Discriminative Dictionary Learning with Motion Weber Local Descriptor for Violence Detection

Tao Zhang, Wenjing Jia, Xiangjian He, *Senior Member, IEEE,* and Jie Yang, *Senior Member, IEEE*

***Abstract*—Automatic violence detection from video is a hot topic for many video surveillance applications. However, there has been little success in developing an algorithm that can detect violence in surveillance videos with high performance. In this paper, following our recently proposed idea of Motion Weber Local Descriptor (MoWLD), we make two major improvements and propose a more effective and efficient algorithm for detecting violence from motion images. First, we propose an improved WLD (IWLD) to better depict low-level image appearance information, and then extend the spatial descriptor IWLD by adding a temporal component to capture local motion information and hence form the MoIWLD. Second, we propose a modified sparse-representation-based classification (SRC) model to both control the reconstruction error of coding coefficients and minimize the classification error. Based on the proposed sparse model, a class-specific dictionary containing dictionary atoms corresponding to the class labels is learned by using class labels of training samples. With this learned dictionary, not only the representation residual but also the representation coefficients become discriminative. A classification scheme integrating the modified sparse model is developed to exploit such discriminative information. Experimental results on three benchmark datasets have demonstrated the superior performance of the proposed approach over the state-of-the-arts.**

***Index Terms*—violence detection, MoIWLD, sparse representation, class-specific dictionary learning.**

I. Introduction

VIOLENT behavior seriously endangers social and personal security [1]. A violence detector has immediate applicability both in the 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 (some sample frames from benchmark datasets are shown in Fig. 1). Visual surveillance systems collect huge amounts of videos but humans must still review most of the data to extract informative knowledge. Our goal is to automatically recognize violent behaviors 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 [2], [3], [4], [5], [6], [7], [8]. Compared with action recognition, relatively little research has been found for detecting action or violent contents. Timely detection of violent outbreaks in crowds may 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 behaviors using computer vision techniques.

Up to now, there have been some developmental systems for detecting violence in videos. Earlier attempts on this have characterized violent scenes by integrating cues obtained from both video and audio tracks. For example, Nam et al. [6] proposed to recognize violent scenes by detecting flame and blood and capturing the degree of motion and the characteristic sounds of violent events. Cheng et al. [9] recognized gunshots, explosions and car-braking in audio. Cristani et al. [7] presented a new method for characterizing audio visual events, where separate audio and video signals were processed in a unique fashion. Lin and Wang [10] presented a novel violent shot detection scheme from both audio and video views (motion, flame and explosion, and blood analysis). 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. [11] exploited an accelerated motion vector to detect the fist fighting and kicking, requiring foreground segmentation to extract the precise silhouettes. Clarin et al. [12] 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. [13] 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. [14] proposed a fast and robust framework (referred to as RVD) for detecting and localizing violence in surveillance scenes, and 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.

In recent studies, some approaches based on spatiotemporal interest points (STIPs) [15] have been proposed for violence detection. Generally, after extracting interest points over the frames, the Bag-of-Words (BoW) approach is used for recognizing violence. This kind of methods compute only in the regions of interest (located around the detected interest points) and are not discriminative sufficiently. Moreover, the BoW model roughly assigns each feature vector to only one visual word and ignores the spatial relationships among the features. To address these problems, Zhou et al. [16] proposed a novel

T. Zhang and J. Yang are with Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 201100, China.(e-mail:zhb827@sjtu.edu.cn; jieyang@sjtu.edu.cn)

W. Jia and X. He are with the Global Big Data Technologies Centre (GBDTC), University of Technology Sydney, PO Box 123, Broadway, NSW, Australia.(E-mail: Wenjing.Jia@uts.edu.au; Xiangjian.He@uts.edu.au).

