Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection

Huamin Ren[1]
hr@create.aau.dk
Weifeng Liu[2]
weifeng.liu@nbi.ku.dk
Søren Ingvor Olsen[3]
ingvor@di.ku.dk
Sergio Escalera[4]
sergio@maia.ub.es
Thomas B. Moeslund[1]
tbm@create.aau.dk

[1] Department of Architecture, Design and Media Technology
Aalborg University
Aalborg, Denmark

[2] Niels Bohr Institute
University of Copenhagen
Copenhagen, Denmark

[3] Department of Computer Science
University of Copenhagen
Copenhagen, Denmark

[4] Dept. Applied Mathematics, University of Barcelona, Computer Vision Center
Barcelona, Spain

Abstract

Abnormal event detection has been an important issue in video surveillance applications. Due to the huge amount of surveillance data, only a small proportion could be loaded during the training. As a result, there is a high chance of incomplete normal patterns in the training data, which makes the task very challenging. Sparse representation, as one of solutions, has shown its effectiveness. The basic principle is to find a collection (a dictionary) of atoms so that each training sample can only be represented by a few atoms. However, the relationship of atoms within the dictionary is commonly neglected, which brings a high risk of false alarm rate: atoms from infrequent normal patterns are difficult to be distinguished from real anomalies. In this paper, we propose *behavior-specific* dictionaries (BSD) through *unsupervised* learning, in which atoms from the same dictionary representing one type of normal behavior in the training video. Moreover, 'missed atoms' that are potentially from infrequent normal features are used to refine these behavior dictionaries. To further reduce false alarms, the detection of abnormal features is not only dependent on reconstruction error from the learned dictionaries, but also on non zero distribution in coefficients. Experimental results on Anomaly Stairs dataset and UCSD Anomaly dataset show the effectiveness of our algorithm. Remarkably, our BSD algorithm can improve AUC significantly by 10% on the stricter pixel-level evaluation, compared to the best result that has been reported so far.

1 Introduction

Abnormal event detection is an important application in video surveillance, which aims to detect potential abnormal behaviors and has an important impact on daily life. A common

solution is to learn normal patterns from the training data first, and then recognize anomalies by measuring their similarities/reconstruction errors based on the learned normal model. Such normal models could rely on trajectory [15], feature distribution density [1], graph model [19] and sparse representation [7, 21, 36, 38, 42], which has reported the best performance.

Sparse representation has gained a great deal of attention since being applied effectively in many image analysis applications and many theoretical achievements are obtained [9, 13, 17, 35, 37, 40]. Its success stems from the discovery of underlying properties from low-level to mid-level human vision: many neurons in the visual pathway are selective for a variety of specific stimuli, such as color, texture, orientation, scale, and even view tuned object images [31]. Existing research on sparse representation can be generally divided into two groups - sparse coding [3, 4, 10, 11, 29] and dictionary learning [7, 16, 21, 35]. Sparse coding deals with finding coefficients for a given dictionary, which requires that each signal be represented sparsely. Dictionary learning, on the other hand, aims to find suitable basis vectors that construct the dictionary. Dictionary learning is vital to abnormality detection, since the dictionary matrix is loaded and involved in anomalies detection, unfortunately, it has gained relatively less attention compared to sparse coding. In this paper, we focus on how to build an optimized dictionary to meet the needs of video abnormality detection.

A dictionary is constructed by a basis (composed of multiple basis vectors), in which each basis vector is called an `atom`. A concatenation of dictionaries is called a `frame` [39]. Supposing we could obtain a frame with all the normal patterns, i.e., each dictionary in the frame contains all the atoms needed to represent a particular behavior, we could then distinguish normal features from anomalies, since normal features could be represented by only a few atoms from the same behavior while an anomaly could not and hence be detected. The question, then, becomes: how to assemble such a frame, especially when the types of normal patterns are unknown?

It is still an open issue, in which finding atoms is crucial. There is consensus that learning atoms from visual features representing original images is a more generalized approach [12, 26], compared to using base functions. However, the question of how to find atoms that best suit particular application involves ongoing research. Learning discriminative dictionaries while achieving multiple classifiers has been on the research agenda recently [16]. Despite good classification results in object recognition and action recognition, however, they cannot be directly applied to abnormality detection. Their prerequisite of class labels in the training data hampers their applicability in video surveillance applications: it is common to have only normal videos for training in video surveillance; therefore, there is only one type of label available. Moreover, due to the huge amount of surveillance data, only a small proportion could be loaded during the learning, which means that training videos could contain incomplete normal patterns.

