Pattern Recognition 47 (2014) 1791–1799

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse representation for robust abnormality detection in crowded scenes

Xiaobin Zhu [a,b], Jing Liu [b,*], Jinqiao Wang [b], Changsheng Li [c], Hanqing Lu [b]

[a] School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c] IBM Research-China, Beijing, Beijing 100193, China

ARTICLE INFO

Article history:
Received 5 June 2013
Received in revised form
9 November 2013
Accepted 20 November 2013
Available online 1 December 2013

Keywords:
Nonnegative matrix factorization
Crowded scene
Abnormality detection
Sparse coding
Earth mover's distance
Wavelet EMD

ABSTRACT

In crowded scenes, the extracted low-level features, such as optical flow or spatio-temporal interest point, are inevitably noisy and uncertainty. In this paper, we propose a fully unsupervised non-negative sparse coding based approach for abnormality event detection in crowded scenes, which is specifically tailored to cope with feature noisy and uncertainty. The abnormality of query sample is decided by the sparse reconstruction cost from an atomically learned event dictionary, which forms a sparse coding bases. In our algorithm, we formulate the task of dictionary learning as a non-negative matrix factorization (NMF) problem with a sparsity constraint. We take the robust Earth Mover's Distance (EMD), instead of traditional Euclidean distance, as distance metric reconstruction cost function. To reduce the computation complexity of EMD, an approximate EMD, namely wavelet EMD, is introduced and well combined into our approach, without losing performance. In addition, the combination of wavelet EMD with our approach guarantees the convexity of optimization in dictionary learning. To handle both local abnormality detection (LAD) and global abnormality detection, we adopt two different types of spatio-temporal basis. Experiments conducted on four public available datasets demonstrate the promising performance of our work against the state-of-the-art methods.

1. Introduction

Abnormality detection plays an important role in video surveillance context. Recently, there has been an increasing interest in modeling activities and detecting abnormal events in crowded scenes. One of the most challenging tasks in computer vision is event analysis (behavior analysis) in crowded scenes, because it requires monitoring an excessive number of individuals and their activities, retaining structural information regarding the entire scene. A central task of event analysis involves detecting or even predicting abnormal events, i.e., finding new patterns in data that do not conform to the expected case. Fig. 1 shows four examples of abnormal events: (a) and (b) are abnormalities in global scale; (c) and (d) are abnormalities in local scale.

Abnormality detection in crowded scenes is thoroughly studied by the computer vision communities, where some well established models have been developed. As a one-class learning problem, most of algorithms [1–6] intended to detect query sample with lower probability as abnormality by fitting a probability model over the training samples. However, as a high-dimensional feature is essential to better represent the event and the required number of training sample increases exponentially with the feature dimension, it is unrealistic to collect enough data for density estimation in practice. Recently, Cong et al. [7] proposed to detect abnormal events via sparse reconstruction over the normal bases (dictionary). Sparse representation is suitable to model high-dimensional samples [7,8]. Normal event is likely to generate sparse reconstruction coefficients with a small reconstruction cost, while abnormal event is dissimilar to any of the normal basis, thus generates a dense representation with a large reconstruction cost. In crowded scenes, low-level visual information (e.g. optical flow, spatio-temporal interest point) extracted is inevitably uncertain and noisy. Recent works [7,1,5,6,9,10] had shown the power of machine learning or statistic approaches can be used to implicitly handle noisy and uncertainty. However, none of previous works had explicitly tackled with feature's noisy and uncertainty.

To address the above issue, we propose a new approach for abnormality detection in crowded scenes based on sparse coding framework, which is specifically tailored to cope with features' uncertainty and noisy. Different from previous works [7,9], we model the task of dictionary learning as a non-negative matrix factorization problem. There is a subtle difference between dictionary learning and matrix factorization. Matrix factorization can be thought of as a special case of dictionary learning, where the size of the dictionary is constrained to be less than or equal to the observed data dimension [11]. Non-negative matrix factorization (NMF) provides an elegant framework to achieve sparsity on the

[*] Corresponding author. Tel.: +86 10 82544507.
E-mail addresses: xbzhu@nlpr.ia.ac.cn (X. Zhu), jliu@nlpr.ia.ac.cn (J. Liu), jqwang@nlpr.ia.ac.cn (J. Wang), csli@nlpr.ia.ac.cn (C. Li), luhq@nlpr.ia.ac.cn (H. Lu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.018

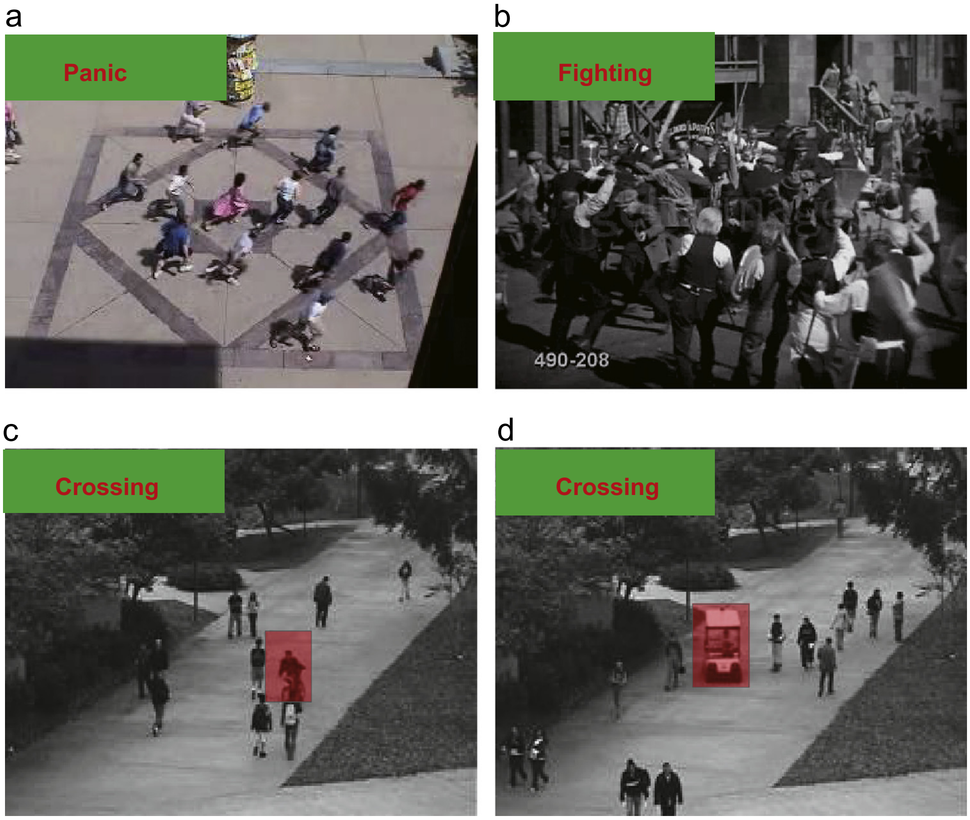

Fig. 1. Abnormal event examples: (a) people are running away in panic; (b) people are fighting; (c) bicycle is crossing the sidewalk; (d) car is crossing the sidewalk. The red masks indicate where the abnormality is. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