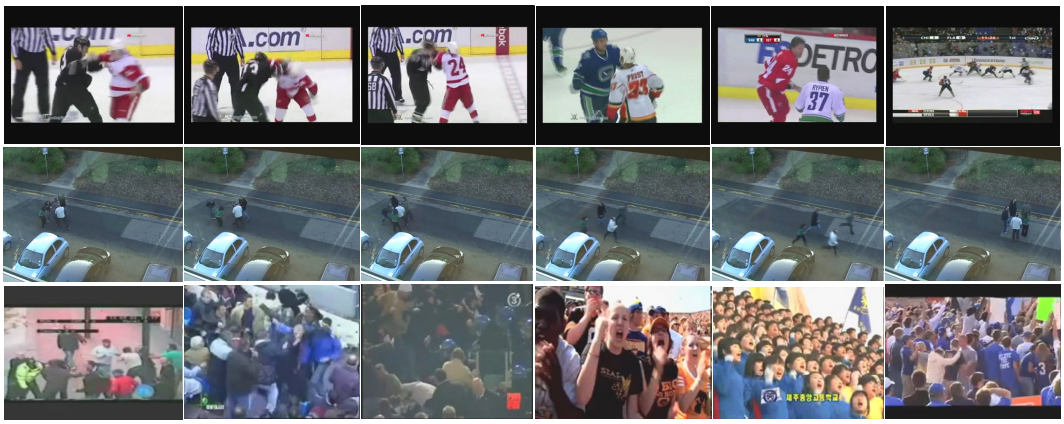


Figure 1: Sample frames from the Hockey Fight dataset (first row), the BEHAVE dataset (second row) and the Crowd Violence dataset (third row). In each row, the left three columns are violent scenes while the right three columns are non-violent scenes.

codebook construction method to encode video feature. This approach suits best for structured videos, instead of the videos in our datasets which are more textural.

Targeting on the above challenges, we focus on effective detection of spatiotemporal interest points and feature representation for detecting violent behaviours in real video scenes.

A straightforward way to detect spatiotemporal interest points is to extend the domain of algorithms used to detect 2D interest points (e.g., the widely adopted Scale Invariant Feature Transform (SIFT) [17]) to motion domain. Since recognizing human motion is more complicated than recognizing other objects, motion recognition requires enhanced local features that provide both shape and motion information. SIFT is a sparse descriptor that considers the regions of interest only, and are not discriminative enough. To include motion information into an SIFT descriptor, many attempts have been reported [18]. Chen et al. [19] proposed the motion SIFT (MoSIFT) for interest point detection. MoSIFT not only encodes local appearance of an interest point but also explicitly models the local motion of the point. In the aspect of action recognition, Chen et al. have also demonstrated the superiority of the MoSIFT over four existing descriptors, i.e., 3D Histogram of Gradients (HOG), Histogram of Optical Flow (HOF), HNF (a combination of HOG and HOF), and grid aggregated HOG and HOF [20].

Another way for violence detection is built on the Weber Local Descriptor (WLD) [21]. Based on Weber's law which states that human perception of a pattern depends on not only the change of a stimulus (such as sound, lighting, etc.) but also the original intensity of the stimulus, Chen et al. [21] proposed the WLD for face detection. Wang et al. [22] further exploited the illumination-insensitive characteristics of the WLD. Li et al. [23] proposed a multi-scale WLD and a multi-level information fusion approach. Since recognizing violent activities is more complicated than recognizing human faces, violence detection requires enhanced local features that can also provide sufficient motion information. WLD-based interest points with sufficient motion information can provide the necessary information for action recognition. Therefore, motivated by the widely used optical flow approach, which detects the movement of a region by calculating where a region moves by measuring temporal differences, we have recently proposed a novel descriptor, called Motion WLD (MoWLD) [24], to measure the spatiotemporal features of dynamic activities effectively. MoWLD can detect a spatially distinctive interest point with a substantial motion. However, we observed from experiments that, the original WLD worked well for image areas with rich texture, but gave zero responses for areas which were rather flat. In this paper, we further improve MoWLD by proposing an improved WLD (IWLD) and then forming a motion IWLD (MoIWLD) accordingly.

