ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Semi-supervised Dictionary Learning via Local Sparse Constraints for Violence Detection

Tao Zhang[a,**], Wenjing Jia[b], Chen Gong[c,**], Jun Sun[a], Xiaoning Song[d]

[a]*Department of Computer Science and Technology, Jiangnan University, Wuxi, China*
[b]*Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, Australia*
[c]*School of Computer Science and Engineering, Nanjing University of Science and Technology , Nanjing, China*
[d]*School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China*

ABSTRACT

In this paper, we propose a novel semi-supervised learning framework for violence detection in video surveillance. With this framework, a classifier which distinguishes violent behavior from normal behavior can be trained using inexpensive unlabeled data with the assistance of human operators. Our approach can learn a single dictionary and a predictive linear classifier jointly. Specifically, we integrate the reconstruction error of labeled and unlabeled data, representation constraints and the coefficient incoherence into an objective function for dictionary learning, which enhances the representative and discriminative power of the established dictionary. This has contributed to that the dictionary and the classifier learned from the labeled set yield very small generalization error on unseen data. Experimental results on benchmark datasets have demonstrated the effectiveness of our approach in violence detection.

1. Introduction

Violent behavior seriously endangers social and personal security. A violence detector has immediate applicability both in surveillance domain and for rating online video contents. Currently, millions of video surveillance devices have been used in public places such as streets, prisons and supermarkets. The primary function of the large-scale surveillance systems deployed in institutions such as schools, prisons and elder care facilities is for alerting authorities to potentially dangerous situations. Visual surveillance systems collect a huge amount of videos but humans still must review most of the data to extract the informative knowledge. Our goal is to automatically detect violent behavior without carefully labeling data over large archives.

Violence detection involves similar techniques to those used in many related computer vision applications, e.g., action recognition, object detection, surveillance, etc [28, 31, 6, 1, 11]. Compared with action recognition, little research has been done for detecting violent action or violent contents. Timely detection of violent outbreaks in crowds often mean the difference between life and death. For this practical consideration, in our work, we focus on the challenging work of detecting violence in surveillance videos and aim to develop a system to effectively detect violent behavior using computer vision techniques.

Up to now, there have been some developmental systems [36] for detecting violence in videos. Earlier attempts on this have characterized violent scenes by integrating cues obtained from both video and audio information. For example, Nam et al. [26] proposed to recognize violent scenes by detecting flame and blood, and captured the degree of motion and the characteristic sounds of violent events. Cheng et al. [9] recognized gunshots, explosions and car-braking in audio using computer techniques. Lin et al. [22] presented a novel violent shot detection scheme from both audio and video views (motion, flame and explosion, and blood analysis). Cristani et al. [11] presented a new method for characterizing audio visual events, where separate audio and video signals were processed in a unique fashion. Approaches described in [22, 10, 18, 27]

**Corresponding author: Tel.: +8613405761069;
e-mail: taozhang@jiangnan.edu.cn (Tao Zhang), chen.gong@njust.edu.cn (Chen Gong)

have addressed the problem of violence detection in the context of typical movies, and these approaches rely on fusing mid-level concept predictions made using multi-layer perceptron classifiers.

Later proposals focused on detecting skin and blood in video sequences, requiring either foreground segmentation or the information of skin color, which performance degraded greatly when the color feature was not discriminating enough. For example, Datta et al. [14] exploited an accelerated motion vector to detect the fist fighting and kicking, requiring foreground segmentation to extract the complete silhouettes. Clarin et al. [10] proposed a novel system to detect skin and blood colored regions in video sequences and checked if these regions had intensified throughout the whole sequence. Based on activity recognition approaches, Hassner et al. [17] proposed the Violent Flow (ViF) descriptor and developed a novel means for efficient crowd violence detection. However, the performance of this method degrades significantly when dealing with crowded scenes. Zhang et al. [43] proposed a fast and robust framework (referred to as RVD) for detecting and localizing violence in surveillance scenes, and their experimental results on several benchmark datasets have demonstrated the superiority of this method over the state of the arts in terms of both detection accuracy and processing speed.

Learning dictionaries for sparse coding has recently led to state-of-art performance in many computer vision tasks [40, 2, 23, 33, 29]. For a comprehensive introduction to sparse representation classification (SRC), please refer to [34]. In our opinion, the good performance of SRC algorithms is mainly attributed to the fact that they can do well in determining the intrinsic similarity of objects embedded in high-dimensional image data. Since the SRC scheme achieved competitive performance in face recognition [34], it has triggered researchers' interest in sparsity-based image classification. Previous research on supervised dictionary learning for sparse coding has been targeted on learning more discriminative sparse models [12, 13, 33]. Based on the predefined relationship between dictionary atoms and class labels, existing supervised dictionary learning works can be divided into three categories: shared dictionary learning [41, 20, 24, 5], class-specific dictionary learning [40, 25, 7, 39] and hybrid dictionary learning [45, 21, 30]. However, most of them are based on iterative batch procedures, which access the whole dataset at each iteration and optimize over all data. For large scale datasets, this has become a big challenge due to high computational complexity and high requirement on computers' memory management.

