objects and to dynamically occluded objects. Unfortunately, such segmentation results are almost inevitable unless there are user interactions, which, out of scope of this paper. Resultantly, at the end of each test run, the total number of valid segmented objects within PETS2000 is 2028, of which 1302 are humans and 726 vehicles, whereas within PETS2001, it is 2764, of which 1983 are humans and 781 vehicles. Notice, furthermore, that the background intensity and size of each segment are set identically to those of training samples.

In the following, we demonstrate the performance of CW2DPCA and compare it to CWPCA. Figures 5 and 6 separately display some examples of correctly classified humans and vehicles in PETS2000 and PETS2001 by CW2DPCA method.

Figures 7 and 8 show the classification accuracies for the moving objects within PETS2000 and PETS2001, respectively, when using CW2DPCA and CWPCA. As observed in Figure 7, the lowest classification accuracies produced by CW2DPCA are 84.10% for humans, 91.18% for vehicles, and 86.83% for total objects (humans and vehicles)

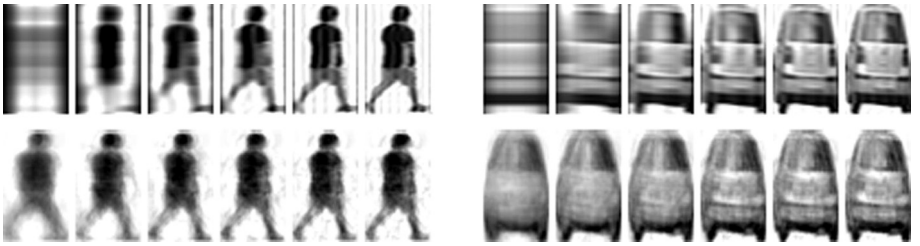

Figure 4.
First row: Two groups of reconstructed images by CW2DPCA; in each group, from left to right: $k=1$; $k=3$; $k=6$; $k=9$; $k=12$; $k=15$. Second row: Two groups of reconstructed images by CWPCA, in each group, from left to right: $d=10$; $d=20$; $d=30$; $d=40$; $d=50$; $d=60$.

when $k=1$, 2, and 1, respectively. Figure 8 also shows that the lowest classification accuracies yielded by CW2DPCA are 78.67% for humans, 91.55% for vehicles, and 82.31% for total objects when k=1. Both of these results indicating that CW2DPCA with a few principal eigenvectors has a strong ability to classify moving objects. Furthermore, it is observed in Figure 7 that the proposed method reached highest classification accuracies of 92.70% for humans, 96.69% for vehicles, and 93.78% for total objects when $k=5$, 8, and 7, respectively. Also, according to Figure 8, the highest classification accuracies achieved with CW2DPCA are 94.35% for humans, 97.95% for vehicles, and 94.79% for total objects when $k=5$, 6, and 5, respectively. Such results further affirming the effectiveness of the proposed method for moving objects classification in these challenging video sequences.

As also noted in Figures 7 and 8, CW2DPCA consistently outperforms CWPCA for each moving object. In Table 1, we report the performance of CW2DPCA and CWPCA methods in terms of average classification accuracy. Results from Table 1 show that CW2DPCA achieves average classification accuracies surpass those of CWPCA by 10% to 14%.

Although the presented method outperforms CWPCA method, mostly benefiting from the efficient representation of original images by CW2DPCA, but not surprisingly their results share some general trends. Particularly, the classification accuracies of both methods tend to increase as the number of principal eigenvectors/components increases. As expected intuitively, the classification accuracies for vehicles are always higher than those for humans. This is fundamentally due to the fact that humans are nonrigid highly deformable objects often appear relatively small within video frames, so the segmentation results may not return their accurate structures. Further, apart from the segmentation performance, the misclassification cases are occurred when the moving objects appear too small to support sufficient features and, in less degree, when their appearances are not well covered in the datasets.

Eventually, we evaluated the computational efficiency of CW2DPCA and CWPCA methods, with unoptimized MATLAB code runs on a laptop with Intel Core i3, 2.26 GHZ CPU, and 4GB RAM. Table 2 provides runtime for individual phases of each method in some of the conducted experiments; more specifically, when both k and d set to 1, 5, 10, and 15. It can be noted that the training times in both methods are very short, since the sample size and

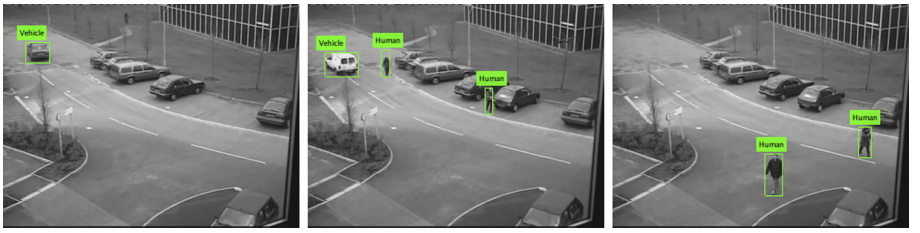

Figure 5.
Examples of correctly classified moving objects in PETS2000.

