Vis Comput (2015) 31:245–255
DOI 10.1007/s00371-013-0915-0

ORIGINAL ARTICLE

Unsupervised kernel learning for abnormal events detection

Weiya Ren · Guohui Li · Boliang Sun · Kuihua Huang

Published online: 5 January 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract In this paper, we propose a method to detect abnormal events using a novel unsupervised kernel learning algorithm. The key of our method is to learn a suitable feature space and the associated kernel function of the training samples. By considering the self-similarity property of training samples, we assume that the training samples will show the distinctly clustering property in the obtained feature space. Non-negative matrix factorization (NMF) is used to learn the feature space, and the support vector data description (SVDD) method is adopted to measure the clustering degree of instances in the feature space. We append the clustering constraints in the process of learning the feature space and use the bases produced by NMF as the projection matrix to construct the kernel function in SVDD. In other words, we incorporate the minimal enclosing sphere constraints within the NMF formulation. In the process of feature space learning, instances in the obtained feature space will be described better and better by an hypersphere. Our algorithm converges to a local optimal solution by applying an alternating optimization approach. Experimental results on three public datasets and the comparison to the state-of-the-art methods show that our method is effective in detecting and locating unknown abnormal behaviors.

W. Ren (✉) · G. Li · B. Sun · K. Huang
College of Information System and Management,
National University of Defense Technology, Changsha 410072,
People's Republic of China
e-mail: weiyren.phd@gmail.com

G. Li
e-mail: guohli@nudt.edu.cn

B. Sun
e-mail: sumboliang@nudt.edu.cn

K. Huang
e-mail: khhuang.nudt@gmail.com

Keywords Kernel learning · One-class learning · Anomaly detection · Non-negative matrix factorization · Support vector data description

1 Introduction

Intelligent visual surveillance has made great progresses in recent years [14–16]. However, we still need to deal with more challenges such as emergent behaviors and self-organizing activities [20] in crowd scene analysis. According to the definition in Oxford English Dictionary, the abnormal events can be regarded as irregular or rarely events in contrast with normal ones. Data on the normal events in video are cheap to obtain, while data of anomalies always require an expensive cost to obtain. Thus, anomaly detection task can be turned into an one-class learning (outlier detection) problem [10] that identifies abnormalities (outliers) based on some normal training samples.

Outlier detection approaches are usually classified into four categories based on the techniques used, which are: distribution-based, distance-based, density-based and deviation-based approaches [29]. There are many representative outlier detection methods such as resolution based outlier factor method [30], influenced outlierness method [31], and angle-based outlier degree method [32]. However, outlier detection is still highly challenging mainly due to three reasons: (1) Defining the normal behavior or region [28]; (2) An existing notion of normal behavior might not be sufficiently representative in the future; (3) The boundary between normal and outlying behavior is often fuzzy.

Kernel methods are famous for its learning ability. Nevertheless, a predefined kernel function might not always be appropriate for a given one-class learning problem. We now

face the problem of how to find a suitable kernel function and the associated feature space. To cope with this problem, we try to learn a kernel matrix (also known as a "Gram matrix") to embed the training data into a Hilbert space where we can easily solve this one-class learning problem.

Non negative matrix factorization (NMF) [1,2] decomposes the non-negative data as a product of two non-negative matrices. The non-negativity constraint leads NMF to a part-based representation of the object in the sense that it only allows additive, not subtractive, combination of the original data. NMF is widely used in many computer vision applications such as face recognition [1], clustering [9] and action recognition [3]. NMF is also used in subspace learning through producing a projection bases matrix [8,24,27]. However, NMF has its shortcomings [8,22] so it is almost impossible to get the desired feature space merely by NMF in one-class learning. To solve this intractable problem, we need a method to control the process of learning feature space.

Abnormal detection task consists of normal examples and abnormal examples. Taking a human cognition view, we can easily represent a normal example by normal ones, but representing an abnormal example that rarely happens by normal ones is hard. Therefore, it is reasonable to assume that normal examples have the self-similarity property, in mathematic way, it is equivalent to assume that the training instances will have good clustering property in some feature space. Thus, we utilize the clustering property to control the direction of learning feature space. In our proposed method, we use support vector data description (SVDD) [4] to measure the degree of instances' clustering property. SVDD gives a description of a set of objects by minimizing the chance of accepting outliers and finding an hypersphere with minimal radius, and it is an effective method to describe one-class data as it finds support vectors and allows outliers.

By appending the clustering constraints in the process of learning feature space, we can get a suitable feature space mapping and the associated kernel function. Our algorithm converges to a local optimal solution by applying an alternating optimization approach. Considering the above factors, a novel unsupervised framework named Unsupervised Kernel Learning with Clustering Constraint (CCUKL) is proposed to solve the unsuspected anomaly detection task in video surveillance.

The rest of the paper is organized as follows. The proposed method is given in Sect. 2. In Sect. 3, we formulate the proposed CCUKL framework and present the method that solves the corresponding optimization problem. We present experimental results on several anomaly detection problems using publicly available datasets in Sect. 4. Finally, we summarize the approach and present some clues for future research work.

2 Proposed method

As we said before, we utilize the self-similarity property of training samples to learn a suitable feature space. As seen in Fig. 1, our main idea is to define an unknown matrix kernel and then solve it by a matrix learning method under clustering constraint. In the process of learning feature space,