Learning a discriminative dictionary usually requires sufficient labeled training data, which can be expensive and difficult to obtain. Insufficient labeled training data yield a dictionary with potentially poor generalization power. By exploiting the information provided by the vast quantity of inexpensive unlabeled data, we aim to develop a novel algorithm to learn a dictionary which is more representative and discriminative than a dictionary trained using only a limited number of labeled samples.

To address the aforementioned difficulties and aim to develop an effective violence detection algorithm, in this paper, we propose a semi-supervised dictionary learning algorithm that integrates dictionary learning and classifier training. We introduce a novel objective function which contains terms representing the reconstruction error of both labeled and unlabeled data, the representation constraints and the coefficient incoherence. Compared to the supervised dictionary learning approaches, our approach improves the representation power of the dictionary by also exploiting the unlabeled data. It considers the reconstruction error of the unlabeled data in its objective function, and treats the unlabeled points with high confidence in label prediction stage. Our approach thus jointly learns a single over-complete dictionary and an optimal linear classifier.

Our main contributions in this paper are three-fold: 1) We propose a novel semi-supervised learning framework for violence detection, which is suitable for very large data sets; 2) The dictionary is learned from labeled samples for discrimination as well as a large number of unlabeled samples and learning from unlabeled data further increases its representative power; 3) A classification scheme integrating the modified sparse model is proposed.

The rest of the paper is organized as follows. Sect. 2 briefly introduces some related work. Sect. 3 presents the proposed semi-supervised dictionary learning algorithm. Sect. 4 presents experimental results where the performance of our proposed approach is compared with the state-of-the-art approaches. We conclude the paper in Sect. 5.

2. Related Work

2.1. Sparse Representation based Classification Model

In [34], Wright et al. proposed a general Sparse Representation based Classification (SRC) scheme, where the training samples of all classes were taken to form the dictionary representing a query face image. The query image was classified by evaluating which class led to the minimal error of reconstructing it.

Given K classes of subjects, let $A = [A_1, A_2, \cdots, A_K]$ be the dictionary formed by A_i, where A_i $(i = 1, 2, \cdots, K)$ is the subset of training samples of class i. Let y be a test sample. The SRC algorithm is summarized as follows:

(1) Normalize each training sample $A_i, i = 1, 2, \cdots, K$.

(2) Define and solve the l_1-minimization problem: $\hat{x} = \arg\min_x\{\|y - Ax\|_2^2 + \gamma \|x\|_1\}$, where γ is a scalar constant.

(3) Label the test sample y by: $Label(y) = \arg\min_i\{e_i\}$, where $e_i = \|y - A_i\hat{\alpha}_i\|_2^2$, with $\hat{\alpha}_i$ representing the coefficient vector associated with class i.

Obviously, the underlying assumption of this scheme is that a test sample can be represented by a weighted linear combination of just those training samples belonging

to the same class. Its impressive performance reported in [34] shows that sparse representation is naturally discriminative.

2.2. Class-Specific Dictionary Learning

In class-specific dictionary learning (DL), the atoms in the learned dictionary $D = [D_1, D_2, \ldots, D_K]$ have class labels corresponding to the subject classes, where D_i is the sub-dictionary corresponding to class i. Once the representation vector $\hat{\alpha} = [\hat{\alpha}_1; \hat{\alpha}_2; \ldots; \hat{\alpha}_K]$ is computed, the class-specific representation residual $\|y - D_i\hat{\alpha}_i\|_2$ can be used for classification. The sub-dictionary D_i can be learned class by class using the algorithm in [38]: $\arg\min_{D_i,Z_i}\{\|A_i - D_iZ_i\|_F^2 + \lambda\|Z_i\|_1\}$, where Z_i is the representation matrix of A_i on D_i. It can be seen as the basic model of class-specific dictionary learning.

Note that, the above basic model trains the class-specific sub-dictionaries separately, and does not consider the relationship between the sub-dictionaries of different classes. To promote the incoherence between the sub-dictionaries and make the whole class-specific dictionary more distinctive, Ramirez et al. [29] used an incoherence promoting term to encourage the sub-dictionaries to be as independent as possible. Recently, a regularization path algorithm for v-support vector classification suffers exceptions and singularities in some special cases. Gu et al. [15] presented a new equivalent dual formulation for v-SVC, theoretical analysis and experimental results verify that this proposed robust regularization path algorithm can avoid the exceptions completely, handle the singularities in the key matrix, and fit the entire solution path in a finite number of steps. To tackle the ordinal regression problems, Gu et al. [16] proposed incremental support vector learning algorithm. This algorithm can handle a quadratic formulation with multiple constraints, where each constraint is constituted of an equality and an inequality. More importantly, it tackles the conflicts between the equality and inequality constraints.

One of the major difficulties here is that, in some applications, obtaining sufficient labeled training samples is unaffordable. Therefore, learning a discriminative dictionary with minimal human supervision becomes an interesting problem.

3. Semi-Supervised Dictionary Learning

3.1. Proposed Sparse Classification Model

To improve the discriminative power of a dictionary, in our proposed model, two terms, i.e., the representation constraint term and the coefficient incoherence term as described below, are introduced to ensure that the learned dictionary is sufficiently discriminative. The representation constraint term is utilized to ensure the class-specific sub-dictionary to have a good capability when reconstructing a query image using training samples having the same class label. On the other hand, the coefficient incoherence term is utilized to ensure the class-specific sub-dictionary to have a poor capability when reconstructing a query image using training samples with different class labels.