To meet the needs of dictionary learning in abnormality detection, we propose an algorithm to learn *behavior-specific* dictionaries (a frame) through *unsupervised learning*, in which each dictionary represents a single type of normal behavior in training videos. Such behavior-specific dictionaries are further refined with 'missed atoms', which refers to atoms from infrequent normal patterns. During detection, abnormal features are recognized based on their reconstruction error from the learned frame, as well as non zero distribution in their coefficients to reduce false alarms.

This paper presents two main contributions. First, we take into consideration the relation of atoms in one dictionary without prior knowledge, which, to the best of our knowledge, has not been done before. Based on this notion, we propose a dictionary in which atoms in the

same dictionary are related to one type of behavior. Second, we propose an algorithm to find atoms with compact non zero coefficients (called missed atoms) in order to better distinguish between anomalies and infrequent behaviors.

2 Related Work

How to obtain good dictionaries and then form a frame? One line of work considers pre-constructed dictionaries, *e.g.*, undecimated wavelets [14], contourlets [8], curvelets [6], the wavelet packets, or the bandelets [33]. In spite of their fast transforms, those approaches are restricted to signals or images of a certain type and cannot be used for a new and arbitrary family of signals of interest [12]. Learning the dictionary instead of using an off-the-shelf basis or a combination of bases has been shown to dramatically improve signal reconstruction. Atoms in the learned dictionaries are tuned to the input image or signals to obtain a small reconstruction error, leading to much better results in practice [25]. Later research focuses on learning discriminative sparse models instead of purely reconstructive ones, such as [23, 24].

One promising dictionary is the overcomplete dictionary, in which the number of basis vectors is greater than the dimension of the input. These basis vectors could be either from one dictionary or a concatenation of basis vectors from multiple dictionaries. To avoid ambiguity, we call the former an overcomplete dictionary, and the latter an overcomplete frame. Overcomplete dictionaries/frames are preferred because they can lead to sparser coefficients and have greater flexibility in matching structure in the data. More importantly, they have proven robust in the presence of noise [20]. K-SVD [2, 28] is a representative dictionary, which handles atoms sequentially to minimize the reconstruction error. More recent researches have attempted to improve discriminativeness of dictionaries based on K-SVD by iteratively updating dictionary atoms based on the results of a linear predictive classifier [34], or by unifying the dictionary and classifier learning process [41]. Nevertheless, the sequence generated by the K-SVD method does not promise global convergence. Bao et al. [5] proposed a *l0* norm dictionary learning algorithm by proximal methods, which promises a convergence to a stationary point with a sub-linear convergence rate.

Despite of the progress these dictionary algorithms have achieved, it is difficult to apply them directly on abnormal event detection, which is mainly due to the unavailability of labels - only normal videos are used as training data due to practical reasons. Recent work propose their dictionaries based on sparse representation, which are designed for abnormality detection purpose and have proven effectiveness, *e.g.*, [7, 21]. In these two approaches, an overcomplete dictionary [7] or frame [21] as well as a sparse coefficient matrix are obtained during the learning based on visual features. A testing feature is identified as an anomaly if its reconstruction error from the dictionary is larger than a threshold. The major difference between them is how to find the atoms. In [7], the basis vectors are a subset of the visual feature pool, so other features could be sparsely and linearly represented by a few visual feature atoms. In [21], they code it directly as a set of possible combinations of basis vectors, and each combination corresponds to a set with basis vectors from multiple dictionaries. During the detection, features with large reconstruction errors are detected as anomalies. However, the relationship of atoms does not contribute to the final detection, which means an anomaly is detected on whether it can be represented by a few atoms or not, no matter how far away representing atoms are. In that case, an infrequent feature is difficult to be distinguished from real anomalies, hence has a high false alarm rate.