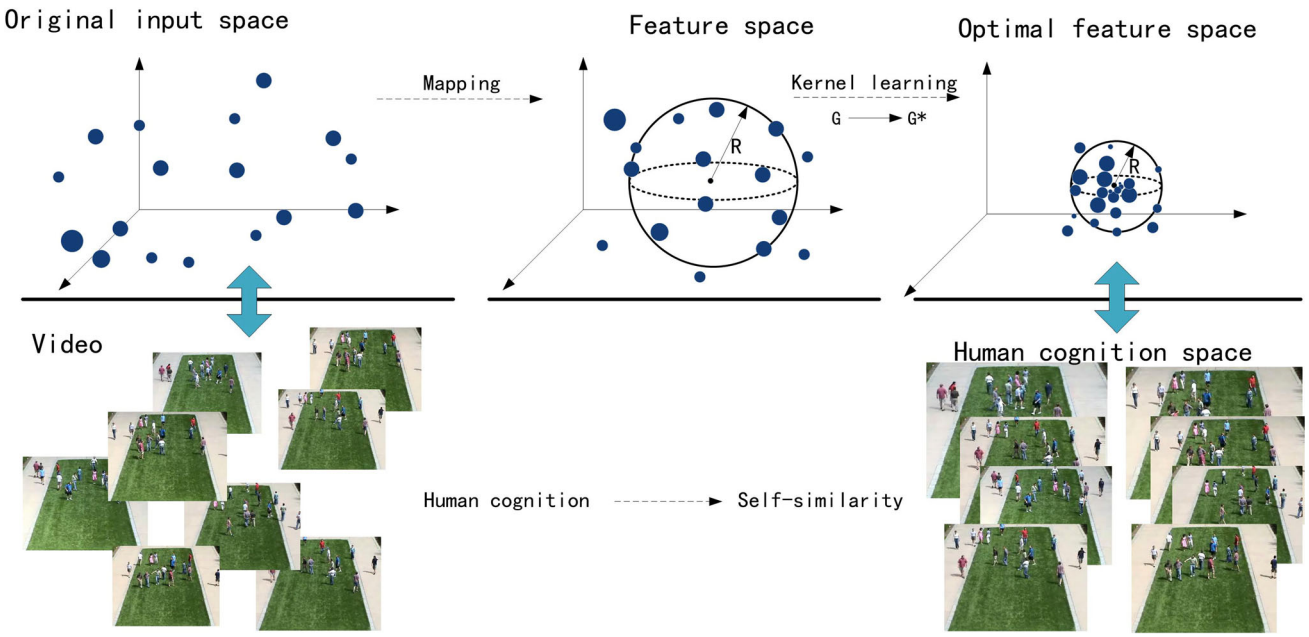


Fig. 1 Unsupervised kernel learning. Normal examples have the self-similarity property in human cognition, in mathematic way, training instances will have good clustering property in the obtained feature space. The clustering degree of instances in the feature space can reflect the quality of the learned feature space. The center of the hypersphere converges in the process of kernel matrix learning

self-similarity property controls the learning direction and evaluates the quality of the learned space.

Given an input training set $X \in \mathbb{R}^{m\times n}$ and an embedding space $\mathcal{F}$, we consider a feature space mapping $\varphi : \mathbb{R}^m \to \mathcal{F}$, where $\varphi(x) = G^{\mathrm{T}}x, x \in \mathbb{R}^{m}, G \in \mathbb{R}^{m\times k}, G \succcurlyeq 0$. The corresponding kernel function can be defined as $\kappa(x, y) = (\varphi(x), \varphi(y)) = x^{\mathrm{T}}GG^{\mathrm{T}}y$, where $\kappa : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. κ is a positive definite kernel for it satisfies Mercer's condition [5], and the symmetric and positive definite matrix $G^{\mathrm{T}}G$ is the kernel matrix which needs to be learned.

We denote the clustering parameter as De, and denote the self-similarity constraint on instances in feature space $\mathcal{F}$ as $C_{\mathcal{F}}(G, De, X) \leq 0$. For two given machine learning method ml_1, ml_2, we minimize the sum of the cost function $F_{ml_1 1}(G)$ and the cost function $F_{ml_2}(De)$ under the non-negative constraint for G and the clustering constraint. Therefore, the Unsupervised Kernel Learning with Clustering Constraint (CCUKL) is actually an optimization problem that can be written as:

$$\underset{G,De}{\operatorname{argmin}} \lambda F_{ml_1 1}(G) + F_{ml_2}(De) \quad s.t.\ C_{\mathcal{F}}(G, De, X) \leq 0, G \succcurlyeq 0 \tag{2.1}$$

Non-negative matrix factorization (NMF) is used as the method ml_1 to produce the bases matrix G as the projection in mapping $\varphi(x)$. Let $X \in \mathbb{R}^{m\times n}$ represent a non-negative matrix having n examples in its columns, NMF aims to find two non-negative matrices: the bases matrix $G \in \mathbb{R}^{m\times k}$ and the coefficients matrix $H \in \mathbb{R}^{k\times n}$ such that $X \approx GH$. The unknown matrices G and H are estimated by minimizing the reconstruction error $||X - \mathrm{GH}||_F^2$ or the Kullback–Leibler divergence $D(X|\mathrm{GH})$ [2]. The bases matrix G is used as the projection matrix in mapping $\varphi(x)$ recently [8,24]. At the same time, classical NMF-based algorithms use $G^{\dagger} = (G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}$ as the projection matrix. The above two choices are equally valid [8], and the former one is easier to work with. Support vector data description (SVDD) is used as the method ml_2 to describe instances' self-similarity and measure their clustering degree.

We form our optimization problem by incorporating the minimal enclosing sphere constraints within the NMF formulation. In the process of solving the optimization problem, the kernel matrix updates iteratively and the obtained hypersphere of training samples becomes smaller. Every two instances will have a close distance as small as they can in the obtained feature space. In respect that the difference between points and the punishment of NMF factorization error, the distance between points will not equal zero.

Our proposed algorithm is solved by applying an alternating optimization approach. For more precise, we solve a convex (quadratic or SVDD-type) sub-problem which contains only a subset of the unknown parameters meanwhile keeping the others fixed at each iteration. Finally, the obtained feature space and the obtained hypersphere can be used for anomaly detection. The proposed method judges anomaly only by a judge function in practice, so CCUKL will be effective when faces mass testing data.

In testing phase, a test sample will be attracted by the hypersphere if it is similar with the training samples. Otherwise, a testing sample will be mapped into the feature space in an uncertain way if it is dissimilar with the training samples. Notice that the hypersphere in the feature space is very small, so the probability of the abnormal samples falling into the hypersphere is quite small.

3 Algorithm

3.1 Optimization problem

The proposed method is expatiated in Sect. 2. In this section, we will present the CCUKL algorithm. In mathematics way, the optimization problem is:

$$\underset{G,H,C,R,\xi_i}{\operatorname{argmin}} \lambda||X - \mathrm{GH}||_F^2 + R^2 + \frac{1}{\upsilon n}\sum_{i=1}^{n}\xi_i \tag{3.1}$$

Subject to

$$||G^{\mathrm{T}}x_i - C||_{\mathcal{H}}^2 \leq R^2 + \xi_i, H \succcurlyeq 0, G \succcurlyeq 0,$$
$$R \in \mathbb{R}^+, C \in \mathbb{R}^k, \xi_1, \xi_2, \ldots, \xi_n \in \mathbb{R}^+ \tag{3.2}$$

At first, we fix some notation, $||\cdot||_F$ denotes the L_2 norm of a vector and the Frobenius norm of a matrix. We use the notation $H \succcurlyeq 0$ to express that the elements of the matrix H are non-negative. $X \in \mathbb{R}^{m\times n}$ denotes training data containing n columns, $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^{m\times 1}$, $G \in \mathbb{R}^{m\times k}$ is the bases matrix and $H \in \mathbb{R}^{k\times n}$ is the coefficients matrix. R is the radius of hypersphere and C is the center point of hypersphere in the feature space.