In class-specific DL, each dictionary atom in the learned dictionary, denoted by $D = [d_1, d_2, \cdots, d_k]$, has a class label corresponding to each subject class. d_i $(i = 1, 2, \cdots, K)$ in D is the sub-dictionary corresponding to class i. To take advantage of the large number of inexpensive unlabeled data, the reconstructive term consists of two parts: one part from labeled training data and the other from unlabeled training data.

Given training feature samples $\{a_{ij}|i = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, N\}$, where a_{ij} is the j-th sample of class i, N denotes the number of training samples in each class, and K is the number of classes. Let $A = [A_1, A_2, \cdots, A_i] \in \mathbb{R}^{n\times N}$, where $A_i = [a_{i1}, a_{i2}, \cdots, a_{iN}]$, $i = 1, 2, \cdots, K$ and n is the feature dimension. We aim to include the classification error as a term in the objective function for dictionary learning in order to make the dictionary be optimal for classification. The sparse code Z can be directly used as a feature for classification. Here, we use a linear predictive classifier $f(Z_i; W) = WZ_i$, where $Z = [Z_1, Z_2, \cdots, Z_i]$. Denote the learned dictionary as $D = [d_1, d_2, \cdots, d_k] \in \mathbb{R}^{n\times k}$ $(k > n$ and $k \ll N)$. The objective function for our dictionary learning is defined as:

$$\begin{aligned}\langle D, W, Z\rangle = \arg\min_{D,W,Z}\{&\|A^u - DZ^u\|_F^2 + \left\|A^l - DZ^l\right\|_F^2 \\ &+\lambda_1\left\|Z^l\right\|_1 + \lambda_2\left\|Z^l - m^l\right\|_F^2 + \gamma_1\left\|WZ^l - B\right\|_F^2 \\ &+\gamma_2\|W\|_F^2\}, \\ &s.t.\ \|d_c\|_2 \le 1, \forall c \in \{1, 2, \cdots, k\}\end{aligned} \tag{1}$$

where the superscripts u and l specify whether the sample is from the unlabeled set or the labeled set, $Z = [Z_1, Z_2, \cdots, Z_i] \in \mathbb{R}^{k\times N}$ is the matrix consisting of the coding coefficients of A_i over D, $m = [m_1, m_2, \cdots, m_i] \in \mathbb{R}^{k\times N}$, m_i denotes the mean vector of Z_i in class i, $\|WZ - B\|_F^2$ represents the classification error, $B = [0, 0\cdots, b_N] \in \mathbb{R}^{m\times N}$ are the class labels of input signals A_i. $b(i) = [0, 0\cdots 1\cdots 0, 0]^T \in \mathbb{R}^m$ is a label vector corresponding to an input signal A_i, where the non-zero position indicates the class of A_i, $\|\cdot\|_F$ denotes Frobenius norm. $W \in \mathbb{R}^{m\times k}$ denotes the matrix of classifier parameters, and λ_1, λ_2, γ_1 and γ_2 are the scalars controlling the relative contributions of the corresponding terms.

Different from the conventional sparse model SRC [34], in our model, the representation constraint term $\phi = \lambda_2\|Z - m\|_F^2 + \gamma_1\|WZ - B\|_F^2$ and coefficient incoherence term $\psi = \gamma_2\|W\|_F^2$ are introduced in Eq. 1. Overall, minimizing the coefficient incoherence term and representation constraint term is efficient for classification, because it allows feature being shared among different classes. We will show that good classification results can be obtained using only a single unified dictionary by a simple extension to the objective function for joint dictionary and classifier construction. It encourages the largest classification parameters of training samples from a different class over D are

associated with different corresponding sub-dictionaries.

A major consideration in choosing a suitable optimization method is that since our problem is to be solved in an online learning setting, we cannot separate the labeled set and the unlabeled set in advance. Supervised learning and the unsupervised learning interleave as new data comes in.

3.2. Optimization

Our algorithm alternates between sparse coding and dictionary updating as the input signals arrive sequentially. We rewrite the objective function in Eq. 2 as:

$$\begin{aligned}\arg\min_{D,W,Z}\{\sum_{i=1}^{N_u}\|A_i^u - DZ_i^u\|_F^2 + \sum_{i=1}^{N_l}(\left\|A_i^l - DZ_i^l\right\|_F^2 + \\ \lambda_1\left\|Z_i^l\right\|_1 + \lambda_2\left\|Z_i^l - m_i^l\right\|_F^2 + \gamma_1\left\|WZ_i^l - b_i\right\|_F^2 + \\ \gamma_2\|W\|_F^2)\}, \\ s.t.\ \|d_c\|_2 \le 1, \forall c \in \{1,2,\cdots,k\}\end{aligned} \tag{2}$$

where N_u and N_l are the number of unlabeled and labeled training samples respectively.