basis or coefficient matrices by using corresponding regularizer. In our algorithm, we adopt the Earth Mover's Distance (EMD) as objective cost function, which is a well-known robust metric in the case of noisy histogram comparison. To further reduce the influence of noisy data, we constrain the feature sample as a sparse linear combination of the elements in training dictionary by a weighted $l_{2,1}$ minimization on weight matrix in dictionary learning. The problem with the EMD is its expensive computation, which prohibits its applications in many vision problems. To tackle this problem, we introduce approximate EMD [12] into our approach to greatly degrade the computation complexity without lose of performance. In addition, the combination of approximate EMD and our approach guarantees the convexity of the optimization problem. To the best of our knowledge, no previous works have addressed abnormality detection for crowed scenes under a non-negative matrix factorization framework with EMD metric. Our method has been tested on four video datasets, all of which are publicly available. The experimental results show that our approach successfully detects abnormality both in global and local crowded scenes, and it is very competitive with respect to state-of-the-art algorithms [7,1].

The rest of the paper is organized as follows. Section 2 overviews the related works. In Section 3, we introduce the details of Earth Mover's Distance. In Section 4, we elaborate our method. The experimental evaluation is given in Section 5, and we draw conclusion in Section 6.

2. Related work

Abnormality detection is an active area of research on its own [13]. Various approaches have been proposed for both crowded and uncrowded scenes. Abnormality in crowded scenes is very challenging, due to the diversity of events and the noise in the scenes. In the following, we will focus on the case of crowded scenes. In crowed scenes, object occlusion can severely affect the accuracy of segmentation or tracking, which can heavily degrade the performance of detection. Additionally, the computational cost will also be tremendous when various objects exist. Consequently, the most popular works of abnormality detection extract motion or spatio-temporal interest points from local 2*D* patches or local 3*D* bricks to avoid tracking [14,7,1,4,15–18].

Depending on applications, the works relevant to the crowded scenes can be broadly divided into two categories: the local abnormality detection (LAD) and the global abnormality detection (GAD) [7]. For LAD, the local behavior is different from its spatio-temporal neighborhoods. Kratz and Nishino [14] proposed a spatial-temporal model, in which a coupled distribution-based Hidden Markov Model (HMM) is used to detect abnormalities in densely crowded scenes. However, this method may work only for one kind of normal behavior type in the scene. Mahadevan et al. [19] modeled the normal crowded behavior using mixtures of dynamic textures (MDT). Abnormality detection is formulated as an outlier detection problem, where temporal anomalies equated to events of low-probability, and spatial anomalies equated to events of discriminant saliency. In [20], a probabilistic framework based on Neyman–Pearson decision rule was developed to detect local anomalies that have infrequent patterns with respect to their neighbors. Kim et al. [6] modeled local optical flow with a mixture of probabilistic PCA models and enforce the consistency by Markov Random Field. Antic [5] parsed video frames by establishing a set of hypotheses that jointly explain all the foreground while, at the same time, trying to find normal training samples that explain the hypotheses. Zhao et al. [9] proposed a fully unsupervised dynamic sparse coding approach for detecting unusual events in videos based on online sparse re-constructibility of test samples from an atomically learned event dictionary, which forms sparse coding bases.

For GAD, the whole scene is abnormal, even though any individual local behavior can be normal. Mehran et al. [1] adopted simplified social force model (SFM) [21] to detect abnormal events. The SFM was used to estimate the interaction force under the belief that it contains significant information for describing crowded behavior. After estimating the so-called force flow, a bag of words method and Latent Dirichlet Allocation (LDA) [22] were employed for abnormality detection, while localizing the abnormal areas as those with the highest force magnitude. In [23], streaklines were introduced and integrated with a particle advection scheme capable of incorporating the spatial change in the particle flow. In [4] chaotic invariant was adopted to analyze events in both coherent and incoherent scenes. In [24], the Particle Swarm Optimization (PSO) method was introduced as a robust algorithm for optimizing the interaction force computed using SFM. However, the threshold method used for detecting anomaly is too arbitrary and scene dependent. In [2], an interaction energy potential based on the Linear Trajectory Avoidance (LTA) method was proposed to represent the behavior state of a subject, and velocity was used as its actions. Finally, SVM was used to find abnormal events.

For accommodating both local and global abnormality detection, we adopt two different types of basis. And we formulate the task of dictionary learning as a non-negative matrix factorization (NMF) problem. Non-negative matrix factorization (NMF) [25] for dictionary learning has been considered in many works in computer vision [11,26,27]. NMF provides an elegant framework to achieve sparsity on the basis or coefficient matrices by using the regularizer theoretically [28,29]. Recently Zen et al. [30,31] considered the problem of complex scene analysis and integrated effectively a sparse constraint into the Earth Mover's Distance matrix factorization. However, our work is very different from [30] as it considers specifically the problem of abnormality detection in crowded scenes. Besides, we intend to impose a sparse constraint on the coefficient factor as the learnt dictionary is used for a sparse representation of a signal or a set of signals, so the sparsest representation is certainly appealing. In addition, wavelet EMD is introduced to greatly save computation complexity and well integrated into our algorithm for transferring the optimization problem into a convex but nonsmooth problem, which can well be solved by Nesterove's method [32].