The first term of the above optimization problem ($\lambda||X - \mathrm{GH}||_F^2, \lambda > 0$) is a classical NMF-type reconstruction error, while sum of the second and the third term ($R^2 + \sum_{i=1}^{n}\xi_i/\upsilon n$) is a SVDD type cost, the parameter $1/\upsilon n$ controls the trade-off between the radius and the errors [4]. Smaller υ means fewer points will be outliers, a more detailed analysis shows that υn is in fact a lower bound on the number of outliers over the training set [13]. Actually, the hypersphere's center point C can be expressed as a linear combination $C = \sum_{i=1}^{n}\varepsilon_i G^{\mathrm{T}}x_i$ using represent theorem [5].

Notice that the inequality constraint $||G^{\mathrm{T}}x_i - C||_{\mathcal{H}}^2 \leq R^2 + \xi_i$ involves slack variables $\xi \in \mathbb{R}^{1\times n}$ which control the misclassification errors. To address this optimization problem, inspired by some literatures [8,9], we solve for subsets of the unknown parameters G, H, R, C, ξ_i by keeping the remaining parameters fixed at each iteration.

3.2 Update strategy

3.2.1 Update strategy for G

In this section, we solve for G, ξ by keeping H, R and C fixed, the optimization problem in Eq. 3.1 can be simplified as:

$$\underset{G,H,C,R,\xi_i}{\operatorname{argmin}} \; ||X - \mathrm{GH}||_F^2 + \frac{1}{\lambda \upsilon n} \sum_{i=1}^{n} \xi_i \tag{3.3}$$

Subject to

$$||G^{\mathrm{T}} x_i - C||_{\mathcal{H}}^2 \leq R^2 + \xi_i, G \succcurlyeq 0, \xi_1, \xi_2, \ldots, \xi_n \in \mathbb{R}^+ \tag{3.4}$$

Considering constraints in Eq. 3.4 as nonlinear, we solve this optimization problem by getting its dual problem, it's Lagrangian function is shown as follows:

$$\begin{aligned}\mathcal{L}(G, \xi_i, \alpha_i, \beta_i) = tr[&-2X^{\mathrm{T}}GH + H^{\mathrm{T}}G^{\mathrm{T}}GH] \\ &+ \frac{1}{\lambda \upsilon n} \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (x_i^{\mathrm{T}} G G^{\mathrm{T}} x_i \\ &- 2x_i^{\mathrm{T}} GC + C^{\mathrm{T}} C - R^2 - \xi_i) \\ &- \sum_{i=1}^{n} \beta_i \xi_i - tr[\varpi^{\mathrm{T}} G]\end{aligned} \tag{3.5}$$

where $\alpha_i, \beta_i, \varpi_{i,j} > 0$ are the Lagrangian multipliers, and $\varpi_{i,j}$ enforces non-negative constraints $G_{i,j} > 0$. Setting the partial derivatives to zero, new constraints are obtained:

$$0 \leq \alpha_i \leq \frac{1}{\lambda \upsilon n}, \quad 1 \leq i \leq n \tag{3.6}$$

$$\varpi = 2GHH^{\mathrm{T}} - 2XH^{\mathrm{T}} + 2(o_m \alpha \circ X)X^T G - 2X\alpha^T C^T \tag{3.7}$$

where $o_m = [1, 1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^{m \times 1}, \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n] \in \mathbb{R}^{1 \times n}$, $\varpi = [\varpi_{ij}] \in \mathbb{R}^{m \times k}$.

Taking under consideration the KKT conditions [11], we get:

$$\begin{aligned}&2(GHH^{\mathrm{T}} - XH^{\mathrm{T}} + (o_m \alpha \circ X)X^{\mathrm{T}} G - X\alpha^{\mathrm{T}} C^{\mathrm{T}})_{i,j} G_{i,j} \\ &= \varpi_{i,j} G_{i,j} = 0\end{aligned} \tag{3.8}$$

Equation 3.8 is a fixed point equation that the solution must satisfy at convergence, then $G_{i,j}$ can be updated, we can directly calculate columns of G directly to save computation time:

$$\tilde{G}_j = G_j \circ \sqrt{\frac{([XH^{\mathrm{T}}]_j + [X\alpha^{\mathrm{T}} C^{\mathrm{T}}]_j)}{([GHH^{\mathrm{T}}]_j + [(o_m \alpha \circ X)X^{\mathrm{T}} G]_j)}} \tag{3.9}$$

where G_j means jth column of G, $\circ$ denotes Hadarmard product(element-wise multiplication).

Denote $\tilde{G} = [\tilde{G}_{ij}]$ and substitute the value of G and ϖ in Eq. 3.7 and simplifying, we get the dual problem:

$$\begin{aligned}&\underset{\alpha}{\operatorname{argmax}}\, tr[(X - \tilde{G}H)^{\mathrm{T}}(X - \tilde{G}H)] \\ &-2tr[(\tilde{G}HH^{\mathrm{T}} - XH^{\mathrm{T}} + (o_m \alpha \circ X)X^{\mathrm{T}} \tilde{G} - X\alpha^{\mathrm{T}} C^{\mathrm{T}})^{\mathrm{T}} \tilde{G}] \\ &+\alpha(X^{\mathrm{T}} \tilde{G} \tilde{G}^{\mathrm{T}} X \circ I_n)o_n - 2\alpha X^{\mathrm{T}} \tilde{G} C + \alpha o_n (C^{\mathrm{T}} C - R^2)\end{aligned} \tag{3.10}$$

Subject to

$$0 \leq \alpha_i \leq \frac{1}{\lambda \upsilon n}, \quad 1 \leq i \leq n \tag{3.11}$$

where I_n denotes an unit matrix of size n.

The above problem is quadratic in α, thus can be solved by using conventional quadratic programming tools. The estimated α is then used to update G using Eq. 3.9, notice that large values of λ (when compared to $1/\upsilon n$) result in getting small α_i.