For unlabeled A_i, the sparse coding problem simply takes this standard form:

$$\arg\min_{Z\in\mathbb{R}}\|A_i - DZ_i\|_2^2 \tag{3}$$

The Orthogonal Matching Pursuit (OMP) algorithm [19] is adopted here to solve the optimization problem in Eq. 3. In addition, Xue et al. [37] proposed a self-adaptive algorithm based on the global best candidate for global optimization. They employed the same initial population for all algorithms on each benchmark function. The results demonstrate that this algorithm is superior to the other algorithms for solving complex optimization problems. It means that it is a new technique to solve the optimization problem in Eq. 3.

For labeled A_i, the sparse coding problem becomes:

$$\begin{aligned}\arg\min_{D,W,Z}\{\|A_i - DZ_i\|_2^2 + \lambda_1\|Z_i\|_1 + \lambda_2\|Z_i - m_i\|_2^2 \\ +\gamma_1\|WZ_i - b_i\|_2^2 + \gamma_2\|W\|_2^2\}\end{aligned} \tag{4}$$

Although the objective function in Eq. 4 is not jointly convex to (D, W, Z), it is convex with respect to each of D, W and Z when the other two parameters are fixed. Therefore, Eq. 4 can be divided into three sub-problems by optimizing D, W and Z respectively, i.e., updating Z while fixing D and W, updating D while fixing W and Z, and updating W while fixing D and Z, detailed as below.

Updating Z: When D and W are fixed, the objective function in Eq. 4 can be regarded as sparse coding problem for solving $Z = [Z_1, Z_2, \cdots, Z_K]$. When Z_i is updated, all $Z_j (j \neq i)$ are also fixed. Thus, for each Z_i, the objective function in Eq. 4 can be replaced by:

$$\begin{aligned}\langle Z_i\rangle = \arg\min_{Z_i}\{\|A_i - DZ_i\|_2^2 + \lambda_1\|Z_i\|_2^2 + \\ \lambda_2\|Z_i - m_i\|_2^2 + \gamma_1\|WZ_i - b_i\|_2^2\}.\end{aligned} \tag{5}$$

By solving Eq. 5, we have:

$$\begin{aligned}Z_i = \{D^TD + (\lambda_1+\lambda_2)I + \gamma_1 W^TW\}^{-1}(D^TA_i + \\ \lambda_2 m_i + \gamma_1 W^T b_i).\end{aligned} \tag{6}$$

Updating D: When Z and W are fixed, Eq. 4 can be regarded as solving $D = [D_1, D_2, \cdots, D_K]$ sparse coding problem. When D_i is updated, all $D_j (j \neq i)$ are fixed. Thus, Eq. 4 can be replaced by:

$$\begin{aligned}\langle D\rangle = \arg\min_{D}\|A_i - DZ_i\|_2^2, \\ s.t.\ \|d_c\|_2 \le 1, \forall c \in \{1,2,\cdots,k\}.\end{aligned} \tag{7}$$

The subproblem in Eq. 7 can also be solved effectively by the Orthogonal Matching Pursuit (OMP) algorithm.

Updating W: When D and Z are fixed, Let $W = [W_1, W_2, \cdots, W_i]$, Eq. 4 can be replaced by:

$$\arg\min_{w_i}\{\gamma_1(\|W_iZ_i - b_i\|_2^2 + \frac{\gamma_2}{\gamma_1}\|W_i\|_2^2)\} \tag{8}$$

It can be observed that Eq. 8 can be solved using the least square method.

Thus, based on the above equations, we can get the optimized values of all parameters for Eq. 1.

3.3. Classification Scheme

Once the dictionary D has been learned, it can be adopted to represent a testing sample y and perform classification.

We propose the following representation model:

$$\hat{\alpha} = \arg\min_{\alpha}\{\|y - D\alpha\|_F^2 + \gamma\|\alpha\|_2\} \tag{9}$$

where γ is a constant value, and $\hat{\alpha} = [\hat{\alpha}^1, \hat{\alpha}^2, \cdots, \hat{\alpha}^K]^T$, where $\hat{\alpha}^i$ is the sub-vector associated with sub-dictionary D_i. In the learning stage, we have enforced the class-specific representation residual to be discriminative. Therefore, if y is from class i, the residual $\left\|y - D_i\hat{\alpha}^i\right\|_2^2$ should be very small; otherwise, $\left\|y - D_j\hat{\alpha}^j\right\|_2^2, j \neq i$ should be large. In addition, the coefficient vector $\hat{\alpha}$ should be far different from the coefficient vector of other classes. Based on the discrimination capability of both representation residual and coefficient vector, the metric for classification can be defined as:

$$l = W\hat{\alpha}. \tag{10}$$

The label for y is assigned by the position corresponding to the largest value in the label vector l.

4. Experimental Results and Discussion

To evaluate the performance of our proposed ideas, we compare our method against the state-of-the-art approaches either implemented by us or cited from literature. These include the BoW based methods [6], the Appearance and Motion DeepNet (AMDN) method in [35], the Violent

Flow (ViF) method in [17], the method in [34], probabilistic semi-supervised (PSS) method [4], structural sparse semi-supervised (SSS) method [32] and our recently published method in [44]. Results are reported with mean prediction accuracy (ACC)±, standard deviation (SD), as well as the area under the ROC curve (AUC). Next, we first briefly introduce the three benchmark datasets, and then present experimental results with discussion.