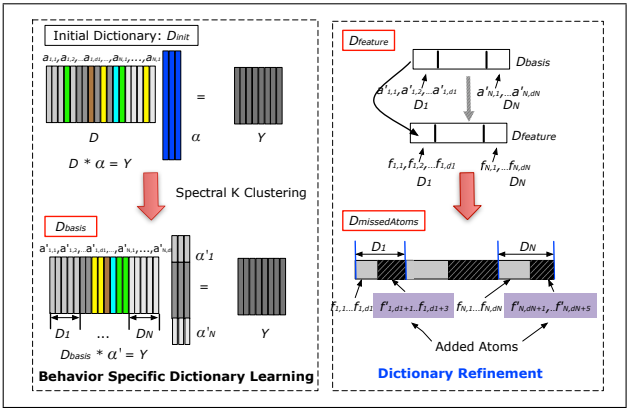


Figure 1: Behavior-specific dictionary learning (left) and refinement procedure (right), illustrated in Sec. 3 and Sec. 4, respectively.

3 Behavior-Specific Dictionary (BSD) Learning

Unlike most existing dictionary learning algorithms, which either learn multiple dictionaries through supervised learning, or learn multiple dictionaries neglecting relationship of atoms within each dictionary, our algorithm explores the relationship of atoms to help constructing behavior-specific dictionaries under unsupervised learning. Our algorithm is illustrated in Fig. 1. It contains two parts: 1) behavior-specific dictionary learning, in which atoms from the same behavior pattern are found to represent that behavior; and 2) dictionary refinement, which improves the dictionary by finding 'missed atoms' of each behavior.

Given a set of training features $\{Y_i \in \mathbb{R}^p\}_{i=1}^n$, which form a feature space F, we assume that there is an overcomplete dictionary that gives rise to the given training samples via sparse combinations, i.e., we assume that there is a matrix $D \in \mathbb{R}^{p\times m}$ with $p \ll m$, so that the training set Y could be represented by the dictionary D and its coefficient $\alpha \in \mathbb{R}^m$. We shall hereafter assume that D is a full-rank matrix to ensure that there are solutions to this underdetermined system, i.e., its columns span the entire space $\mathbb{R}^p$. $\|\cdot\|_0$ norm constraint (the number of non zeros) is applied on the coefficients, hence the objective is an optimization solution to Eq. 1, which preserves reconstructiveness of the training data and sparsity of their coefficients. Note that all vectors in this paper are column vectors.

$$\min ||Y - D\alpha||_2 + ||\alpha||_0. \tag{1}$$

For our behavior-specific dictionaries D (a frame, $D = [D_1, D_2, \cdots, D_i, \cdots, D_N]$, where N denotes the number of dictionaries), it should satisfy the above requirements while preserving two desirable properties: 1) for each D_i of d_i-dimension, the information of a particular behavior is preserved with atoms $[a_{i,1}, \cdots, a_{i,j}, \cdots, a_{i,d_i}], a_{i,j} \in \mathbb{R}^p$, and for this dictionary, 2) the feature with a similar behavior pattern should be finely reconstructed using a few atoms of the dictionary. To formulate our objective, we define $\Lambda_j \subset \{1, \cdots, m\}$ as a subset of indices indicating the index of each atom in the dictionary D_j; for example, $\Lambda_2 = \{10, 11, 12\}$ means that there are 3 atoms in the second dictionary, referring to 10^{th},11^{th} and 12^{th} columns of D. With the help of Λ_j, the coefficient of training sample Y_i to each dictionary could be easily denoted as $\alpha_{\Lambda_j,i}$. Then, our problem of behavior-specific frame construction can be

formulated as Eq. 2, where $\|\alpha_{\Lambda_j,i}\|_0$ is an accumulated number of non zeros corresponding to the j^{th} dictionary in the coefficient vector of Y_i.

$$\begin{aligned} &\arg\max_j \frac{\|\alpha_{\Lambda_j,i}\|_0}{\sum_{k=1}^{N}\|\alpha_{\Lambda_k,i}\|_0}, j=1,2,\cdots,N \\ &s.t. \|Y_i - D_j\alpha_{\Lambda_j,i}\|_2 \leq \varepsilon, \|\alpha_{\Lambda_j,i}\|_0 \leq T, i=1,2,\cdots,n. \end{aligned} \tag{2}$$