Figure 6.
Examples of correctly classified moving objects in PETS2001.

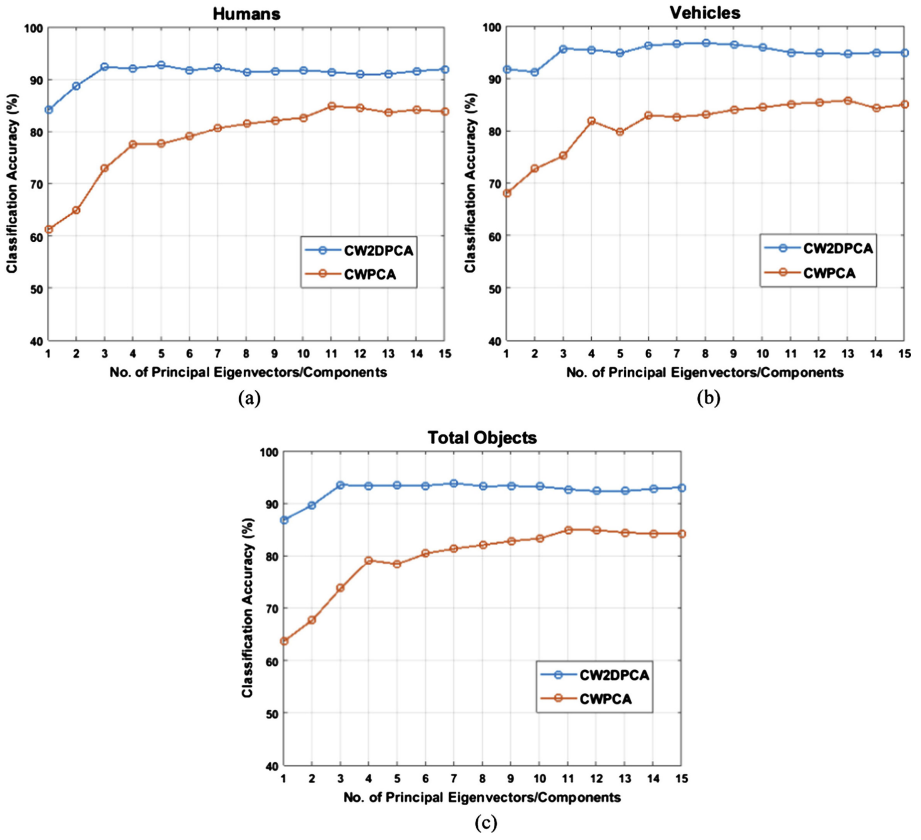

Figure 7. Classification accuracies of CW2DPCA and CWPCA for (a) humans, (b) vehicles, and (c) total objects within PETS2000.

dimension of training datasets are relatively small. Even so, CWPCA training times are slightly longer compared to CW2DPCA. As for segmentation and testing times, CWPCA method also takes slightly longer times than the CW2DPCA method. Table 2 also clearly shows that the computational time of each individual phase increases as the number of principal eigenvectors/components increases. Moreover, both CW2DPCA and CWPCA methods in their primal forms are able to achieve 8 to 10 fps.

5. Conclusions

In this paper, we have proposed CW2DPCA-based framework for classifying dynamic objects in video sequences. The basic idea of CW2DPCA is to construct category-wise optimal projection matrices from object-specific training datasets, and then derive feature space for each object category. As a result, CW2DPCA ensures early separation between different object categories and meanwhile produces compact and discriminative features to characterize training datasets and test objects samples. Unlike other methods, our classification framework able to accommodate the variability in objects appearance by the virtue of CW2DPCA, and it is inherently insensitive to objects' motion patterns. The experimental results on two challenging video sequences confirm the performance of the presented framework. Although we have addressed human/vehicle classification