4.1. Datasets

Experiments of our method were conducted on three challenging benchmark datasets, i.e., the Hockey Fight dataset [17], the BEHAVE dataset [3] and the Crowd Violence dataset [17].

The Hockey Fight dataset contains 1000 video clips of actions from hockey games of the National Hockey League (NHL), of which 500 are randomly selected as training data. Each clip consists of 50 frames (with a resolution of 360×288 pixels).

The BEHAVE dataset contains more than $70,000$ frames (with a resolution of 640×480 pixels) and various scenarios, including walking, running, chasing, discussing in groups, driving or cycling across the scene, fighting and so on. We partition the dataset into clips with various activities. Each clip consists of at least 200 frames. Finally, we pick 80 clips for violence detection, including 20 violence clips and 60 non-violence clips.

The Crowd Violence dataset is assembled for testing violent crowd behavior detection. All video clips are collected from YouTube, representing a wide range of scene types, video qualities and surveillance scenarios. The dataset consists of 246 video clips including 123 violent clips and 123 normal clips with a resolution of 320×240 pixels. The whole dataset is split into five sets for 5-fold cross-validation. Half of the footages in each set presents violent crowd behavior and the other half presents non-violent crowd behavior.

4.2. Experimental Settings

To evaluate the classification accuracy, we employ the 5-fold cross validation test on each dataset. Since the classification accuracy depends on the number of labeled training samples, it is tricky to do a fair comparison with other methods unless we fix our settings. To address this issue, we conducted the experiments in two folds: (1) Split the training set into labeled set and unlabeled set. While our method takes advantage of both sets due to our learning strategy, the competing methods can only take the labeled set for training since the unlabeled samples are useless to them. (2) To compare our best recognition rate with the state-of-the-arts, we assumed all the training samples are labeled.

In order to determine the number of labeled training samples, we first compare our method with the original SRC classification algorithm on the three databases. The experimental results shown in Figs. 1– 3 demonstrate the effect of the number of labeled samples on our performance in comparison with others.

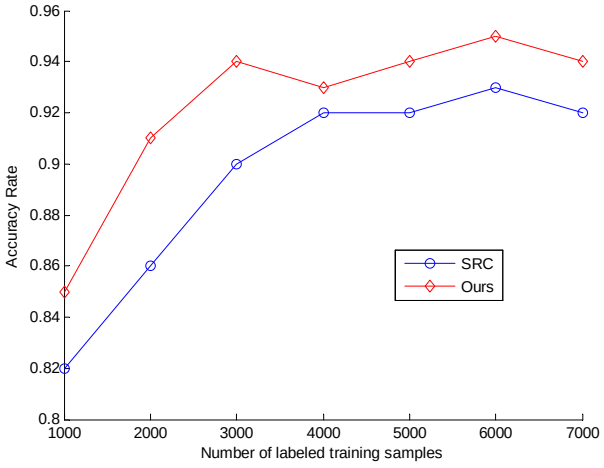

Fig. 1: Classification accuracy with varying number of labeled samples on Hockey Fight dataset.

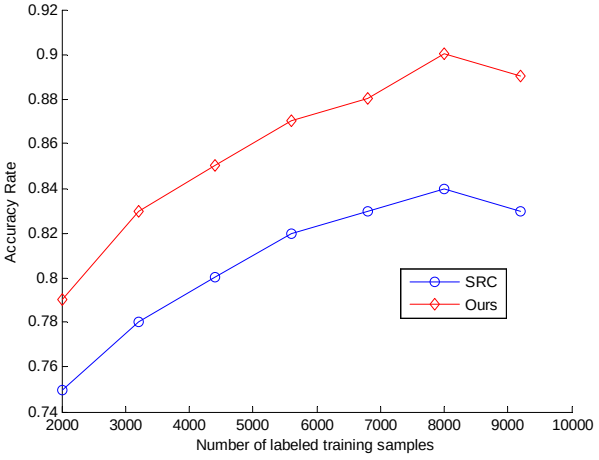

Fig. 2: Classification accuracy with varying number of labeled samples on BEHAVE dataset.

Furthermore, in our proposed model, there are two stages, i.e., dictionary learning stage and classification stage. In the dictionary learning stage, we set $\lambda_1 = 0.005, \lambda_2 = 0.01, \gamma_1 = 0.05, \gamma_2 = 0.1$; and in classification stage, we set $\gamma = 0.05$. All of the experiments are executed on a workstation with 2.8GHz CPU and 16GB RAM.

4.3. Feature Extraction

In this work, we adopt the Motion Weber Local Descriptor (MoWLD) that we proposed in our recent works [44, 42] in the spirit of the well-known Weber Local Descriptor (WLD) [8] as features for sparse representation-based dictionary learning. We extended the original WLD spatial descriptions by adding a temporal component to the appearance descriptor, which implicitly captured local motion information as well as low-level image appearance information. This has made MoWLD a powerful and robust descriptor for motion images.