3. The Earth Mover's Distance

The Earth Mover's Distance (EMD) is motivated by the following intuitive observation. The distance between two signatures which may be considered as small local deformations should be less than that of other pair of signatures, which differ in non-adjacent bins. Therefore, an intuitive metric would be the sum of the changes required to transform one signature into the other with low cost caused by local deformations and high cost by nonlocal ones. Let b, p be two histograms, with D-dimension, and are all normalized to unit mass. The EMD is obtained as the solution of the transportation problem:

$$\min_{f_{ij} \geq 0} \sum_{i=1}^{D} \sum_{j=1}^{D} d_{i,j} f_{i,j}, \quad \text{s.t.} \quad \sum_{i=1}^{D} f_{i,j} \leq p_j, \quad \sum_{j=1}^{D} f_{i,j} \leq b_i \tag{1}$$

where $f_{i,j}$ denotes the flow between b_i and p_j, and $d_{i,j}$ denotes the ground distance between i and j. In general, the ground distance $d_{i,j}$ can be any distance (such as L_1 and L_2 distance) and will be chosen according to the particular problem. The problem of Eq. (1) is a Linear Program (LP) problem which can be solved efficiently, due to the special structure of its sparse constraints [33,34]. However, in the case of high dimensional histograms, solving Eq. (1) can be very time consuming due to the large number of flow variables involved. For a D-dimension histogram, the computational complexity is $O(D^3 \log D)$.

Many efforts had been devoted to speed up the calculation of EMD [12,33,35–37]. In [37], Holmes and Taylor used partial signature matching based on the EMD for identifying mammogram structures. They embedded histograms into a learned Euclidean space to speed up computation. In [36], Pele and Werman adopted a thresholded ground distance in EMD computation. The algorithm transformed the flow-network of the EMD so that the number of edges is reduced by an order of magnitude. In [33], Ling and Okada adopted L_1 as the ground distance. They showed that if the points lie on a Manhattan network (e.g. image), the number of variables in the LP problem can be reduced from $O(D^2)$ to $O(D)$. A short survey of other methods suggested for efficient EMD calculation can be found in [12].

Shirdhonkar and Jacobs [12], in another attempt, proposed a new distance metric to approximate original EMD. They applied wavelet decomposition on the dual program of EMD and eliminated the parameters on small waves. In their paper, they proved that the result of optimization of Eq. (1) is approximated very well by

$$d(b,p)_{WEMD} = \sum_{\beta} \alpha_{\beta} |WAV_{\beta}(b-p)| \tag{2}$$

where $WAV_{\beta}(b-p)$ are the wavelet transform coefficients of the D dimensional difference $b-p$ for all shifts and scales β, and $\alpha_{\beta} = 2^{-2*\beta}$ are the scale dependent coefficients. The different underlying metrics are characterized by the choice of different scale weighting and different wavelet kernels. The new distance can be efficiently calculated in linear time with respect to the number of bins in the histograms, while the comparison is about as fast as for normal Euclidean distance or χ^2 statistic. The local minima of Wavelet EMD and are generally co-located, and thus the accuracy of the WEMD approximation of the actual EMD is less important in our algorithm.

4. Our method

4.1. Overview

In this paper, we propose a novel abnormality detection method for crowded scenes using sparse representation for both LAD and GAD. The framework of our approach is summarized in Fig. 3. First of all, feature chosen for LAD or GAD is dependent on the application, and it is concatenated by Multi-scale Histogram of Optical Flow (MHOF), as shown in Fig. 2(a). Then, we conduct dictionary learning based on non-negative matrix factorization, with a constraint of sparsity on the weight matrix. Considering the noise of feature extracted from crowded scenes, the robust Earth Mover's Distance (EMD) is adopted as a distance metric. For saving computation complexity, wavelet EMD is introduced and well integrated into our algorithm for keeping the convexity of the optimization problem. And this is the key part of our algorithm. Given the learned dictionary, a reconstruction weight vector is learned for each query sample and a normality measure is computed from the reconstruction vectors.

4.2. Feature extraction

As in [7], we adopt the Multi-scale Histogram of Optical Flow (MHOF) as a feature descriptor. We first filter out noise optical flow with extremely large amplitude. Our MHOF has $K=24$ bins including three scales, for more precisely preserve motion direction information and motion energy information, as shown in Fig. 2(a). The first scale uses the first 8 bins to denote 8 directions with motion magnitude $r \le T_1$, the second scale uses the next 8 bins with motion magnitude $T_1 < r \le T_2$, while the third scale uses the final 8 with motion magnitude $r > T_2$. After estimating the motion field by optical flow [38], we partition the image into a few $2D$ patches or $3D$ bricks, and then extract MHOF from each unit.

To handle both local abnormality detection (LAD) and global abnormality detection (GAD), we adopt two types of bases with different spatio-temporal structures, which are shown in Fig. 2(b). For GAD, we select the spatial basis covering the whole frame, that contains overall information. For LAD, we extract the spatio-temporal basis that contains spatio-temporal contextual information of surroundings. This type of basis is flexible, you can take different number of units in spatial surroundings and time axis into basis, for different scenes. The final feature vector is concatenated by MHOF from basic units selected.

4.3. Dictionary learning with sparse EMD matrix factorization

In this section, we address the problem of learning dictionary by non-negative matrix factorization. Given a training set of feature pool $B = \{b_1, b_2, \dots, b_N\} \in R^{M \times N}$, where each column vector $b_i \in R^M$ denotes a normal feature vector. Our goal is to find a set of basis (dictionary) $P = \{p_1, p_2, \dots, p_K\} \in R^{M \times K}$ $(K \gg N)$, and a matrix of mixing coefficients $W = \{w_1, w_2, \dots, w_N\}$, $w_i \in R^K$, such that B can be well reconstructed by the weighted sum of computed basis P according to the Earth Mover's Distance. More formally, we formulate the problem as follows:

$$\begin{aligned} &\min_{P,W} \quad \| B - PW \|_{EMD} \\ &\text{s.t.} \quad P \ge 0, W \ge 0 \end{aligned} \tag{3}$$

Here, the EMD between two matrices with N columns is defined as a sum of EMDs between each column in the source matrix and the corresponding column in the target matrix:

$$\| B - PW \|_{EMD} = \sum_{i=1}^{N} EMD\left(b_i, \sum_k w_i^k p_k \right) \tag{4}$$