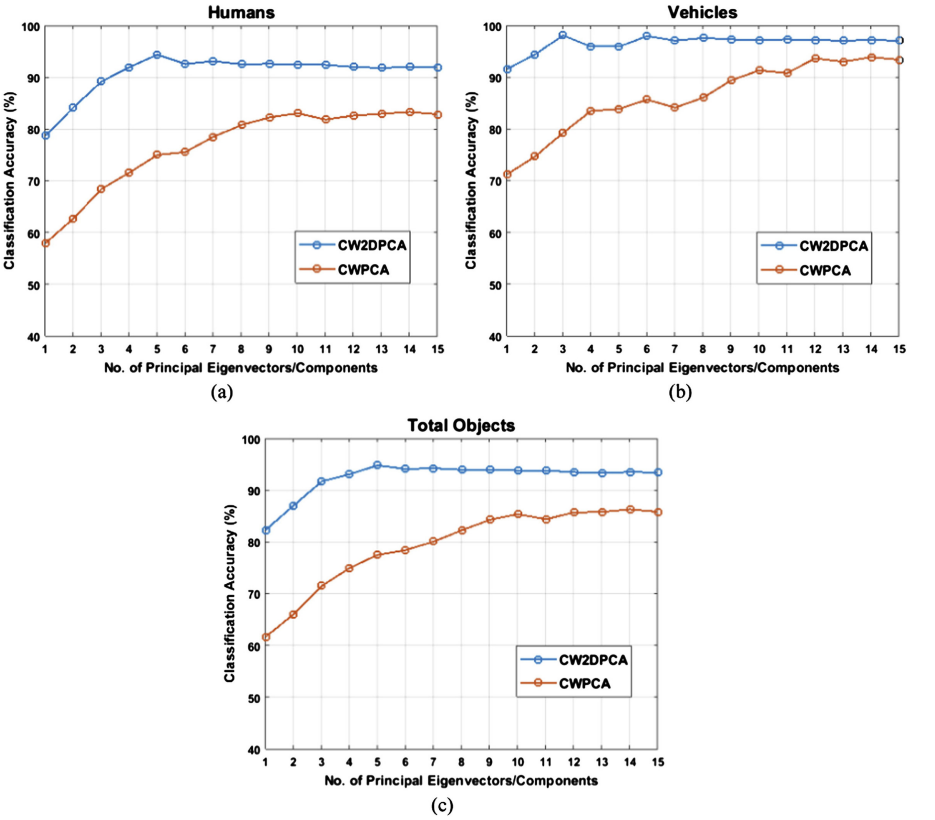

Figure 8.
Classification accuracies of CW2DPCA and CWPCA for (a) humans, (b) vehicles, and (c) total objects within PETS2001.

Table 1.
Comparison of the average classification accuracy (%) of CW2DPCA versus CWPCA.

	CW2DPCA			CWPCA		
Sequence	Humans	Vehicles	Total Objects	Humans	Vehicles	Total Objects
PETS2000	91.02	94.97	92.43	78.75	81.35	79.68
PETS2001	90.77	96.58	92.41	76.61	86.23	79.33

Table 2.
Computational time (s) required for each phase of CW2DPCA and CWPCA methods.

	Training Time		Segmentation and Testing Time			
			PETS2000 (1422 frames)		PETS2001 (2380 frames)	
$k=d$	CW2DPCA	CWPCA	CW2DPCA	CWPCA	CW2DPCA	CWPCA
1	0.1128	0.1847	164.5231	167.8043	256.6595	265.3945
5	0.1183	0.1920	165.3460	169.4325	261.2318	270.1269
10	0.1251	0.1975	167.6328	172.4360	266.9264	276.0366
15	0.1302	0.2050	171.9067	177.0871	272.7050	284.0124

in this paper, it is straightforward to extend CW2DPCA to handle multiple objects classification.

References

[1] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005, vol. I, pp. 886–893, 2005.

[2] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, Cvpr (2000) 746–751.

[3] R.T. Collins et al., A system for video surveillance and monitoring, VSAM Final Rep., pp. 1–68, 2000.

[4] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proc. 1999 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Cat No PR00149, vol. 2, no. c, pp. 246–252, 1999.

[5] S. Palazzo, C. Spampinato, D. Giordano. Gamifying Video Object . Segmentation 39 (10) (2016) 1942–1958.

[6] AJ. Lipton, H. Fujiyoshi, RS. Patil, Moving target classification and tracking from real-time video, Proc. Fourth IEEE Work. Appl. Comput. Vision. WACV'98 (Cat. No.98EX201) 98 (2) (1998) 8–14.

[7] B.U. Töreyin, U. Güdükbay, a.E. Çetin, Silhouette-Based Method for Object Classification and Human Action Recognition in Video, Eccv 3979 (2006) 64–77.

[8] H.-Y. L. H.-Y. Lin, J.-Y. W. J.-Y. Wei, A Street Scene Surveillance System for Moving Object Detection, Tracking and Classification, in: 2007 IEEE Intell. Veh. Symp., pp. 1077–1082, 2007.

[9] R.N. Hota, V. Venkoparao, A. Rajagopal, Shape Based Object Classification for Automated Video Surveillance with Feature Selection, in: 10th Int. Conf. Inf. Technol. (ICIT 2007), pp. 97–99, 2007.

[10] A.J. Lipton, Local Application of Optic Flow to Analyse Rigid versus Non-Rigid Motion, Int. Conf. Comput. Vis. Work. Fram. Appl. (1999).