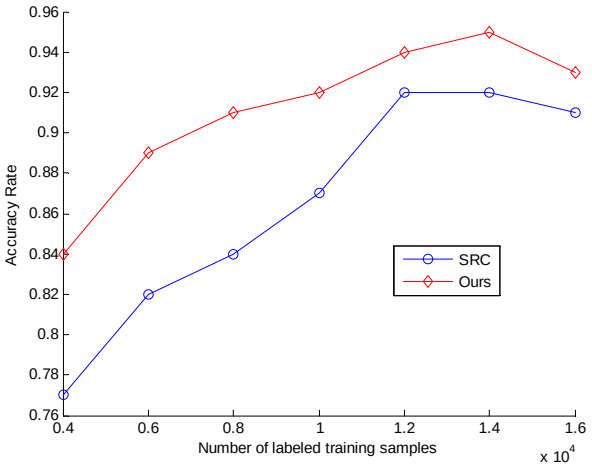

Fig. 3: Classification accuracy with varying number of labeled samples on Crowd Violence dataset.

4.4. Results and Discussion

Results on the Hockey Fight dataset. We compare our approaches with the state-of-the-art approaches on the Hockey Fight dataset, including HOG, HOF, and MoSIFT [6]. For our algorithm, 120 video clips of this dataset are randomly selected as labeled training samples, and the others are unlabeled samples. As shown in the table, MoSIFT and HOG based BoW models perform comparably, with slight better results achieved by HOG compared with MoSIFT. Our recently proposed MoWLD (noted as "MoWLD+BoW" in the table) outperforms all above approaches.

Table 1 shows the results obtained by the proposed approach after adopting the BoW approach and the proposed sparse classification model into the MoWLD approach. We compare the results with not only HOG, HOF, HNF, MoSIFT and MoWLD paired with BoW, but also ViF, PSS, SSS, AMDN and MoWLD paired with sparse coding method in [44]. The method in [44] adopts a MoWLD paired with sparse coding method to detect violent behavior, which has reduced many noise disturbances, so its performance is better than the BoW-based approaches. The AMDN approach [35] utilizes optical flow as the input image feature, and there exist many redundant and interference features, so its performance is not very good. The results of structural sparse semi-supervised (SSS) method in [32] is very close to our method. It can be seen from Table 1 that our proposed semi-supervised sparse classification model has performed the best. Furthermore, our approach outperforms the competing SRC, PSS, SSS and MoWLD+Sparce Coding with the same size of dictionary. This is due to the fact that our proposed semi-supervised class-specific dictionary learning framework incorporates representation constraints and coefficient incoherence terms resulting in the highest recognition rate.

Results on the BEHAVE dataset. We compare our approaches with the state-of-the-art approaches implemented by us on the BEHAVE dataset, where 80 clips

Table 1: Detection results on the Hockey Fight dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [6]	88.77±0.73%	0.9123
HOF+BoW [6]	86.07±0.59%	0.8843
HNF+BoW [6]	89.27±0.79%	0.9294
ViF [17]	90.07±0.99%	0.9429
MoSIFT+BoW [6]	88.8±0.75%	0.9052
MoWLD+BoW	89.28±0.93%	0.9112
AMDN [35]	89.7±1.13%	0.9198
SRC [34]	94.2±1.07%	0.9528
MoWLD+Sparse Coding [44]	93.8±1.08%	0.9618
PSS [4]	95.5±1.07%	0.9628
SSS [32]	96.1±1.04%	0.9709
Proposed method	96.5±1.04%	0.9758

Fig. 4: Examples of false alarms on the BEHAVE dataset.

of this dataset are randomly picked for training. For our algorithm, 20 video clips of this dataset are randomly picked up to form the labeled training samples, and the others are unlabeled samples. Table 2 presents the results obtained with the above mentioned methods on this dataset. As it can be seen from the table, our proposed semi-supervised class-specific dictionary learning method outperforms other approaches. This again demonstrates that the proposed approach is significantly superior in performance to all other approaches. The performance of the MoWLD paired with sparse coding method in [44] and the AMDN [35] on this dataset is consistent with their performance on the Hockey Fight dataset. Furthermore, our semi-supervised sparse model method outperforms SRC and PSS methods. It validates that the representation constraints term and coefficient incoherence term can improve the discriminative ability of sparse representation model. The results of structural sparse semi-supervised (SSS) method in [32] is better than our method. The results on this dataset demonstrate that SSS algorithm is more effective for detecting violence in a group fighting scene.

False alarms only happen when a group of people get together to do some strenuous non-violence activities (example frames are shown in Fig. 4).

Results on the Crowd Violence dataset. This dataset is more challenging than the other two datasets

Table 2: Detection results on the BEHAVE dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [6]	58.97±0.34%	0.6394
HOF+BoW [6]	60.03±0.28%	0.5923
HNF+BoW [6]	58.24±0.31%	0.6113
ViF [17]	83.62±0.19%	0.8632
MoSIFT+BoW [6]	62.78±0.23%	0.6679
MoWLD+BoW	81.65±0.18%	0.8324
AMDN [35]	84.22±0.17%	0.8562
SRC [34]	82.7±0.14%	0.8538
MoWLD+Sparse Coding [44]	85.27±0.13%	0.8758
PSS [4]	85.15±0.13%	0.8727
SSS [32]	89.07±0.10%	0.9096
Proposed method	88.26±0.11%	0.9028