In this paper, the learnt dictionary is used for a sparse representation of a testing sample or a set of testing samples. As the sparsest representation is certainly appealing, we consider imposing a sparse constraint on the coefficient factor W. l_0-norm is often used as a sparseness measure, and it can be replaced by l_1-norm for the convenience of optimization in the real applications, hence the NMF problem with sparsity constraint of our algorithm can be solved as follows:

$$\min_{P,W} \| B - PW \|_{EMD} + \lambda \| W \|_1 \tag{5}$$

where λ is a regularization parameter. Enforcing sparsity for reconstructing usual events is necessary due to the fact that basis P is learned to maximize the sparsity of reconstruction vectors for abnormal events in the video. On the other hand, for abnormal events, although a fairly small reconstruction error could be probably achieved, a large number of bases should be needed for this reconstruction, resulting in a dense reconstruction weight vector. We need to require the consistency of the sparsity on the solution, i.e. the solution needs to contain some "0" rows, which means that the corresponding features in P are not selected to reconstruct any data samples. We thus substitute the l_1 norm constraint in Eq. (5) with $l_{2,1}$ norm. And the problem is now formulated as

$$\min_{P,W} \| B - PW \|_{EMD} + \lambda \| W \|_{2,1} \tag{6}$$

The problem with the EMD is its expensive computation, which prohibits its applications in many vision problems. To tackle this problem, we introduce approximate EMD [12] into our approach. Another advantage of wavelet EMD is that it can change the optimization problem of our algorithm into a convexity. By replacing the definition of EMD with Wavelet EMD in Eq. (2), the following optimization problem must be solved:

$$\min_{P,W} \sum_{\beta} \alpha_\beta | WAV_\beta (B - PW) | + \lambda \| W \|_{2,1} \tag{7}$$

Next, we show how to solve this optimization problem in Eq. (7), which is not convex. However, if one of the variable matrices P or W is given, the problem becomes linear. Thus, by consecutively fixing either P or W, we can improve the solution by the following two phase in Algorithm 1.

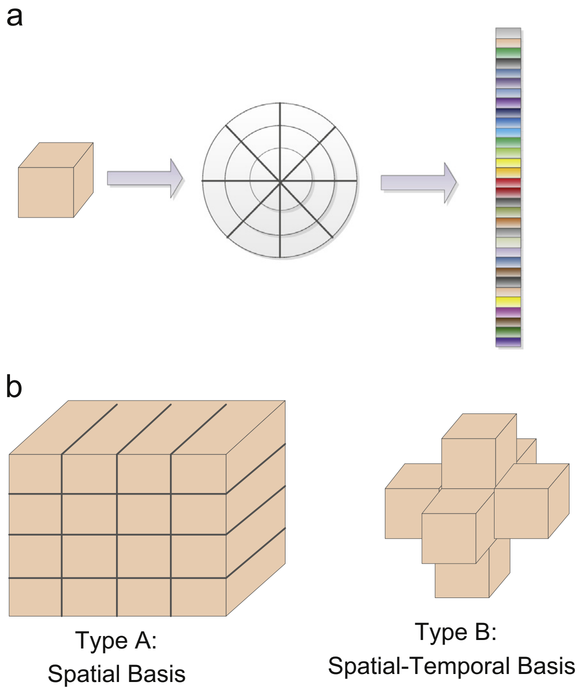

Fig. 2. (a) The multi-scale HOF is extracted from a basic unit ($2D$ image patch or $3D$ brick) with 24 bins. (b) The flexible spatio-temporal basis for sparse representation, concatenated by MHOF from basic units. Type A is tailored to global abnormality detection, and type B is tailored to local abnormality detection.

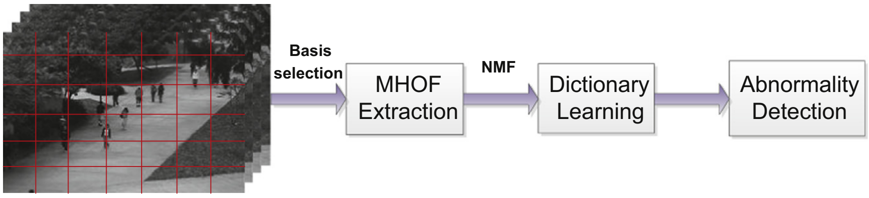

Fig. 3. The framework of our approach.

Algorithm 1. Sparse EMD matrix factorization.

```
Input: The objective matrix $B \in R^{M \times N}$, and an initial value for
   the basis $P^0 \in R^{M \times K}$
$i=0$
repeat
   $i=i+1$
   Find $P^i$ using Eq. (7) with $W$ fixed.
   Find $W^i$ using Eq. (7) with $P$ fixed.
until converge
Output: $P^i$ and $W^i$.
```

The task of finding P^i and W^i in each step in Algorithm 1 is:

$$P^i = \arg\min_P \sum_\beta \alpha_\beta |WAV_\beta(B-PW^{i-1})| \tag{8}$$

$$W^i = \arg\min_W \sum_\beta \alpha_\beta |WAV_\beta(B-P^iW)| + \lambda \| W \|_{2,1} \tag{9}$$

We can solve both minimizations in Eqs. (8) and (9) with gradient based optimization. We consider an objective function $f(w)+g(w)$, where $f(w)$ is convex and smooth and $g(w)$ are convex but not smooth. Here, we elaborate the problem of solving minimization for Eq. (9), because Eq. (8) is a special case of Eq. (9) with $g(w)=0$. The key technique of Nesterove's method [32] is to use

$$p_{Z,L}(w) = f(Z) + \langle \nabla f(Z), w-Z \rangle + \frac{L}{2} \| w-Z \|_F^2 + g(w) \tag{10}$$

to approximate the original function $f(w)$ at the point Z, where L is the Lipschitz constant. At each iteration, we need to solve $\arg\min_w p_{Z,L}(w)$. In our case, we define $f(W) = \sum_\beta \alpha_\beta |WAV_\beta(B-PW)|$, $g(W) = \lambda \| W \|_{2,1}$. However, gradient methods naturally require knowledge of the gradient for the optimization variables. In WEMD, we need to calculate the gradient separately. Consider $h^s = B_m$ and $h^t = (PW)_m$, where m is the column index of matrix. The gradient is

$$\nabla f(W_m) = \sum_\beta \alpha_\beta \, \mathrm{sign}(WAV_\beta(h^s-h^t)) \nabla WAV_\beta(h^s) \tag{11}$$

where the explicit expression for the gradient $\nabla WAV_\beta(h^s)$, with respect to either h^s or h^t, is lengthy but straightforward.