Once the representation of an action feature is made, action classification can be performed by employing one of the well-known pattern recognition techniques such as nearest neighbor methods based on dimensionality reduction [25], artificial neural networks [26], Support Vector Machines (SVMs) [20], and Sparse Rrepresentation-based Classification (SRC) [27]. Approaches based on local spatiotemporal descriptors are traditionally combined with models of Bag-of-Words (BoWs) and have achieved promising performance in violence detection [2], [15]. However, the conventional BoW methods rely on the discriminative power of local spatiotemporal descriptors and how often these descriptors occur in the video. Moreover, the performance of a BoW model can be degraded significantly due to a high quantization error. Currently, methods based on sparse coding have been successfully utilized in the areas of action recognition and image classification [28], [18], [29]. The sparse coding methods transform a low-level descriptor to a linear combination of a few atoms in a well-trained dictionary. They usually generate fewer reconstruction errors and can achieve a more discriminative feature representation compared with the BoW models. Wright et al. [27] proposed a general classification scheme based on a sparse representation and applied it to face recognition. However, how to learn a discriminative dictionary for both representation and classification of sparse data is still an open problem.

Addressing the above challenges, this paper proposes an effective and robust violence detection algorithm. We propose a novel discriminative holistic descriptor for action representation and a novel supervised learning algorithm based on a

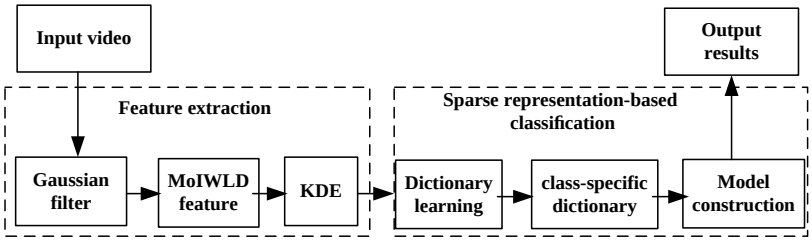


Figure 2: The framework of the proposed method.

modified sparse model to learn a discriminative dictionary for violence classification. We combine the MoIWLD descriptor with the sparse coding method in order to generate a more discriminative representation in a video for violence detection. The framework of our approach is illustrated in Fig. 2. Firstly, we extract the MoIWLD features from an input video and employ the Kernel Density Estimation (KDE) based feature selection method [30], [31] to eliminate redundant and irrelevant features from the extracted MoIWLD descriptor. Then, in order to perform violence classification, we propose a modified sparse model to learn a class-specific dictionary, consisting of the dictionary atoms that correspond to the class labels.

Compared with the MoWLD recently proposed by us and published in [24], the new contributions presented in this paper are mainly in the following two aspects:

- First, in order to detect a sufficient number of interest points containing the necessary information to recognize a violent activity, we follow our recent proposal of MoWLD [24] and present an improved version of MoWLD (i.e., MoIWLD) to extract the low-level image and motion features of a query video more effectively. The proposed MoIWLD detects spatially distinctive interest points having substantial motions. In a sense, this descriptor takes the advantages of both SIFT in terms of computing the histogram using the magnitude and orientation of a gradient, and Local Binary Pattern (LBP) [32] in terms of computational efficiency.
- Secondly, we propose a modified sparse model which minimizes both the reconstruction error of coding coefficients and the classification error. Based on this model, a class-specific dictionary, i.e., the dictionary atoms that correspond to the class labels, is learned by using the class labels of training samples. With this learned dictionary, not only the representation residual but also the representation coefficients become discriminative. A classification scheme integrating the modified sparse model is then developed to exploit such discriminative information.

Experimental results on three challenging benchmark datasets demonstrate the superior performance of our proposed approach over the state-of-the-arts.

The remaining of this paper is organized as follows. Related works on action recognition, anomaly detection and dictionary learning are discussed in Section II. Section III introduces our new MoIWLD feature descriptor and the corresponding feature selection method. Section IV details the proposed modified sparse framework and the supervised class-specific dictionary learning method for classification. In Section V, experimental results and analysis are presented. Finally, conclusions are drawn in Section VI.

II. Related Works

In this section, we present recent works related to our research and the proposed method on action recognition, anomaly detection as well as on supervised dictionary learning.

A. Related Works on Action Recognition

Violent behaviour indicates rapid and significant movement of persons and objects. It is a specific problem within the greater problem of action recognition. We refer readers to two recent survey articles [3], [33]. We classify the most recent research works on action recognition into three groups: statistical, spatiotemporal and description-based approaches.