Let T denote the degree of sparsity in coefficients, ε be the reconstruction error due to the noise that exists in images, n be the number of training samples, and N be the number of dictionaries. By solving Eq. 2, each training feature should be reconstructed, while, at the same time, its coefficients should be as sparse as possible. Note that reconstructiveness and sparsity in Eq. 2 is different from existing algorithms: it should ideally be reconstructed only by a few atoms in its belonging dictionary, which means non zeros in coefficients only exist in this dictionary, implying there are more zeros over the frame compared to existing dictionaries. It is important to note that an even stricter constraint is put on the objective: a block-wise property in the coefficient, as well as the capability of being reconstructed with sparse coefficients, i.e., for each visual feature assigned to the dictionary D_j, the majority of non zeros in its coefficient $\alpha_{\bullet,i}$ only exist in the location corresponding to the dictionary D_j. However, it is potentially very difficult to solve this problem. Ordinary solutions to an optimization problem with multiple parameters can not be applied in Eq. 2. The main reason is that ordinary solutions try to solve one parameter by fixing others, and then optimize these fixed parameters one by one with the learned parameter. However, the two constraints in Eq. 2 - sparsity in coefficients and block wise non zero distribution - could cause a location interaction, which complicates the problem.

Rather than optimize two interacting parameters simultaneously, we learn our dictionary based on an initial dictionary D_{init}, which can be any $\|\cdot\|_0$ solution of the convex optimization problem: $\|y - D\alpha\|_2 \leq \varepsilon, \|\alpha\|_0 \leq T$. Therefore, sparsity and reconstructiveness are achieved in this initial dictionary. We then run a clustering algorithm on atoms in D_{init} to categorize them into N groups, where each group presents one possible behavior pattern and atoms within the group form a new basis. Through this, a roughly behavior-specific frame D_{basis} is constructed and its refinement is illustrated in Section 4. Fig. 1 (left) shows our rough behavior-specific dictionary learning procedure. Concretely, we use the K-SVD algorithm [2] to generate a initial dictionary due to its simplicity and effectiveness. We then treat each atom in D_{init} as a sample and apply spectral clustering [30] to segment them into N groups. Next, atoms in each group are reordered based on their reconstruction error to represent a basis for a dictionary D_j. Finally, a frame D_{basis} is formed by concatenating N dictionaries: $D_{basis} = [D_1, D_2, \cdots, D_N]$. Note that each dictionary D_i is not necessarily overcomplete, but the combination of dictionaries should be overcomplete, i.e., $p \ll \sum_i d_i$.

4 Dictionary Refinement

The sparse coding via the frame D_{basis} cannot guarantee a block wise non zero distribution; therefore, we search for atoms from infrequent visual information, and supplement them as 'missed atoms' to better express the behavior, and meanwhile to achieve this block wise property. We first transfer atoms in D_{basis} into its original feature space, because the atoms

that stem from visual features are more interpretable and have proven effective (see [7] for details). Each feature y can be assigned to a dictionary D_j by calculating Eq. 3, and features that belong to the same dictionary are regarded as a new basis. In this way, a new frame is formed and is denoted as $D_{feature}$. To keep the same size dictionaries in two frames, we reorder features in each dictionary in descending order by their reconstruction error, and only choose first d_i features as basis vectors. Then coefficients based on $D_{feature}$ can be obtained by using sparse coding algorithms, such as OMP algorithm [32]. In this setting, we seek optimal behavior-specific dictionaries.

$$\begin{aligned} &\arg\min_j R_j, j = 1, \cdots, N \\ &R_j = \|y - D_j \alpha_j\|_2. \end{aligned} \tag{3}$$

Note that atoms in $D_{feature}$ are selected if they satisfy a small reconstruction error as well as sparsity (see Eq. 2). The constraint of this sparsity only ensures that there is a small number of non zeros in the coefficients, but has nothing to do with their location. For this reason, we search for coefficients that already have a block wise sparsity - and then treat them as 'missed atoms' to the existing dictionary. We call them 'missed atoms' because these atoms have a high response with some atoms in the same dictionary but also have a large reconstruction error. This implies that some atoms in the current dictionary are missing, which could be infrequent visual information and difficult to track. Therefore, we introduce these atoms and add this missing information. Let x denote the coefficient vector of the sample Y_i, i.e $x = \alpha_{\bullet,i}, i \in \{1,2,\cdots,n\}, x \in \mathbb{R}^m$, g_i indicates the closest dictionary to the sample, and Λ_{g_i} the index of the basis in $D_{feature}$ for the g_i^{th} dictionary. For example, given $g_{10} = 1$ and $\Lambda_{g_{10}} = \{1,2,3,4\}$, we could easily figure out that the training sample Y_{10} belongs to dictionary D_1, and the first 4 columns in the concatenated matrix $D_{feature}$ are its atoms. Define $\{supp(x)|supp(x) = k : x_i \neq 0\}$ as the support of x. The number of concentrated coefficients is more interesting than the cardinality of coefficients; therefore, we use $|supp_i(x)|$ to denote the the number of non zeros of x corresponding to the i^{th} dictionary. We call a coefficient a compact coefficient if it satisfies:

$$\frac{|supp_{\Lambda_{g_i}}(x)|}{\sum_{i=1}^{N} |supp_i(x)|} > th, \tag{4}$$

where th is a threshold to control the degree of non zero concentration, N is the number of dictionaries, and Λ_{g_i} is the coefficient entries corresponding to the closest dictionary D_{g_i} of the training sample. A compact coefficient implies that the majority of non zeros concentrates on one dictionary D_i. We treat a sample as a 'missed atom' if its coefficient is compact and its reconstruction error based on D_i is the smallest among other dictionaries. This procedure is illustrated in Algorithm 1 with more details.

Finally, we detect abnormal features based on their reconstruction error R from the learned frame. In contrast to Lu13 [21] and Sparse [7], we only return the reconstruction error corresponding to the dictionary where the non zeros in its coefficient concentrate - if no concentration is found, the reconstruction error over all dictionaries is returned. See E.q. 5, R_i refers to reconstruction error of the test sample y_i.

$$R_i = \begin{cases} \|D_{\bullet,\Lambda_{g_i}} \cdot \alpha_{\Lambda_{g_i},i} - y_i\|_2^2, & \text{if E.q. 4 is satisfied,} \\ \|D \cdot \alpha_{\bullet,i} - y_i\|_2^2, & \text{otherwise.} \end{cases} \tag{5}$$

Algorithm 1 Behavior-Specific Dictionary Refinement

Input: $y \in \mathbb{R}^p, N, m, n, D_{init} \in \mathbb{R}^{p\times m}(D_{init} = [D_1, \cdots, D_N])$, $\alpha \in \mathbb{R}^{m\times n}$, $D_{basis} \in \mathbb{R}^{p\times m}$
Transfer basis vectors into feature space
for each y_i, find its dictionary index g_i
for i=1 : N
Find indices of samples in dictionary D_i: Λ_i
Calculate $R = \|Y_{\Lambda_i} - D_{init}\alpha\|_2$
Order R descendingly
Find first d_i features as a basis of D_i
end for
$D_{feature}$ is represented as: $[D'_1, \cdots, D'_N]$
Updating $D_{feature}$ **with new features**
Find features y' with compact coefficients
Add y' as missed atoms
$D_{missedAtoms}$ is represented as:
$[D'_1, D'_{missedAtoms1}, \cdots, D'_N, D'_{missedAtomsN}]$
Output: $D_{feature}, D_{missedAtoms}$

5 Experiments

5.1 Datasets and Settings

We carry out experiments on two datasets: UCSD Ped1 dataset [22] and Anomaly Stairs dataset. UCSD Ped1 is a frequently used public dataset on abnormal event detection. It includes clips of groups of people walking towards and away from the camera with some amount of perspective distortion. Abnormal behaviors in testing videos are either non-pedestrian entities in the walkways or anomalous pedestrian motion patterns. However, there are not many obvious patterns in the training data; therefore, it is difficult to estimate the effectiveness of algorithms on real surveillance applications, in which the training data could only be a very small proportion of the surveillance data and patterns are incomplete. For this reason, we build a RGB dataset: Anomaly Stairs dataset at the stairs case. There are 9 types of normal behaviors performed by 12 persons, which are further divided into four groups. Each group contains 3 actors/actresses, performing normal and abnormal behaviors with different postures. Normal behaviors include: single person (groups of persons) walking and running upstairs and downstairs and one person going upstairs while another person is going downstairs. Abnormal behaviors include slipping or falling. See representative normal and abnormal images in Fig. 2.

We divide frames in UCSD Ped1 into 23 $\times$ 15 patches, and use 5 consecutive frames to form spatial-temporal cubes. We then compute 3D gradient features on these cubes, following the setting in [21, 22]. One thing to mention is that they resize the image into three resolutions and detect features in three layers. In order to speed up online extraction and detection and better meets the needs of real surveillance applications, we simplified the feature with only one layer. Feature dimension is reduced to 100 using PCA from initial 500. For the Anomaly Stairs dataset, we resize frames to 480 $\times$ 219 and extract 3D gradient features of cubes on each 48 $\times$ 21 patch.