Then we can get the closed form solution [7] of Eq. (10):

$$\arg\min_W p_{Z,L}(W) = D_{\lambda/L}\left(Z - \frac{1}{L} \nabla f(Z) \right) \tag{12}$$

where $\nabla f(Z)$ is defined in Eq. (11), and the transformation function $D_\tau(.)$:

$$W_{i.} = \begin{cases} 0, & \| M_{i.} \| \le \tau \\ (1-\tau / \| M_{i.} \|) M_{i.} & \text{otherwise} \end{cases} \tag{13}$$

where $\tau = \lambda/L$, $M = Z-(1/L)\nabla f(Z)$, $M_{i.}$ is the source data of i's row and $W_{i.}$ is the i's row of matrix computed.

4.4. Abnormality detection

With the dictionary P at hand, we will introduce how to determine a test sample y to be normal or not. As mentioned previously, the features of a normal sample can be linearly constructed by only a few bases in the dictionary P, while an abnormal sample cannot. Now, we are ready to formulate this sparse representation problem as

$$w^* = \min_w \| y - Pw \|_{EMD} + \lambda \| w \|_1 \tag{14}$$

This can be solved by the gradient based method described in above subsection. The $l_{2,1}$-norm is adopted for W during dictionary learning. Here, l_1-norm is adopted for w. The $l_{2,1}$-norm is indeed a general version of the l_1-norm, since if W is a vector, then $\| W \|_{2,1} = \| W \|_1$. After we learn the optimal reconstruction weight vector w^*, we compute the Sparsity Reconstruction Cost (SRC) [7] as follows:

$$S(y, w^*, P) = \| y - Pw^* \|_{EMD} + \lambda \| w^* \|_1 \tag{15}$$

y is detected as an abnormal event if the following criterion is satisfied

$$S(y, w^*, P) > \varepsilon \tag{16}$$

where ε is a user defined threshold that controls the sensitivity of the algorithm to abnormality event.

5. Experimental evaluation

To validate the effectiveness of our proposed algorithm, we conduct experiments on four publicly available datasets: the dataset from University of Minnesota (UMN)[1] and the web dataset [1] are used to test the GAD; the UCSD Ped1 dataset [19] and the Subway dataset [3] are used to test LAD. Details are shown in the following subsections. In wavelet EMD, the standard Mallat filter back algorithm [39] (in Section 7.3.1) for computing the wavelet transforms starts with fine level wavelet coefficients as input. Throughout all the experiments, we empirically set the regularization parameter $\lambda = 2/\sqrt{M}$ in all our experiments (where M is the dimension of feature vector). And parameter β in wavelet EMD is set to 6 in our experiments. The experimental results of the other algorithms are all borrowed from their corresponding papers directly. The ROC curve, based on true false positive and false positive rates on average, is adopted to evaluate performance for global abnormality detection. The EER (equal error rate) and RD (Rate of Detection) is adopted to evaluate performance for local abnormality detection.

5.1. Global abnormality detection

In our experiments, all the video frames from two datasets are resized to a fixed resolution with 480 × 360 pixels. The type A basis in Fig. 2(b) is selected. We compare our method with the pure optical flow based method (denoted as *Optical Flow*), Social Force Model based method in [1] (denoted as *SFM*), Interaction Energy Potential based method [2] (denoted as *LTA*), and Sparse Coding based method in [7] (denoted as *SRC*).

5.1.1. UMN dataset

This dataset comprises the videos of 11 different scenarios of an escape event. The videos are captured in 3 different indoor and outdoor scenes, and the total frame number is 7740. Fig. 4(a) shows some selected frames of these scenes. Each video clip starts with an initial part of normal behaviors and ends with sequences of abnormal behaviors. Scenes in this dataset are crowded, with about 20 people walking around.

We follow the same setup as in [7]. We initialize the dictionary learning from the first 400 frames of each scene, and leave others for testing. Each video is split into 4 × 5 sub-regions, and MHOF from each sub-region is extracted. We then concatenate them to build a basis with a dimension $m=480$. Because the abnormal events cannot occur only in one frame, a temporal smooth is applied. Fig. 4(b) shows the experimental results. The results of *SFM*, *Optical Flow*, *LTA* and *SRC* are directly obtained from their papers [1,7,2]. Although, method [2] achieve good performance in UMN Dataset. However, in more crowded scenario, the tracking accuracy in [2] will greatly affect the anomaly detection

[1] mha.cs.umn.edu/movies/crowd-activity-all.avi

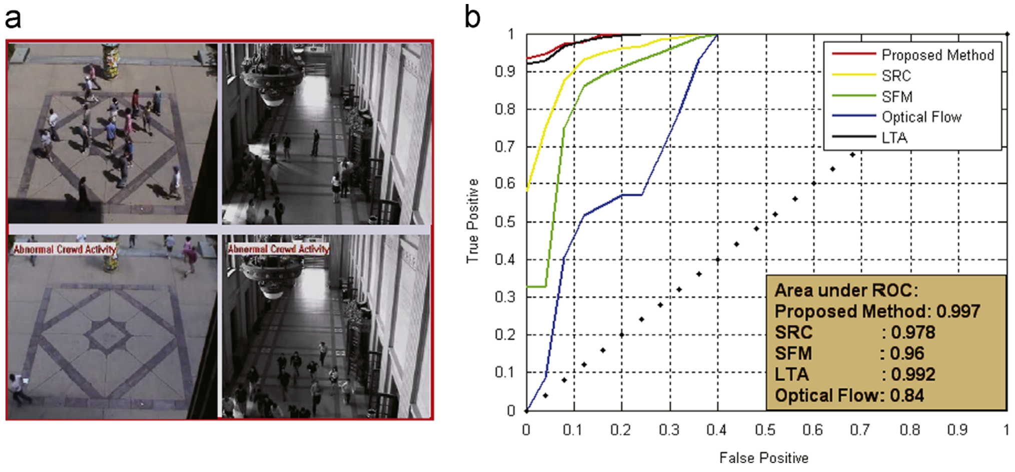

Fig. 4. (a) Top line: samples from normal events; bottom line: samples from abnormal events. (b) The ROCs for abnormality detection on UMN dataset.