Statistical approaches use statistical, state-based models to recognize activities. Zhang et al. [34] proposed a two-layer Hidden Markov Model (HMM) framework. First, by decomposing the problem hierarchically, learning is performed on low-dimensional observation spaces, in order to produce and apply simple models. This proposed framework is easy to interpret, as both individual actions and group activities have clear meanings. However, the impact of the size of each parameter space for the layered decomposition in this approach was not given. Nguyen et al. [35] presented the use of the shared-structure Hierarchical HMM (HHMM) to recognize people's behaviors. They have used both the exact approximate inference algorithm and the Rao-Blackwellised particle filter (RBPF) to infer behaviors at different levels. However, it failed to recognize complex behaviors. Shi et al. [36] associated the P-Net and the D-Condensation algorithm, and provided a natural and efficient way to integrate temporal and logical relationships in daily activities. However, its performance strongly relied on the manually labeled training data. Dai et al. [37] introduced an event-based dynamic context model to address the problems of context awareness in the analysis of group interaction scenarios. However, the applicable scenario was very limited. Damen and Hogg [38] presented a novel framework for jointly recognizing and linking visually ambiguous events. However, it failed to recognize complex and ambiguous events.

Spatiotemporal approaches represent activities in volume trajectories as a set of features. These approaches were used to match input with their representative models to determine

the activity classes. Bobick and Davis [39] presented a novel representation and recognition technique for identifying movements. This approach was based upon temporal templates for dynamic matching in time. However, it was only applicable in those situations where the motion of an object could be easily separated into various simple movements. In [33], an approach based on optical flows was used to represent apparent velocities of movements of brightness patterns in an image, and similar approaches were employed for modeling typical motion patterns [19], [5]. However, this approach would fail in extremely crowded scenes. This issue could be solved by a dense, local sampling of optical flows proposed in [40]. Baysal and Duygulu [41] introduced a line based pose representation and explored its ability in recognizing human actions. They encapsulated a human pose into a collection of line-pairs, preserving the geometrical configurations of the components forming the human figure. Saghafi and Rajan [42] proposed a novel embedding which was optimum in the sequence recognition framework based on the Spatiotemporal Correlation Distance (SCD) as the distance measure. However, its performance mainly relied on the key poses chosen equidistantly from one action period and could not work well in a complex environment. Oikonomopoulous et al. [43] represented a human action as a collection of short trajectories extracted in the areas having significant amounts of visual activities in a scene. Vishwakarma and Agrawal [44] considered multiclass activities fused in a three dimensional (spatial and time) coordinate system to achieve maximum accuracy. This method worked well in semantically varying events and was robust to scale and view changes.

Description-based approaches explicitly maintain spatiotemporal structures for human activities. They represent a high-level human activity using the simpler activities, composing the activity, and their temporal, spatial and logical relationships. Yang et al. [45] proposed to use a scheme of multi-feature learning via hierarchical regression for multimedia semantics understanding. The algorithm could be applied to a wide range of multimedia applications. Gao et al. [46] proposed to learn an optimal graph from multiple cues (i.e., partial labels and multiple features) to embed the relationships among data points more precisely. However, this approach would fail in extremely crowded scenes.

B. Related Works on Anomaly Detection

Recently, more and more research attention is given for anomaly detection in video [4], i.e., detecting irregular patterns that are different from regular video events. Although there are many existing works on video anomaly detection [40], [4], few of them can work well in crowded scenes. Vijay Mahadevan et al. [6] proposed a novel framework for anomaly detection in crowded scenes. Three tasks were deemed to be important for the design of a localized video representation suitable for anomaly detection. The three tasks are jointly modeling of the appearance and dynamics of a scene, detection of temporal abnormalities and the detection of spatial abnormalities. Marco Bertini et al. [40] constructed a multi-scale local descriptor for anomaly detection and achieved real-time performance in video surveillance applications. Mehrsan et al. [47] presented a novel approach for video parsing and simultaneous online learning of dominant and anomalous behaviors in surveillance videos. Xu et al. [48] presented a novel unsupervised deep learning framework for anomalous event detection in complex video scenes. To exploit the complementary information of both appearance and motion patterns, they introduced a novel double fusion framework, combining both the benefits of traditional early fusion and late fusion strategies.