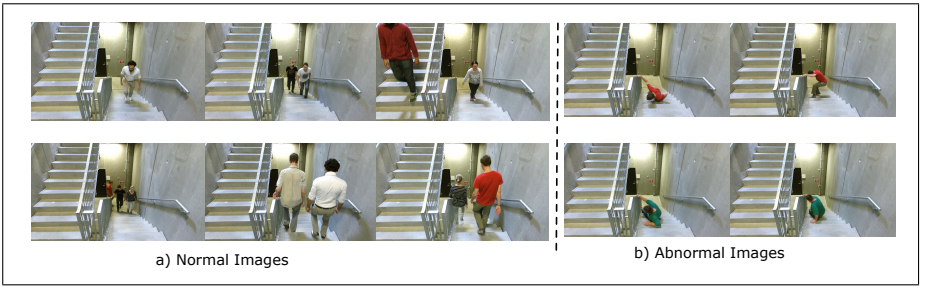


Figure 2: Representative normal and abnormal images on Anomaly Stairs dataset. The resolution is 1920×1076.

5.2 Dictionary Validation

We first evaluate our BSD algorithm in terms of non zero concentration for normal features and non zero spreading over for abnormal features in coefficients. We then compare the performance of different dictionaries when varying the size of the training data.

First, we randomly select one normal frame and one abnormal frame from the UCSD Ped1 dataset, pick 25 features in each frame, and use the OMP algorithm [32] to generate their coefficients according to our three behavior-specific dictionaries - D_{basis}, $D_{feature}$ and $D_{missedAtoms}$ - and compare these with coefficients based on Lu13 [21], which is also a combination of dictionaries. For each feature, we accumulate non zeros corresponding to each dictionary in its coefficient, and normalize the accumulated number to [0 1], which is used to draw a gray image. A white dot implies that there are non zeros in this dictionary, and the whiter the pixel in the image is, the more non zeros in the coefficient. Observing Fig. 3, the y-axis refers to dictionary id (20 dictionaries in total), and the x-axis refers to the feature id. A point (x_i, y_j) refers to the normalized number of non zeros in the coefficient of x_i to the j^{th} dictionary. For our three BSD frames, normal features tend to have concentrated non zero coefficients, and abnormal features tend to spread out over dictionaries. For the normal coefficients, We can see that there are some non zeros spreading out in D_{basis}, $D_{feature}$, which are then improved by introducing missed atoms in $D_{missedAtoms}$. This property is mainly attributed to the missed atoms. Despite of their relatively large reconstruction errors, their high response in some representative atoms also implies that they could be from infrequent normal patterns. By adding them as new atoms, normal features that could be linearly represented by these representative atoms could also be represented by the newly added atoms and, as a result, more non zeros will appear in this dictionary. Due to the sparsity constraint, fewer non zeros will show in the remaining dictionaries. Atoms in Lu13 [21] are not related; therefore, non zeros spread out in both normal and abnormal features.

Next, we compare our BSD algorithm with two dictionaries when varying the size of the training set on the Anomaly Stairs Dataset: Bao14 [5] and Lu13 [21]. They are selected because of their proven effectiveness either in abnormal event detection or in a general sense. We vary the training set by tuning two factors: patterns and samples. Each group contains normal patterns performed by 3 actors/actresses; therefore, the pattern can be controlled by varying the number of groups, and for each group, samples can be varied by changing its size. To evaluate, we apply an event-level evaluation, i.e., an anomaly testing video is reported if an accumulated anomaly score in consecutive frames is larger than a predefined threshold. Results are shown in Tab. 1. Taking the first row for example, only 1% of the nor-

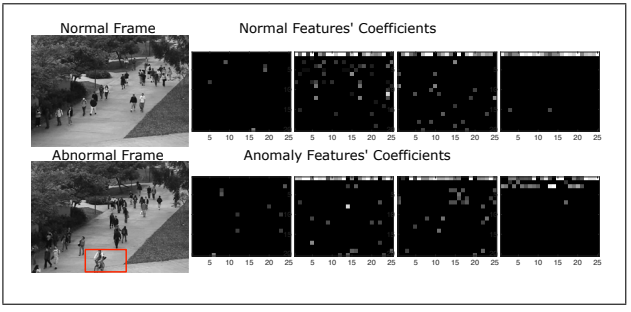


Figure 3: Non zero distribution in coefficients of features from normal and abnormal frames. From left to right: Lu13 [21] and our three frames: D_{basis}, $D_{feature}$ and $D_{missedAtoms}$. Accumulated non zeros in each dictionary are normalized in [0,1]; therefore, the whiter, the more non zeros.