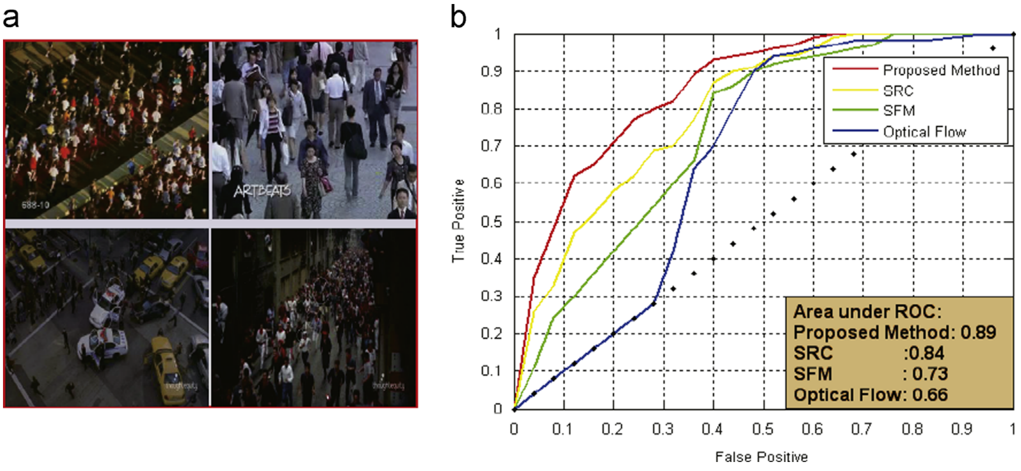

Fig. 5. (a) Top line: samples from normal events; bottom line: samples from abnormal events. (b) The ROCs for abnormality detection on web dataset.

performance. The ROC curves in Fig. 4(b) demonstrate the performance of the proposed method that outperforms these state-of-the-art methods in this dataset. The fail detections happen in the end of some videos.

5.1.2. *Web dataset*

To evaluate the effectiveness of our algorithm, we also conduct an experiment on a more challenging dataset, namely web dataset from [1]. This dataset contains 12 sequences of normal crowded scenes, such as pedestrian walking, marathon running, and 8 scenes of abnormal scenes such as people fighting, escaping. Fig. 5(a) shows some selected frames of these scenes.

The setup is similar to the above subsection. Each video is split into 4×5 sub-regions, and MHOF from each sub-region is extracted. We then concatenate them to build a basis with a dimension $m = 480$. To learn the dictionary, we randomly exclude 2 sequences from the normal set and train on the rest. In the testing phase we add the excluded sequences to the rest. We do this experiment 10 times and construct the ROC by averaging the results of these experiments. In this experiment, our approach fails in classifying the scene of Marathon, as shown in Fig. 5(a). However, the performance of our approach outperforms these state-of-the-art methods, as shown in ROC curves in Fig. 5(b).

5.2. *Local abnormality detection*

To test localization accuracy, detections are compared with pixel level ground-truth masks. If at least 40% of the truly anomalous pixels are detected, the frame is considered to be detected correctly, and counted as a false positive otherwise. For each spatial location in the ROI (Region of Interest, as shown in Fig. 6), we learn a dictionary and use it to determine whether a testing sample is normal or not.

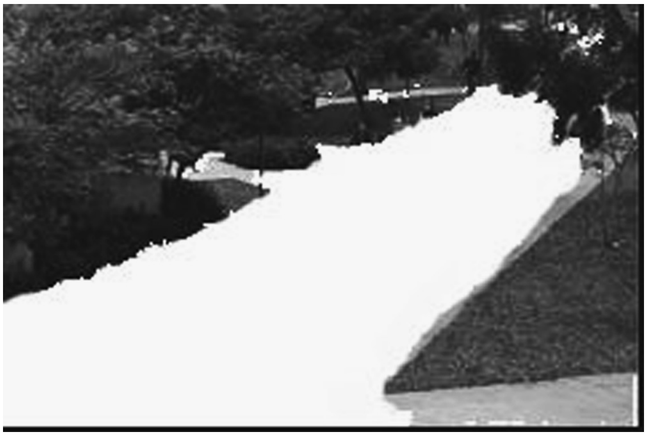

Fig. 6. Region of interest (ROI) of UCSD Ped1 datasets.

5.2.1. *UCSD Ped1 dataset*

The video sequences feature a pedestrian walkway acquired by a stationary camera with low resolution. The crowded density varies

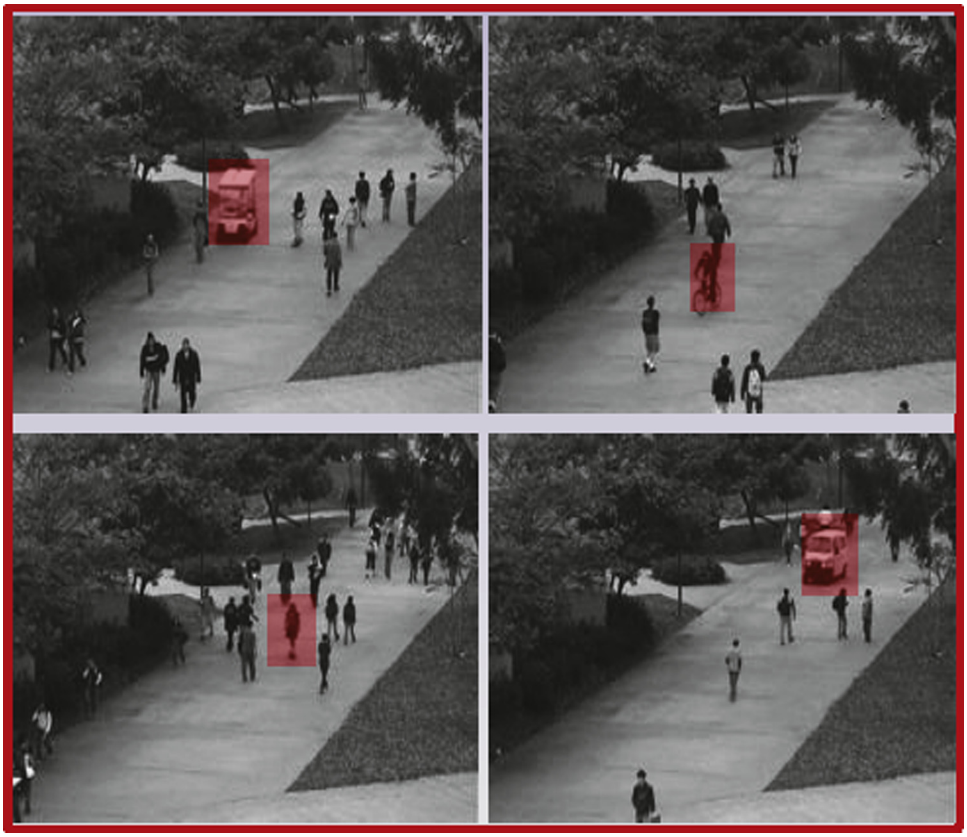

Fig. 7. Abnormal event detections for UCSD Ped1 datasets. The objects such as cars, bicycles, skaters are all well detected. The red masks indicate where the abnormality is. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