3.2.2 Update strategy for R, C

In this section, we proceed in solving for the minimal enclosing sphere that minimizes the radius of the enclosing sphere by keeping the bases matrix G and weights matrix H fixed, so the optimization problem in Eq. 3.1 is simplified to a classical SVDD problem:

$$\underset{C,R,\xi_i}{\operatorname{argmin}} \; R^2 + \frac{1}{\upsilon n} \sum_{i=1}^{n} \xi_i \tag{3.12}$$

Subject to

$$\left|\left|G^{\mathrm{T}} x_i - C\right|\right|_{\mathcal{H}}^2 \leq R^2 + \xi_i, C \in \mathbb{R}^d, \xi_1, \xi_2, \ldots, \xi_n \in \mathbb{R}^+ \tag{3.13}$$

We introduce the Lagrangian function:

$$\begin{aligned}\mathcal{L}(R, \xi_i, C, \tau_i, \varrho_i, \zeta) = R^2 &+ \frac{1}{\upsilon n} \sum_{i=1}^{n} \xi_i \\ &+ \sum_{i=1}^{n} \tau_i (-2x_i^{\mathrm{T}} GC + C^{\mathrm{T}} C - R^2 - \xi_i) - \sum_{i=1}^{n} \varrho_i \xi_i\end{aligned} \tag{3.14}$$

where $\tau_i, \varrho_i > 0$ are the Lagrangian multipliers, $\tau = [\tau_1, \tau_2, \ldots, \tau_n] \in \mathbb{R}^{1 \times n}$, setting the partial derivatives to zero, new constraints are obtained:

$$0 \leq \tau_i \leq \frac{1}{\upsilon n}, \quad \sum_{i=1}^{n} \tau_i = 1, 1 \leq i \leq n$$

$$C = \sum_{i=1}^{n} \frac{\tau_i (G^T x_i)}{\sum_{i=1}^{n} \tau_i} = \sum_{i=1}^{n} \tau_i (G^T x_i) \tag{3.15}$$

Resubstituting gives to maximize with respect to τ:

$$\underset{\tau}{\operatorname{argmax}}\, \tau (X^{\mathrm{T}} G G^{\mathrm{T}} X \circ I_n) o_n - \tau X^{\mathrm{T}} G G^{\mathrm{T}} X \tau^{\mathrm{T}} \tag{3.16}$$

Subject to

$$\sum_{i=1}^{n} \tau_i = 1, 0 \leq \tau_i \leq \frac{1}{\upsilon n}, 1 \leq i \leq n \tag{3.17}$$

Then, R can be updated as follows:

$$R^2 = x_k^{\mathrm{T}} G G^{\mathrm{T}} x_k - 2\tau X^{\mathrm{T}} G G^{\mathrm{T}} x_k + \tau X^{\mathrm{T}} G G^{\mathrm{T}} X \tau^{\mathrm{T}} \tag{3.18}$$

where $G^{\mathrm{T}} x_k$ is one of the support vectors on the ball boundary. In this way, the minimal enclosing sphere parameters R and C are obtained.

3.2.3 Update strategy for H

After above two steps, the next task is to update weights matrix H for we have already acquired the values for G,R and C. By keeping all the other variables fixed, the objective function Eq. 3.1 is simplified as:

$$\underset{H}{\operatorname{argmin}} ||X - \mathrm{GH}||_F^2 \tag{3.19}$$

Subject to

$$H \succcurlyeq 0$$

The Lagrangian of the above cost function is:

$$\mathcal{F}(H, \tau) = tr[-2X^{\mathrm{T}} G H + H^{\mathrm{T}} G^{\mathrm{T}} G H - \eta^{\mathrm{T}} H] \tag{3.20}$$

where $\eta_{i,j} > 0$ are the Lagrangian multipliers, and $\eta_{i,j}$ enforce non-negative constraints $H_{i,j} > 0$. Setting the partial derivatives to zero, new constraints are obtained:

$$\eta = -2G^{\mathrm{T}} X + 2G^{\mathrm{T}} G H \tag{3.21}$$

Taking under consideration the KKT conditions [11], we get:

$$2(-[G^{\mathrm{T}} X] + [G^{\mathrm{T}} G H])_{i,j} H_{i,j} = \eta_{i,j} H_{i,j} = 0 \tag{3.22}$$

Equation 3.22 is a fixed point equation that the solution must satisfy at convergence, then $H_{i,j}$ can be updated, notice that h_j (jth column of H) contributes only to the j-th data point x_j, so columns of Hcan be solved independently as follows:

$$H_j = H_j \circ \sqrt{\frac{([G^{\mathrm{T}} X]_j)}{([G^{\mathrm{T}} G H]_j)}} \tag{3.23}$$

Experiments show that this kind of update strategy is better than optimizing the primary problem which is more time consuming and less accurate in practice.

3.3 Judge rule

In this section, we discuss the criterion rule of anomaly alarming. Given an instance vector x which denotes a patch in a frame, we map it into the feature space by mapping $\varphi(x)$ and then determine whether to alarm. The judge functions is:

$$\mathscr{F}(x) = ||G^{\mathrm{T}} x - C||_{\mathcal{H}}^2 - R^2 \tag{3.24}$$

The new patch's t history normal patches in t frames are denoted as $X_{\mathrm{h}} = [x_{\mathrm{h}1}, x_{\mathrm{h}2}, \ldots, x_{\mathrm{ht}}]$. Given a patch x and its t history normal patches X_{h}, we can judge x using permutation hypothesis testing method [25].

Now, we address the hypothesis testing problem by denoting $H_0 : F_{\mathrm{N}} = F_{\mathrm{A}}$, where F_{N} and F_{A} are distribution functions of $\mathscr{F}(X_{\mathrm{h}})$ and $\mathscr{F}(x)$, respectively, i.e. $\mathscr{F}(x_{\mathrm{h}1}), \mathscr{F}(x_{\mathrm{h}2}), \ldots, \mathscr{F}(x_{\mathrm{ht}}) \sim F_{\mathrm{T}}$ and $\mathscr{F}(x) \sim F_{\mathrm{A}}$. To judge whether a patch is normal or not, we compute the p value which is the minimal significance level to refuse H_0. Algorithm of permutation hypothesis testing method includes three steps:

1. $T_{\mathrm{obs}} = T(\mathscr{F}(x_{\mathrm{h}1}), \mathscr{F}(x_{\mathrm{h}2}), \ldots, \mathscr{F}(x_{\mathrm{ht}}), \mathscr{F}(x))$
   $= |\frac{\sum_i \mathscr{F}(x_{\mathrm{hi}})}{t} - \mathscr{F}(x)|$.
2. Permutate $\mathscr{F}(x_{\mathrm{h}1}), \mathscr{F}(x_{\mathrm{h}2}), \ldots, \mathscr{F}(x_{\mathrm{ht}}), \mathscr{F}(x)$ randomly B times and compute T at the same time, denoted as $T_1, T_2, \ldots, T_B$.
3. $$p \approx \frac{1}{B} \sum_{i=1}^{B} I(T_i > T_{\mathrm{obs}}). \tag{3.25}$$

By setting the significance level a, we can judge a patch by computing the p value, more precisely, alarm when $p <$ a. During the real-time testing, we judge each patch in each frame, so we can not only detect anomalies but also locate them.