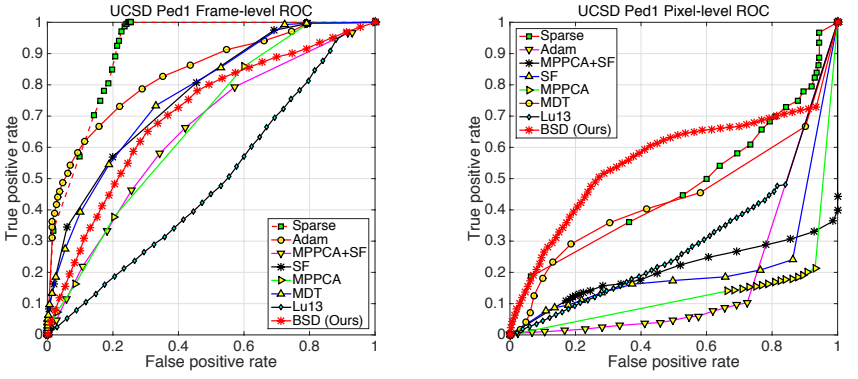
	# Correctly detected abnormal videos (TP)			# False alarms (FA)		
Pattern and training size	Lu13 [21]	Bao14 [6]	BSD (ours)	Lu13 [21]	Bao14 [6]	BSD (ours)
One group (1%)	11	12	**15**	**13**	16	14
One group (10%)	**17**	12	**17**	21	16	**15**
Two groups (1%)	14	13	**17**	**15**	19	17
Two groups (10%)	14	**19**	15	17	18	**12**
Three groups (1%)	14	**18**	**18**	**15**	19	16
Three groups (10%)	14	**19**	**19**	**17**	18	**17**
Average TP / FA	14.0	15.5	**16.8**	16.3	17.7	**15.2**
Average TPR / FPR	70.0%	77.5%	**84.2%**	77.8%	84.1%	**72.2%**

Table 1: Results on Anomaly Stairs dataset by varying the size of the training data.

mal features in group one (normal behaviors are performed by 3 actors) is used for learning. The test videos contain 41 clips, among which 21 are normal and 20 are abnormal. Note that the test videos contain abnormal behaviors performed by the total 12 actors/actresses. Therefore, dictionaries tested in our experiments are challenged by incomplete training patterns. The same empirical threshold is set on event-level evaluation in three algorithms, and then correctly detected abnormal videos (true positives, TP) and false alarms (FA) are reported. When only limited normal patterns and a small number of samples (1%) are used for training, our BSD algorithm performs best in terms of TP and FA. All of these algorithms are improving when more patterns and samples are learned. Averaged True Positive Rate (TPR) and False Positive Rate (FPR) are also reported to obtain comparable results. As can be seen, our BSD achieves the best detection with the least false alarms - our averaged TPR is 84.2%, compared to Lu13 [21](70.0%) and Bao14 [6](77.5%); our averaged FPR is 72.2%, much smaller than Lu13 [21] (77.8%) and Bao14 [6](84.1%). Anomaly Stairs dataset is very challenging, since slippery happens during the running, some people respond very quickly, and it is very difficult to judge whether they are running or slipping. That is why all three dictionaries have relatively high false alarms.

5.3 Abnormal Event Detection

We compare our BSD learning algorithm with state-of-the-art abnormal event detection algorithms, i.e., SF [27], MPPCA [18], SF-MPPCA [22], Adam et al. [1], MDT [22], Sparse [2], and Lu13 [21]. Note that only the source code of Lu13 [21] is available to us, the remaining data are extracted from their published papers. Following the setting in [22], we report frame-level and pixel-level evaluations in Fig. 4. Abnormal frames are marked as positives.



(a) Comparison on frame-level evaluation.

(b) Comparison on pixel-level evaluation.

Figure 4: Comparative ROC curves on frame-level 4(a) and pixel-level 4(b) evaluation with state-of-the-art methods on UCSD Ped1 dataset. Note that frame-level evaluation do not take the location of anomaly feature into considerations, therefore there is a chance of coincide, for example, a normal feature in an anomaly frame is wrongly detected as anomaly, which contributes to a high ROC score. More accurate detection evaluation is shown in Fig. 4(b).