Table 3: Detection results on the Crowd Violence dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [6]	57.98±0.37%	0.6252
HOF+BoW [6]	58.71±0.12%	0.5931
HNF+BoW [6]	57.05±0.32%	0.6154
ViF [17]	82.13±0.21%	0.8595
MoSIFT+BoW [6]	57.09±0.37%	0.6073
MoWLD+BoW	88.16±0.19%	0.9028
AMDN [35]	84.72±0.17%	0.8891
SRC [34]	89.6±0.18%	0.9288
MoWLD+Sparse Coding [44]	89.38±0.13%	0.9216
PSS [4]	89.5±0.13%	0.9298
SSS [32]	91.9±0.12%	0.9357
Proposed method	92.25±0.12%	0.9408

Fig. 5: Examples of false alarms on the Crowd Violence dataset.

because it contains many crowded scenes. The set contains 246 clips divided into five splits, each containing 123 violent and 123 non-violent scenes. For our algorithm, 35 video clips of this dataset are randomly selected to constitute the labeled set, and the others are unlabeled samples. Table 3 presents the results obtained using different methods mentioned above on this dataset. In this dataset, the performance of AMDN method is still very stable. However, because of the introduction of optical flow noise, the performance of AMDN is not very good. Our proposed semi-supervised learning approach still outperforms other approaches. MoWLD descriptor is still significantly superior in performance to HOG, HOF, HNF, AMDN, PSS and SSS. It confirms that our selected MoWLD is a more effective descriptor for describing action feature. Consistent with the results on the previous two datasets, our proposed semi-supervised algorithm outperforms the SRC methods. It indicates that the proposed classification model has a smaller classification error rate compared with the original SRC. Results on this dataset demonstrate that our algorithm is also effective for detecting violence in a crowded scene. Some false alarms (some examples are shown in Fig. 5) are caused by people waving flags, vigorously clapping hands, or sharply and disorderly waving hands.

By verifying the results, we can conclude that our proposed system is effective and robust for detecting violence with complex scenarios, such as various distances from cameras, severe occlusions between people and crowed scenes.

5. Conclusion

We proposed a novel semi-supervised dictionary learning approach for violence detection. It is particularly suitable for large scale datasets where a batch mode does not work well. Moreover, by adding the representation constraints and the coefficient incoherence, our algorithm actively seeks for the critical points for labeling, and identifies the easily classified points as labeled data. In this way we reduce the manual labeling effort to the minimum without sacrificing the performance too much. The dictionary and the classifier are jointly learned to further enhance the discriminative power. Experimental results showed that our approach have achieved state-of-art performance. Possible future work includes updating the learned discriminative dictionary for input signals from a new sample.

Acknowledgment

This research was partly supported by National Science Foundation, China (No: 61672263,21365008), the Natural Science Foundation of Jiangsu Province (Grant no. SBK2017042111, BK20161135), the Open Project Program of the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University, No. MJUKF201709).

References

[1] Aggarwal, J.K., Ryoo, M.S., 2011. Human activity analysis: A review. ACM Computing Surveys 43, 1–43.

[2] Aharon, M., Elad, M., Bruckstein, A., 2006. An algorithm for designing overcomplete dictionaries for sparse representation. IEEE T. SP 54, 5311–5322.

[3] Andrade, E., Fisher, R., 2006. Modelling crowd scenes for event detection, pp. 175–178.

[4] Behnam, B.M., Ali, Z., Mohammadreza, Z., Mahdieh, S.B., 2013. Pssdl: Probabilistic semi-supervised dictionary learning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases 8190, 192–207.

[5] Bengio, S., Pereira, F., Singer, Y., Strelow, D., 2009. Group sparse coding. NIPS , 791–804.

[6] Bermejo, E., Deniz, O., Bueno, G., Sukthankar, R., 2011. Violence detection in video using computer vision techniques. In: Proceedings of the 14th international conference on computer analysis of images and patterns , 332–339.

[7] Castrodad, A., Sapiro, G., 2008. Sparse modeling of human actions from motion imagery. Int.J.Comput.Vis. , 1–15.

[8] Chen, J., Shan, S., He, C., Zhao, G., Chen, X., Gao, W., 2010. Wld: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 1705–1720.

[9] Cheng, W.H., Chu, W.T., Wu, J.L., 2003. Semantic context detection based on hierarchical audio models. In: Proceedings of the ACM SIGMM workshop on Multimedia information retrieval , 109–115.

[10] Clarin, C., Dionisio, J., Echavez, M., Naval, P.C.S., 2005. Detection of movie violence using motion intensity analysis on skin and blood. Tech. rep., University of the Philippines .

[11] Cristani, M., Bicego, M., Murino, V., 2007. Audio-visual event recognition in surveillance video sequences. Multimedia, IEEE Transactions on , 257–267.

[12] Dai, P., Di, H., Dong, L., Tao, L., Xu, G., 2008. Group interaction analysis in dynamic context. IEEE TRANS. SYST., MAN, CYBERN., SYST. 38, 275–282.

[13] Damen, D., Hogg, D., 2009. Recognizing linked events: searching the space of feasible explanations. In: Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on , 927–934.

[14] Datta, A., M., S., N., D.V.L., 2002. Person-on-person violence detection in video data. Proceedings of IEEE International Conference on Image Processing (ICIP2002) , 433–438.