3.4 CCUKL analysis

Inspired by EM algorithm and alterative projections, we solve a set of convex sub-problems each of which is guaranteed to converge, so Eq. 3.1 which can be regarded as a monotonic function converges to a local minimum value. We show the experiment result of the convergence issues in Sect. 4.

The parameter selection is important in our framework. Tradition NMF method often tries to find a low-dimensional representation so that the dimension k should be small and a small k will also reduce the computational complexity, i.e.

Table 1 A finite grid of parameter values

Parameter	Values
υ	$2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}$
λ	$2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}$
k	$1.2^{-3}m, 1.2^{-2}m, 1.2^{-1}m, 1.2^{0}m, 1.2^{1}m, 1.2^{2}m, 1.2^{3}m$

We find the best performance of our algorithm on this finite grid

$k < \min(m, n)$. However, we find that the proposed algorithm converges faster when k gets larger and a large k ensures the mapped space has a necessary large dimension. Therefore, k should be a trade-off based on the above two factors. Large λ and υ will reduce the effect of similarity constraint and improve the efficiency of getting feature space. Besides, the parameters υ has similar effects on generalization as in the original SVDD.

We select combinations of the parameter values on a finite grid, based on the idea in [33], as seen in Table 1. In this way, it is sufficient to perform algorithm comparisons.

In practice, one has to try many choices of parameters during the model selection. We use tenfold cross-validation method [34] to evaluate the choices of parameters. In our experiments, λ and υ are set to 1, and k is set to $1.2m$. Significance level a in Sect. 3.3 is set to 0.05, a small a predicates a low tolerance to make the second-type error. In this way, we avoid threshold setting which may be different in different scenes.

CCUKL framework mainly embraces four part: inputs, initialization, training and testing, notice that the training part do not need too many normal samples. Moreover, CCUKL will not take too much time for training and testing in practice, actually the most time-consuming part is the computing of optical flow. Algorithm process of CCUKL is shown as follows:

Inputs: Several normal frames; Parameter setting of λ, υ.
Initialization: Key parameters in CCUKL–G,H,R,C.
Training: Solve parameters solved in an iterative manner. Solve α, G, R, C, H in order and repeat the process until satisfy the terminate conditions.
Testing: Receive a frame and judge each patch of it using proposed judge function and permutation hypothesis testing method.

4 Experiment

Training samples X can be quite fiexible, such as image patches or spatio-temporal subvolumes. Actually, X correspond to the bases selection, such as spatial bases, temporal bases and spatial-temporal bases [6]. We show our algorithm process in practice using spatial bases and divide each frame into several patches.

Optical flow [21] method is used as the feature extraction, and then we use MHOF [6] method to extract feature of patches. By encoding each patch as a column of $X (X = [x_1, x_2, \ldots, x_n])$, the whole feature data X are generated.

4.1 Tests on UMN dataset

The UMN dataset is a publicly available dataset of normal and abnormal crowd videos from University of Minnesota [26] which consists of 11 video segments. Each video segment consists of an initial part of normal behavior and ends with sequences of the abnormal behavior. There are total 7739 frames with a 320 × 240 resolution. We split each frame into 4 × 6 local patches with no pixel overlapping and extract the MHOF from each sub-region.

In Fig. 2, we show the proposed algorithm in a video segment which contains 625 frames. We use only 10 pairs of adjacent frames for training and test the whole segment by computing their judge function values. A new frame will be decided whether to alarm by considering a certain number of history normal frames and utilizing the permutation hypothesis testing method.

We use only 10 pairs of adjacent frames in one video segment for training and then test all 11 video segments, and the result of our method is shown in Fig. 3. We also display the ground truth and the result using the sparse reconstruction cost (SRC) method [6]. The comparison of results shows that our proposed method performs better.

Table 2 provides the quantitative comparisons to the state-of-the-art methods. The AUC of our method is 0.98 which outperforms [19], and the anomalies can be detected earlier than [7,9].

Figure 4 shows that the factorization error in CCUKL is declining while searching the feature space. Figure 5 displays the hypersphere's center point Cat all iterations. The center point C of training samples in feature space converges. Figure 6 shows the change of hypersphere's radius Rat all iterations. Change of R means that the hypersphere is becoming smaller and clustering degree is becoming higher. In the end, a suitable feature space and a stable hypersphere gradually emerge.

4.2 Tests on recessive walking sequence

We download a sequence about recessive walking [19] and use it as a type of local abnormal event to evaluate our model's performance. There are total 583 frames with a 270 × 480 resolution. Each frame is divided into 4 × 6 local patches with no pixel overlapping. Then, we extract features from each patch. Training data consist of 10 pairs of normal frames. The normal scene can be described as pedestrians walk in the same direction. Figure 7 shows the experiment's result

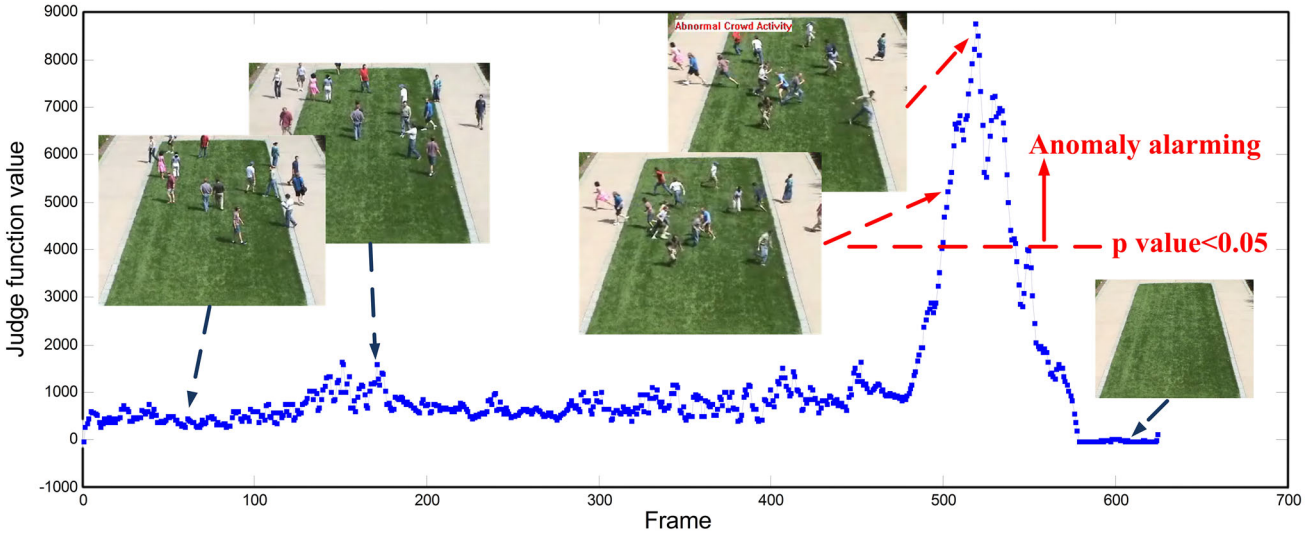


Fig. 2 A temporal smooth is applied for we assume that abnormal events cannot occur only in one frame. 10 normal frames are used for training and the whole video segment is used for testing