[11] R. Cutler, L.S. Davis, Robust real-time periodic motion detection, analysis, and applications. . IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 781–796.

[12] O. Javed, M. Shah, Tracking and Object Classification for Automated Surveillance, Eur. Conf. Comput. Vis. (2002) 343–357.

[13] R. Jyothi Priya, S.M. Mansoor Roomi, B. Yogameena, V. Abhaikumar, S. Raju, People/vehicle classification by recurrent motion of skeleton features, IET Comput. Vis. 6 (5) (2012) 442–450.

[14] J.L. Landabaso, L.Q. Xu, M. Pardas, Robust tracking and object classification towards automated video surveillance, in: Image Anal. Recognition, Pt 2, Proc., vol. 3212, pp. 463–470, 2004.

[15] J.P. Renno, D. Makris, G.A. Jones, Object Classification in Visual Surveillance Using Adaboost, in: 2007 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8, 2007.

[16] P. Remagnino, G.A. Jones, Classifying Surveillance Events from Attributes and Behaviour, . Procedings Br. Mach. Vis. Conf. (2001), p. 70.1-70.10, 2001.

[17] Q. Zhou, J.K. Aggarwal, Tracking and classifying moving objects from video, Proc. IEEE Work. Perform. Eval. Track. Surveill. 12 (2001).

[18] Z. Zhaoxiang, C. Yinghao, H. Kaiqi, T. Tieniu, Real-time moving object classification with automatic scene division, Proc. - Int. Conf. Image Process. ICIP 5 (2007) 149–152.

[19] B. Bose, E. Grimson, Improving object classification in far-field video, Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit 2 (2004) pp. II–II.

[20] S. Boragno, B. Boghossian, D. Makris, S. Velastin, Object classification for real-time video-surveillance applications, in: 5th Int. Conf. Vis. Inf. Eng. (VIE 2008), no. 1, pp. 192–197, 2008.

[21] L. Zhang, S.Z. Li, X. Yuan, S. Xiang, Real-time Object Classification in Video Surveillance Based on Appearance Learning, in: 2007 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8, 2007.

[22] C.W. Liang, C.F. Juang, Moving object classification using local shape and HOG features in wavelet-transformed space with hierarchical SVM classifiers, Appl. Soft Comput. J. 28 (2015) 483–497.

[23] A. Miller, A. Basharat, B. White, J. Liu, M. Shah, Person and vehicle tracking in surveillance video, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4625 LNCS, pp. 174–178, 2008.

[24] C. Longbin, R. Feris, Z. Yun., L. Brown, A. Hampapur, An integrated system for moving object classification in surveillance videos, in: Proc. - IEEE 5th Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2008, pp. 52–59, 2008.

[25] Y. Gurwicz, R. Yehezkel, B. Lachover, Multiclass object classification for real-time video surveillance systems, Pattern Recognit. Lett. 32 (6) (2011) 805–815.

[26] M. Civelek, A. Yazici, Object Extraction and Classification in Video Surveillance Applications, Eur. Rev. 25 (2) (2017) 246–259.

[27] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[28] S. Munder, D.M. Gavrila, An Experimental Study on Pedestrian Classification, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1863–1868.

[29] C. Zhang, X. Chen, W. Chen, A PCA-Based Vehicle Classification Framework, in: 22nd International Conference on Data Engineering Workshops, (ICDEW'06), 2006, pp. 17–26.

[30] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (1) (2004) 131–137.

[31] M. Turk, A. Pentland, Eigenfaces for Recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.

[32] B. Luo, H. Li, F. Meng, Q. Wu, C. Huang, Video Object Segmentation via Global Consistency Aware Query Strategy, IEEE Trans. Multimed. 19 (7) (2017) 1482–1493.

[33] Y. Zhang, X. Chen, J. Li, W. Teng, H. Song, Exploring Weakly Labeled Images for Video Object Segmentation with Submodular Proposal Selection, IEEE Trans. Image Process. 27 (9) (2018) 4245–4259.

[34] A. Sobral, A. Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Comput. Vis. Image Underst. 122 (2014) 4–21.

[35] A Elgammal, Figure-ground segmentation - pixel-based, Vis. Anal. Humans (2011) 31–51.

[36] P. Datasets, No Title. [Online]. Available: http://www.cvg.reading.ac.uk/datasets/.

[37] L. Sun, X. Liang, Q. Zhao, Recursive templates segmentation and exemplars matching for human parsing, Comput. J. 57 (3) (2014) 364–377.

[38] A. Opelt, A. Pinz, The TU Graz-02 database, 2002. [Online]. Available: http://www.emt.tugraz.at/~pinz/data/GRAZ_02/.

Corresponding author
Falah Alsaqre can be contacted at: alsaqre@ieee.org