[15] Gu, B., Sheng, V.S., 2016. A robust regularization path algorithm for ν-support vector classification. IEEE Transactions on Neural Networks and Learning Systems , 1–8.

[16] Gu, B., Sheng, V.S., Tay, K.Y., Romano, W., Li, S., 2015. Incremental support vector learning for ordinal regression. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 26, 1403–1416.

[17] Hassner, T., Y., I., Kliper-Gross, O., 2012. Violent flows: Real-time detection of violent crowd behavior. 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) , 1–6.

[18] Ionescu, B., Schlu-ter, J., Mironica, I., 2013. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. ACM Conference on International Conference on Multimedia Retrieval , 215–222.

[19] Jian, W., Kwon, S., Shim, B., 2011. Generalized orthogonal matching pursuit. IEEE Transactions on Signal Processing 60, 6202–6216.

[20] Jiang, Z.L., Lin, Z., Davis, L.S., 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35, 791–804.

[21] Kong, S., Wang, D.H., 2012. A dictionary learning approach for classification: Separating the particularity and the commonality. Proc. ECCV , 186–199.

[22] Lin, J., Wang, W., 2009. Weakly-supervised violence detection in movies with audio and video based co-training. In the 10th IEEE Pacific-Rim Conference on Multimedia, Dec , 990–935.

[23] Mairal, J., Bach, F., Ponce, J., 2012. Task-driven dictionary learning. IEEE Trans.Pattern Anal. Mach. Intell. 34, 791–804.

[24] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2009. Supervised dictionary learning. NIPS , 792–800.

[25] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zissserman, A., 2008. Learning discriminative dictionaries for local image analysis. Proc. CVPR , 1233–1240.

[26] Nam, J., Alghoniemy, M., Tewfik, A., 1998. Audio-visual content-based violent scene characterization. Proceedings of IEEE International Conference on Image Processing (ICIP1998) , 353–357.

[27] Penet, C., Demarty, C., Gravier, G., Gros, P., 2012. Multimodal information fusion and temporal integration for violence detection in movies. IEEE International Conference on Acoustics, Speech and Signal Processing , 2393–2396.

[28] Popoola, O.P., Wang, K.J., 2012. Video-based abnormal human behavior recognition - a review. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42, 865–878.

[29] Ramirez, I., Sprechmann, P., Sapiro, G., 2010. Classification and clustering via dictionary learning with structured incoherence and shared feature. Proc. IEEE Int. Conf. CVPR , 3602–3611.

[30] Shen, L., Wang, S.H., Sun, G., Jiang, S.Q., Huang, Q.M., 2013. Multi-level discriminative dictionary learning towards hierarchical visual categorization. Proc. CVPR , 1320–1327.

[31] Tran, D., Sorokin, A., 2008. Human activity recognition with metric learning. European Conference on Computer Vision (ECCV), 2008 , 548–561.

[32] Wang, D., Zhang, X., Fan, M., Ye, X., 2016. Semi-supervised dictionary learning via structural sparse preserving. Thirtieth AAAI Conference on Artificial Intelligence , 2137–2144.

[33] Wang, H., Yuan, C., Hu, W., Sun, C., 2012. Supervised class-specific dictionary learning for sparse modeling in action recognition. Pattern Recognit. 45, 3902–3911.

[34] Wright, J., Yang, A.Y., Ganesh, A., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Patt. Mach. Intel. 31, 210–227.

[35] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N., 2015. Learning deep representations of appearance and motion for anomalous event detection. In: The British Machine Vision Conference (BMVC) , 1–12.

[36] Xu, L., Gong, C., Yang, J., 2013. Violent video detection based on mosift feature and sparse coding. ICASSP 2013 , 3538–3542.

[37] Xue, Y., Jiang, J.M., Zhao, B.P., Ma, T.H., 2017. A self-adaptive articial bee colony algorithm based on global best for global optimization. Soft Computing , 1–18.

[38] Yang, A.Y., Zhou, Z., Balasubramani, A.G., Sastry, S.S., 2010. Fast l1-minimization algorithms for robust face recognition. IEEE Trans.Image Process. 19, 3234–3246.

[39] Yang, M., Zhang, L., Feng, X.C., Zhang, D., 2015. Sparse representation based fisher discrimination dictionary learning for image classification. Int.J.Comput.Vis. , 1–15.

[40] Yang, M., Zhang, L.and Feng, X.C., Zhang, D., 2011. Fisher discrimination dictionary learning for sparse representation. Proc. ICCV , 654–662.

[41] Zhang, Q., Li, B.X., 2010. Discriminative k-svd for dictionary learning in face recognition. Proc. IEEE Int. Conf. CVPR , 255–264.

[42] Zhang, T., Jia, W., He, X., Yang, J., 2017a. Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Transactions on Circuits and Systems for Video Technology 27, 696–709.

[43] Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X., 2016. A new method for violence detection in surveillance scenes. Mach.Vision.Appl. 93, 1408–1425.

[44] Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X., 2017b. Mowld: a robust motion image descriptor for violence detection. Mach.Vision.Appl. 76, 1419–1438.

[45] Zhou, N., Fan, J.P., 2012. Learning inter-related visual dictionary for object recognition. Proc. CVPR , 3490–3497.