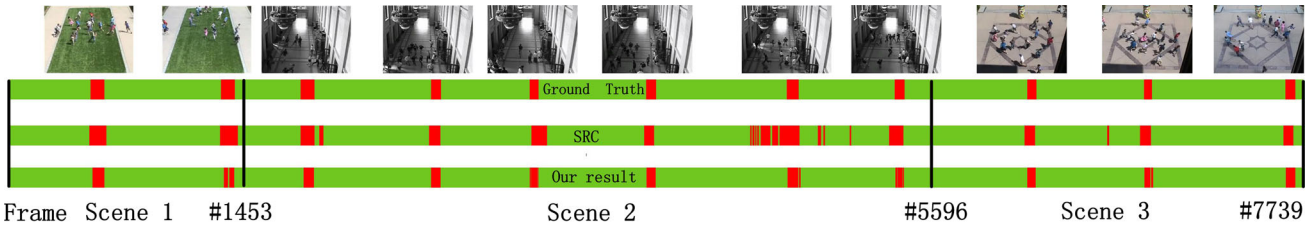


Fig. 3 The qualitative results of the abnormal event detection for 11 video segments from UMN dataset. The *top row* represents the ground truth bar where *green color* denotes the normal frames and *red* corresponds to abnormal frames. The *middle row* is the result using sparse reconstruction cost (SRC) method [6]. At the *bottom*, we show the result using our proposed method

Table 2 The comparison of the use of the proposed method and the state-of-the-art methods on the UMN dataset

Method	Area under ROC
CCUKL	0. 98
Sparse reconstruction [6]	0.98
Chaotic invariants [7]	0.99
Social force [19]	0.96
Pure optical flow [21]	0.84

and we can see that the detecting time and localization effect are all that could be desired (Table 3).

4.3 Tests on UCSD Ped1 dataset

The UCSD Ped1 dataset [6,12,21] contains 70 short clips and each clip has 200 frames, with a 158 × 238 resolution. The training set contains 34 short clips for learning normal patterns. The testing set contains 36 short clips for testing and there is a subset of 10 clips in testing set provided with pixel-level binary masks, which identify the regions containing abnormal events. We split each frame into 7 × 7 local patches with a 22 × 34 resolution, then utilize the judge rule in Sect. 3.3 to determine whether a local patch is normal or not.

In testing phase, patches are mapped into the feature space. Normal ones will be attracted by the hypersphere and abnormal ones will fall into the feature space in an unpredictable way. For the hypersphere is a very small region compared to the feature space, the probability that the abnormal sample falling into the hypersphere is quite small. Some image results are shown in Fig. 8. The proposed method can detect and locate abnormal events which are different from training samples such as biker, skater and vehicle.

As defined in [21], a frame is considered as a detection if it contains at least one abnormal pixel when utilizing frame-level measurement and a frame is considered detected correctly if at least 40 % of the truly anomalous pixels are detected when utilizing pixel-level measurement. We compare the proposed method with sparse reconstruction cost model (denoted SRC) [6], MDT [21], original SVDD with RBF kernel [4], social force model (denoted SF) [21] and SF-MPPCA [21]. In Fig. 9, performance of the approaches tested for the anomaly detection task and performance of

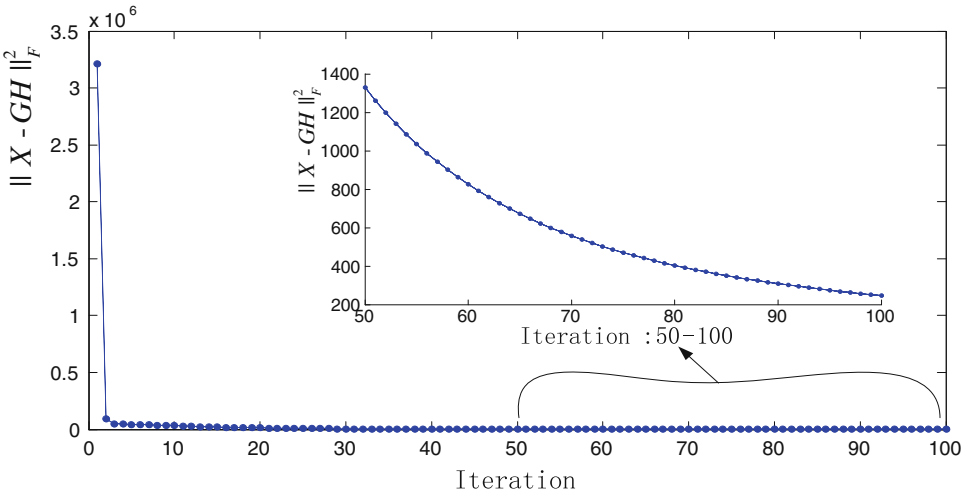


Fig. 4 Factorization error $||X - GH||_F^2$ decreases with iteration and we magnify factorization error's value with iterations change from 50 to 100

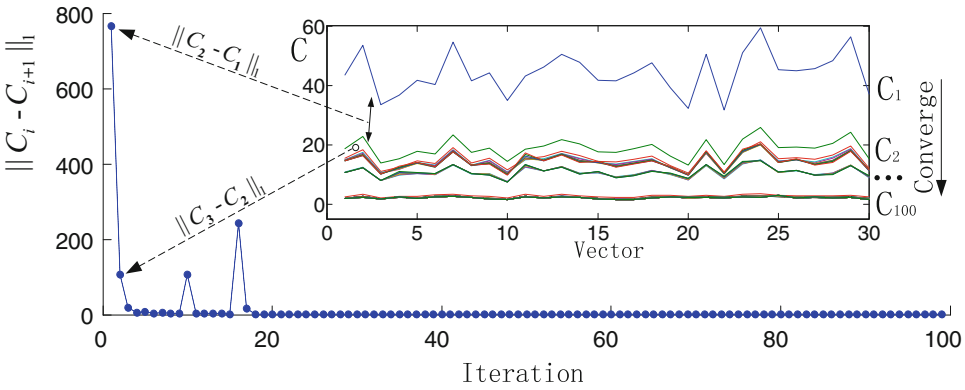


Fig. 5 Hypersphere's center point C converges with iteration, C is expressed as a vector, we calculate the "distance" (using norm 1) between C_i and C_{i+1} and the result shows the convergence of C