and there are numerous sequences that are very crowded and with severe occlusions. Abnormalities are naturally occurring events but are unusual in the present surroundings, e.g. cars running, people cycling, or skaters on the walkway. *Ped1* (resolution 158 × 238) contains 34 training video clips of normal events, and 36 testing clips of various abnormal events. We split each frame into 15 pixels × 15 pixels local pathes with 7 pixels overlapping. The patch size is chosen such that each block does not contain too many objects, which may interfere with one another. The type B basis in Fig. 2(b) is selected to take both local spatial and temporal context information into consideration, with a dimension $m = 7 \times 24 = 168$. A spatio-temporal smooth is adopted here to eliminate noise.

We compare our method with Markov Random Field based method proposed in [6] (denoted as *MPPCA*), Social Force Model based method in [1] (denoted as *SFM*), SRC proposed in [7] (denoted as *SRC*), and MDT proposed in [19]. Fig. 7 shows some selected frames containing abnormality of these scenes. Our algorithm can detect different types of anomalies, such as bicycles, skaters, and cars. In Table 1, some evaluation results which borrowed from [19] are presented: the Equal Error Rate (EER) (ours $15\% < 19\%$ [7], and ours $15\% < 25\%$ [19]), the Rate of Detection (RD) (ours $53\% > 46\%$ [7], and ours $53\% > 45\%$ [19]). It is easy to see that our method outperforms all the other state-of-the-art algorithms.

Table 1
Quantitative comparison of our algorithm with other methods on UCSD Ped1 dataset. EER is equal error rate and RD is rate of detection. The results of MPPCA and SFM are borrowed from paper [19].

Method	EER (%)	RD (%)
MPPCA [6]	40	18
SFM [1]	31	21
SRC [7]	19	46
MDT [19]	25	45
Ours	15	53

5.2.2. Subway dataset

The subway dataset is obtained from Adam et al. [3]. In our experiments, we use the “entrance gate” video (resolution 512 × 384) which is 1 h 36 min long with 1,44,249 frames in total. In our experiments, we resize the frames to 320 × 240 and split each frame into 15 × 15 pixels local pathes with 7 pixels overlapping. The first 15 min are collected to estimate an optimal dictionary. Fig. 2(b) is selected to incorporate both local spatial and temporal information, with a dimension $m=7 \times 24=168$.

We compare our method with the method proposed in [3] (denoted as *Adam*), SRC proposed in [7] (denoted as *SRC*). Fig. 8 shows some selected frames containing abnormality of these scenes. In addition to wrong direction event, the no-payment events are also detected, which are very similar to normal “check in” action. The event-level evaluation is show in Table 2, our method detects all the wrong direction events, and has a higher accuracy for no-payment event, also has a low false alarm, compared with others.

All the experiments are run on a computer with 3 GB RAM and a 3.1 GHz CPU. The average computation time is 1.2 s/frame for GAD, and 4.6 s/frame for UCSD dataset, and 5.1 s/frame for the Subway dataset.

6. Conclusions

In this paper, we propose a novel approach for abnormality detection in crowed scenes. In our algorithm, we conduct the

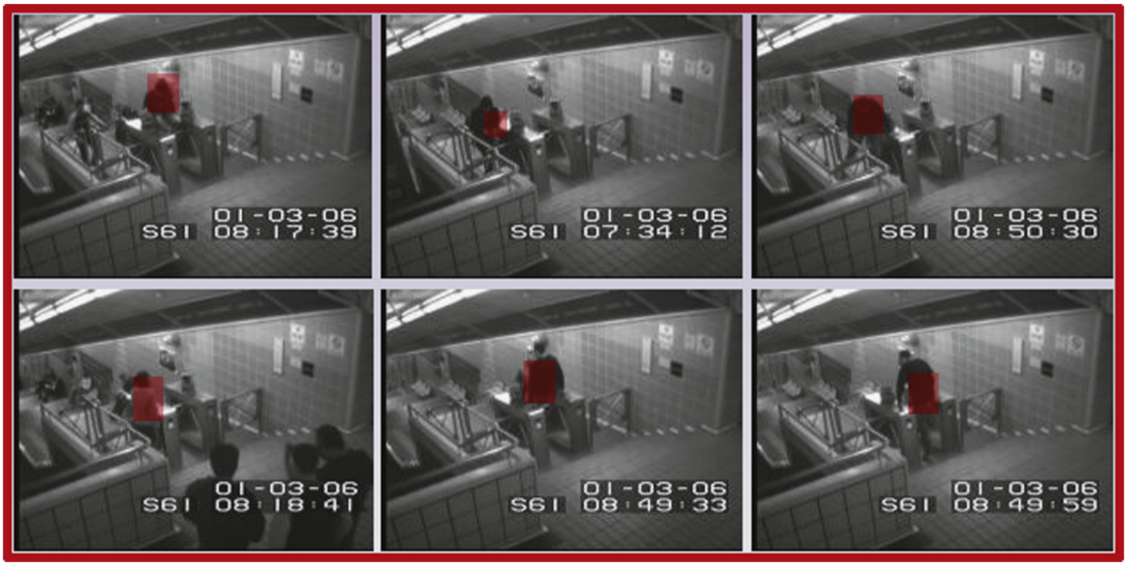

Fig. 8. Abnormal event detections for subway entrance gate. The top row is abnormal events of wrong directions, and the bottom row is abnormal events of no-pay. The red masks indicate where the abnormality is. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 2
Quantitative comparison of our algorithm with other methods on subway entrance video.