On frame-level evaluation, a frame is marked as a positive if at least one abnormal feature is found; while on pixel-level evaluation, stricter criteron is applied - a frame can only be marked as a correctly detected positive if at least 40% of truly abnormal features are reported. Frame-level evaluation has a risk of coincidence - if a normal feature in an abnormal frame is incorrectly detected as an anomaly, it still contributes to a higher AUC score. Due to this reason, pixel-level criterion is stricter but more accurate. Our BSD obtains a AUC of 56.17% on pixel-level evaluation, which achieves the best result (promoting AUC by 10% compared to the best score (46.1%) published so far [7]), yet has a satisfactory result on frame-level evaluation (AUC is 70.69%).

6 Conclusion

We proposed an unsupervised behavior-specific dictionary learning algorithm. Each dictionary contains atoms representing a particular normal behavior. By finding 'missed atoms', the dictionary is further improved to avoid high false alarms. As a result, all the dictionaries are combined as a frame to represent normal behaviors in the training data, and anomalies are detected based on their non zero distribution in coefficients and reconstruction error. The experiments proved the effectiveness of the proposed method, especially when the data available during the training is limited.

7 Acknowledgement

This work is sponsored by the Danish National Advanced Technology Foundation (HTF). We give our thanks to every member in VAP lab for their assistance in recording Anomaly Stairs dataset, especially to Chris Bahnsen. We also give our sincere appreciation to Andreas Møgelmose for his suggestions to our early version and all reviewers for their valuable comments.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(3):555–560, 2008.

[2] M. Aharon, M. Elad, and A. Bruckstein. Svdd: An algorithm for designing overcomplete dictionaries for sparse representation. *Trans. Sig. Proc.*, 54(11):4311–4322, 2006.

[3] M.S. Asif and J. K. Romberg. Sparse recovery of streaming signals using l1-homotopy. *CoRR*, abs/1306.3331, 2013.

[4] S. B., E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

[5] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *CVPR*, pages 3858–3865, June 2014.

[6] E. J. CandÃĺs and D. L. Donoho. Recovering edges in ill-posed inverse problems: optimality of curvelet frames. *The Annals of Statistics*, 30(3):784–842, 2002.

[7] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011.

[8] M. N. Do and M. Vetterli. Contourlets. In *Beyond Wavelets*. Academic Press, 2003.

[9] D. L. Donoho and M. Elad. *Optimally Sparse Representation in General (nonorthogonal) Dictionaries Via L1 Minimization*. Department of Statistics, Stanford University, 2002.

[10] D. L. Donoho and M. Elad. On the stability of the basis pursuit in the presence of noise. *Signal Process.*, 86(3):511–532, 2006.

[11] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[12] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[13] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15:3736–3745, 2006.

[14] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[15] F. Jiang, J. Yuan, and and A. K. Katsaggelos S. A. Tsaftaris. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115(3):323–333, 2011.

[16] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*, pages 1697–1704, 2011.

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

[18] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, pages 2921–2928, 2009.

[19] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453, 2009.

[20] M. S. Lewicki, T. J. Lewicki, and Sejnowski. Learning overcomplete representations. *Neural computation*, 12:337–365, 2000.

[21] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013.

[22] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2008.

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, pages 1–8, 2008.

[25] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[27] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009.

[28] A. Michal, E. Michael, and B. Alfred. K-svd: Design of dictionaries for sparse representation. In *SPARS'05*, pages 9–12, 2005.

[29] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):310–316, 2010.

[30] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances In Neural Information Processing Systems*. MIT Press.

[31] B. A. Olshausen and D. J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

[32] Y. C. Pati, R. Rezaiifar, Y. C. Pati R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[33] E. L. Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

[34] D. S. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *CVPR*, 2008.

[35] Q. Qiu, Z. Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *ICCV*, pages 707–714, 2011.

[36] H. Ren and T. B. Moeslund. Abnormal event detection using local sparse representation. In *AVSS*, pages 125–130, 2014.

[37] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastr, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[38] C. Yang, J. Yuan, and J. Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, 2013.

[39] Gitta Kutyniok Yonina C. Eldar, editor. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[40] J. Zepeda, C. Guillemot, and E. Kijak. Image compression using sparse representations and the iteration-tuned and aligned dictionary. *Selected Topics in Signal Processing, IEEE Journal of*, 5(5):1061–1073, 2011.

[41] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010.

[42] B. Zhao, F. Li, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, pages 3313–3320, 2011.