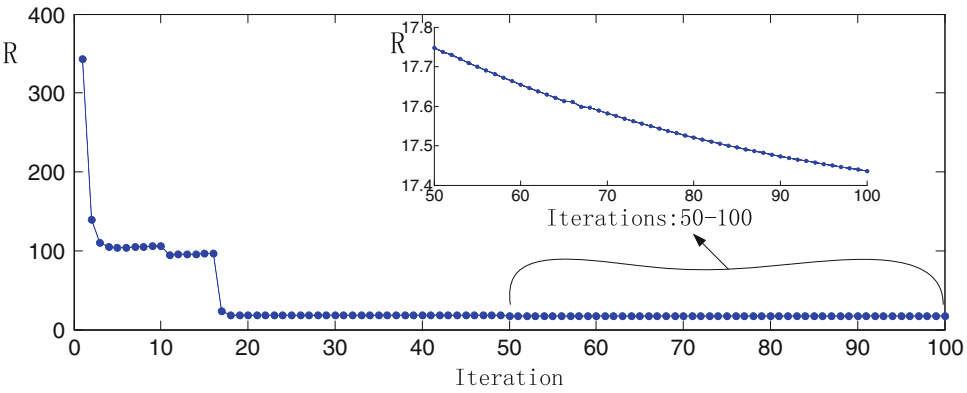


Fig. 6 Hypersphere's radius R decreases with iteration and we magnify R's value with iterations change from 50 to 100

the approaches tested on the anomaly localization with pixel level groundtruth on the UCSD Ped1 dataset are displayed. It is clear that the proposed method outperforms the state-of-the-art methods.

In Table 4, some evaluation results [6,21] are presented: the Equal Error Rate (EER) is the percentage of misclassified frames when the false positive rate is equal to the miss rate (ours 19 % < 25 % [21]), Rate of detection (RD) in anomaly localization experiment (ours 51 % > 46 % [6]) and area under curve (AUC) (ours 50.3 % > 46.1 % [6]), the results show that the CCUKL anomaly detection outperforms the state-of-the-art methods.

5 Conclusions and discussion

The proposed method provides a novel unsupervised kernel learning method to solve the one-class learning problem by utilizing the "self-similarity" property of training samples. In human cognition, the self-similarity property means that similar samples will show a clearly clustering property in a

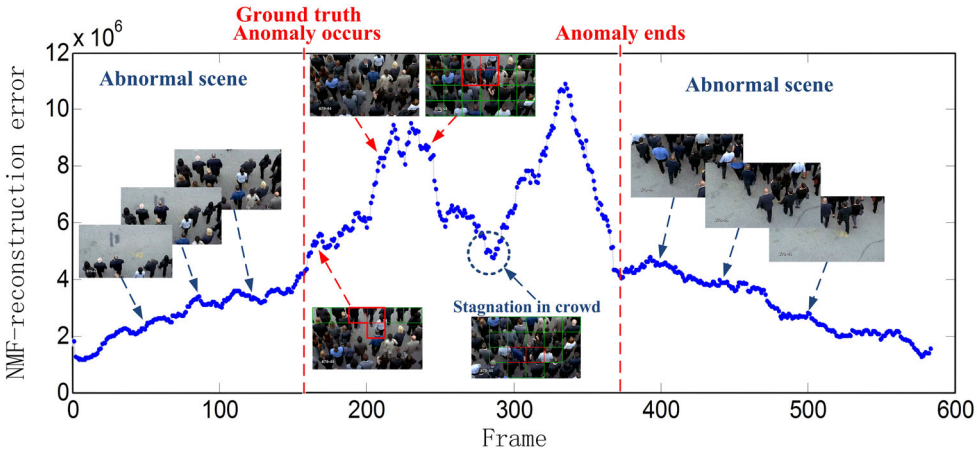


Fig. 7 Examples in recessive walking sequence. Pedestrian who walks in an opposite direction is abnormal and the anomaly is detected fast and localized well

Table 3 The comparison of the use of the proposed method and the state-of-the-art methods on the recessive walking dataset

Method	Area under ROC
CCUKL	0. 98
Sparse reconstruction cost [6]	0.98
Chaotic invariants [7]	0.99
Social force [19]	0.96
Pure optical Flow [21]	0.84

Fig. 8 Examples of abnormal event detections and localizations for UCSD Ped1 datasets. The abnormal events such as biker, skater and vehicle are all well detected and located

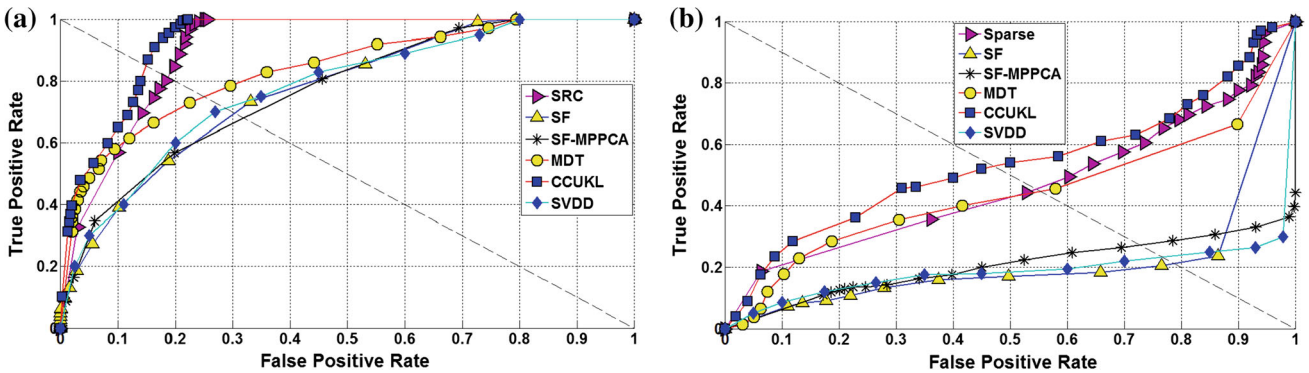


Fig. 9 The detection and localization results of UCSD Ped1 dataset. **a** Frame-level ROC curve for Ped1 dataset. **b** Pixel-level ROC curve for Ped1 dataset

certain feature space in mathematic way. We apply CCUKL to accomplish the anomaly detection and localization task without abnormal samples.