Method	Wrong direction	No-pay	Total	False alarm
Ground truth	21	10	31	–
Adam [3]	17	–	17	4
SRC [7]	21	6	27	4
Ours	21	8	29	2

dictionary learning as a task of non-negative matrix factorization problem. For coping with the uncertainty and the noise in complex crowded scenes, we adopt the Earth Mover's Distance (EMD) as an objective function instead of L_2 distance. For degrading the computation complexity of EMD, wavelet EMD is introduced and well combined into our approach to greatly degrade the computation complexity. In addition, wavelet EMD is well combined into our approach to guarantee the convexity of the optimization problem. Due to the flexibility of our basis selection, our method can tackle both local abnormality detection (LAD) and global abnormality detection (GAD). Experimental results on four benchmarks demonstrate the effectiveness of our approach. As our future work, we will attempt to update our dictionary in an online manner, and apply it to some other applications, such as event or action recognition.

Conflict of interest statement

None declared.

Acknowledgements

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61272329, 61273034 and 61070104).

References

[1] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.

[2] X. Cui, Q. Liu, M. Gao, D. Metaxas, Abnormal detection using interaction energy potentials, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3161–3167.

[3] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 555–560.

[4] S. Wu, B. Moore, M. Shah, Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060.

[5] B. Antic, B. Ommer, Video parsing for abnormality detection, in: IEEE International Conference on Computer Vision, 2011, pp. 2415–2422.

[6] J. Kim, K. Grauman, Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2921–2928.

[7] Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3449–3456.

[8] T. Zhang, B. Ghanem, S. Liu, N. Ahuja, Robust visual tracking via structured multi-task sparse learning, Int. J. Comput. Vis. 101 (2013) 367–383.

[9] B. Zhao, F.-F. Li, E.P. Xing, Online detection of unusual events in videos via dynamic sparse coding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3313–3320.

[10] T. Zhang, J. Liu, S. Liu, C. Xu, H. Lu, Boosted exemplar learning for action recognition and annotation, IEEE Trans. Circuits Syst. Video Technol. 21 (2011) 853–866.

[11] M.D. Gupta, J. Xiao, Non-negative matrix factorization as a feature selection tool for maximum margin classifiers, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2841–2848.

[12] S. Shirdhonkar, D.W. Jacobs, Approximate earth mover's distance in linear time, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 23–28.

[13] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. 41 (2009) –.

[14] L. Kratz, K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1446–1453.

[15] Y. Benezeth, P. marc Jodoin, V. Saligrama, C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurrences, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2458–2465.

[16] A. Zaharescu, R. Wildes, Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing, in: European Conference on Computer Vision, 2010, pp. 563–576.

[17] S. Ali, M. Sha, A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[18] J. Cao, DianhuiMao, Q. Cai, H. Li, J. Du, A review of object representation based on local features, J. Zhejiang Univ. Sci. C 14 (2013) 495–504.

[19] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.

[20] V. Saligrama, Z. Chen, Video anomaly detection based on local statistical aggregates, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2112–2119.

[21] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Phys. Rev. E 51 (1995) 42–82.

[22] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 34 (1981) 993–1022.

[23] R. Mehran, B. Moore, M. Shah, A streakline representation of flow in crowded scenes, in: European Conference on Computer Vision, 2010, pp. 439–452.

[24] R. Raghavendra, D. Alessio, C. Macro, M. Vittorio, Optimizing interaction force for global anomaly detection in crowded scenes, in: ICCV Workshops, 2011, pp. 136–143.

[25] D.D. Lee, H.S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999) 788–791.

[26] H. Zhang, Y. Zhang, T.S. Huang, Simultaneous discriminative projection and dictionary learning for sparse representation based classification, Pattern Recognit. 46 (2013) 346–354.

[27] Y. Li, A. Ngom, Supervised dictionary learning via non-negative matrix factorization for classification, in: IEEE Conference on Machine Learning and Applications, 2012, pp. 439–443.

[28] M. Heiler, C. Schnorr, Learning sparse representations by non-negative matrix factorization and sequential cone programming, J. Mach. Learn. Res. 7 (2006) 1385–1407.

[29] M. Heiler, C. Schnorr, Non-negative matrix factorization with sparseness constraints, J. Mach. Learn. Res. 5 (2004) 1457–1469.

[30] G.Z. end Elisa Ricci, N. Sebe, Exploiting sparse representations for robust analysis of noisy complex video scenes, in: European Conference on Computer Vision, 2012, pp. 199–213.

[31] G. Zen, E. Ricci, Earth mover's prototypes: a convex learning approach for discovering activity patterns in dynamic scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3225–3232.

[32] Y. Nesterov, Gradient methods for minimizing composite objective function, in: Technical Report, 2007.

[33] H. Ling, K. Okada, An efficient earth mover's distance algorithm for robust histogram comparison, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 840–853.

[34] Y. Rubner, C. Tomasi, L. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2000) 99–121.

[35] M. Werman, S. Peleg, A. Rosenfeld, A distance metric for multidimensional histograms, Comput. Vis., Graph., Image Process. 32 (1985) 328–336.

[36] O. Pele, M. Werman, Fast and robust earth mover's distance, in: IEEE International Conference on Computer Vision, 2009, pp. 460–467.

[37] A.S. Holmes, C.J. Rose, C.J. Taylor, Transforming pixel signatures into an improved metric space, Image Vis. Comput. 20 (2002) 701–707.

[38] M. Black, P. Anandan, The robust estimation of multiple motions: parametric and piecewise-smooth flow fields, in: Computer Vision and Image Understanding, 1996, pp. 75–104.

[39] S. Mallat, A Wavelet Tour of Signal Processing, second edition, Academic Press, 1998.

Xiaobin Zhu received his B.S. degree in 2003 from the Hangzhou Institute of Electronic Engineering, Hangzhou, China. And he received his M.E. degree in 2006 from Beijing Normal University. He is currently pursuing Ph.D. degree in Institute of Automation, Chinese Academy of Sciences. His research interests include machine learning, video analysis and object tracking, etc.

Jing Liu received her B.E. degree in 2001 and M.E. degree in 2004 from Shandong University, and her Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2008. Currently she is an associate professor in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include machine learning, image content analysis and classification, multimedia information indexing and retrieval, etc.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently an Assistant Professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Changsheng Li received his B.S. degree in 2008 from University of Electronic Science and Technology of China. He is currently pursuing Ph.D. degree in Institute of Automation, Chinese Academy of Sciences. His research interests include machine learning, big data, multimedia content analysis, etc.

Hanqing Lu received his B.E. degree in 1982 and his M.E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China in 1992. Currently he is a Professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, object tracking, recognition and image retrieval, etc. He has published more than 200 papers in those areas. He is a senior member of the IEEE.