CCUKL judges anomalies just by a judge function so it will be effective when facing mass testing data. The probability of anomalies falling into the hypersphere is quite small

Table 4 Quantitative comparison of the proposed method with the state-of-art methods on the UCSD Ped1 dataset

Method	Equal error rate (%)	Rate of detection (%)	Area under pixel-level ROC (%)
MDT [21]	25	45	44.1
SF [21]	31	21	17.9
SF-MPPCA [21]	32	28	21.3
SRC [6]	19	46	46.1
CCUKL	19	51	50.3
SVDD (RBF kernel) [4]	30	25	19.6

since the hypersphere is just a very small region in the feature space. For instance, $les0.01$ in experimental 4.2.

In future work, the proposed algorithm can be extended to its semi-supervised manner by considering both normal and abnormal samples. Anomalies have different types and differ from each other, so we can apply the CCUKL algorithm to learn any anomaly's feature space if we are interested. Since anomaly clustering is an interesting task, we can do a matching job to recognize one special anomaly if we have an anomaly database. Last but not the least, expanding CCUKL online learning algorithm by updating parameters incrementally is also a meaningful task.

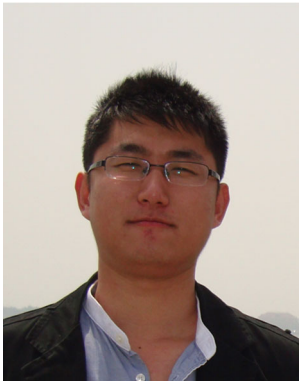
Acknowledgments This paper is supported by College of Information System and Management, National University of Defense Technology and subsidized by National Natural Science Foundation of China(Grant No. 61170158).

References

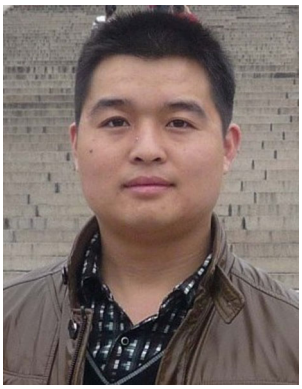
1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
2. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. NIPS **13**, 629–634 (2001)
3. Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images, CVPR, pp. 1–8 (2008)
4. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54**, 45–66 (2004)
5. Christoph, H., Lampert, C.H.: Kernel methods in computer vision. Found. Trends Comput. Graph. Vis. **4**(3), 193–285 (2009)
6. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. CVPR, pp. 3449–3456 (2011)
7. Wu, S., Moore, B., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: CVPR (2010)
8. Vijay Kumar, B.G., et al.: Max-margin non-negative matrix factorization. Image Vis. Comput. **30**, 279–291 (2012)
9. Ding, C.H.Q., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2010)
10. Ritter, G., Gallegos, M.T.: Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern Recognit. Lett. **18**, 525–539 (1997)
11. Pers, J., et al.: Histograms of optical flow for efficient representation of body motion. Pattern Recognit. Lett. **31**, 1369–1376 (2010)
12. http://www.svcl.ucsd.edu/projects/anomaly
13. Smola, A.: Learning: with Kernels: support vector machines. MIT Press, Cambridge, MA (2002)
14. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. CVPR, pp. 2442–2449 (2009)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR, pp. 886–893 (2005)
16. Cong, Y., Gong, H., Zhu, S., Tang, Y.: Flow mosaicking: real-time pedestrian counting without scene-specific learning. CVPR, pp. 1093–1100 (2009)
17. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental up-dates. In: CVPR (2009)
18. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR (2009)
19. Ramin Mehran, M.S., Oyama, A.: Abnormal crowd behavior detection using social force model. In: CVPR (2009)
20. Beibei Zhan, P.R.S.V., Monekosso, D., Xu, L.-Q.: Crowd analysis: a survey. Mach. Vis. Appl. **19**, 345–357 (2008)
21. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. CVPR, pp. 1975–1981 (2010)
22. Pascual-Montano, J., Carazo, K., Kochi, D., Lehmann, R.D.: Pascual-Marqui, Nonsmooth nonnegative matrix factorization (NSNMF). IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 403–415 (2006)
23. Zhang, D., Zhou, Z., Chen, S.: Non-negative matrix factorization on kernels, PRICAI. pp. 404–412 (2006)
24. Bociu, I., Pitas, I.: A new sparse image representation algorithm applied to facial expression recognition, machine learning for signal processing, 2004. Proceedings of the 2004 14th IEEE signal processing society workshop, pp. 539–548 (2004)
25. Wasserman, L.: All of statistics: a concise course in statistical inference. (2004)
26. http://mha.cs.umn.edu/movies/crowdactivity-all.avi
27. Tropp, J.: Literature survey: non-negative matrix factorization, university of Texas at Austin, preprint (2003)
28. Singh, K., Upadhyaya, S.: Outlier detection: applications and techniques. Int. J. Comput. Sci. Issues **9**(1), 3 (2012)
29. Han, J., Kamber. M.: Data Mining: Concepts and Techniques, 2nd edn. Elsevier, Amsterdam, pp. 451–460
30. Fan, H., Zaïane, O. R., Foss, A., Wu, J.: A nonparametric outlier detection for efficiently discovering top-n outliers from engineering data. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore (2006)
31. Jin, W., Tung, A., Han, J., and Wang, W. Ranking outliers using symmetric neighborhood relationship. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore (2006)
32. Kriegel, H.-P., Schubert, M., Zimek, A.: Angle-based outlier detection. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV (2008)
33. Chapelle, O., Zien, A.: Semi-supervised classification by low density separation. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, pp. 57–64 (2005)
34. McLachlan, G.J., Do, K.A., Ambroise, C.: Analyzing microarray gene expression data. Wiley, London (2004)

Weiya Ren received the MSc degree in Management Science from the National University of Defense Technology, China, in 2012. He is currently a Ph.D. student in the College of Information Systems and Management at the National University of Defense Technology (China). His current research interests include machine learning, feature selection, pedestrian tracking, face detection, and image processing.

Guohui Li received the B.S., M.S., and Ph.D. degrees in system engineering from National University of Defense Technology, Changsha, China, in 1983, 1986, and 2001, respectively. He is currently a professor in the department of system engineering, National University of Defense Technology. His research interests include computer vision, information system engineering, data mining, and virtual reality technology.



Boliang Sun received his B.S. and M.S. degrees in control science and engineering from National University of Defense Technology, Changsha, China, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in system engineering. His research interests include machine learning, pattern recognition, and data mining.



Kuihua Huang received his B.S. and M.S. degrees in control science and engineering from National University of Defense Technology, Changsha, China, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in system engineering. His research interests include computational photography, coded exposure photography, and hybrid imaging.