C. Supervised Dictionary Learning

Sparse representation has received a lot of attention in action recognition area, but most sparse models mainly consider minimizing the reconstruction error, and little attention is paid to better classification. Recent research on supervised dictionary learning for sparse coding has been targeted on learning more discriminative sparse models [49], [50], [51]. According to the predefined relationship between dictionary atoms and class labels, we can divide the existing supervised dictionary learning works into three categories: shared dictionary learning, class-specific dictionary learning and hybrid dictionary learning.

In shared dictionary learning, a dictionary shared by all classes is learned, and the discriminative power of the representation coefficients is mined [49], [50], [52], [53], [54]. In [53], Marial et al. proposed a scheme which learned discriminative dictionaries while training a linear classifier over coding coefficients. Based on KSVD [55], Zhang and Li [50] proposed a joint learning framework called discriminative KSVD (DKSVD) to learn a dictionary for face recognition. Following the work in [50], Jiang et al. [52] proposed to enhance the discriminative power of a learned dictionary via adding a label consistent term. Recently, Mairal et al. [49] proposed to learn a dictionary by minimizing different risk functions over the coding coefficients for different tasks, and this learning method is called a task-driven dictionary learning. Generally, in this scheme, a shared dictionary and a classifier over the representation coefficients are learned together. However, there is no relationship between a dictionary atom and a class label, so no class-specific representation residuals are obtained so that this learning approach is not ideal for classification.

In class-specific dictionary learning, a dictionary, in which atoms are predefined to correspond to subject class labels, is learned and thus the class-specific reconstruction error could be used for classification [51], [56], [57], [58], [59], [60]. By adding a discriminative reconstruction penalty term in the KSVD model [50], Mairal et al. [58] presented a dictionary learning algorithm for texture segmentation and scene analysis. Yang et al. [57] proposed to learn a structural dictionary by imposing the Fisher discrimination criterion on the sparse coding coefficients to enhance the class discrimination power. In [59], by adding non-negative penalties to both dictionary atoms and representation coefficients, Castrodad and Sapiro proposed to learn a set of action-specific dictionaries. Ramirez et al. [56] introduced an incoherence promotion term to the dictionary learning model for ensuring the dictionaries

representing different classes to be as independent as possible. Wang et al. [51] presented a modified sparse framework to learn a dictionary by minimizing the similarity constraint term and the dictionary incoherence term.

Hybrid dictionary learning which combines shared dictionary learning and class-specific dictionary learning has been proposed. Zhou et al. [61] proposed to learn a hybrid dictionary by using a Fisher-like penalty term on the coding coefficients. Kong et al. [62] proposed to learn a hybrid dictionary by introducing coherence penalty terms on different sub-dictionaries. Shen et al. [63] proposed to learn a dictionary with a hierarchical category structure instead of a flat category structure. However, how to balance the shared part and class-specific part in a hybrid dictionary is not a trivial task.

III. Feature Extraction and Selection

We aim to develop an effective feature representation method, which can generate a more powerful and robust local descriptor to capture the cues for classifying action types in video sequences. In our recent work [24], we proposed a novel Motion Weber Local Descriptor (MoWLD) to extract the low-level features of a query video in spatio-temporal domains. Then, a kernel density estimation (KDE) [24] process was performed in order to eliminate the redundant and irrelevant features. The original WLD with motion information has demonstrated superior performance on action recognition. However, it produces zero response for relatively flat areas.

In this section, for integrity purpose, we first introduce the original WLD, and then propose a more discriminative IWLD. Lastly, similar to the work to construct MoWLD shown in [24], details of constructing MoIWLD are presented.

A. The Original WLD

Weber's law indicates a fact that, for a stimulus, the ratio between the smallest perceptual change and the background is a constant, and it implies that stimuli are not perceived in absolute terms but in relative terms. In another word, only if the ratio of the change of a stimulus to the original stimulus is big enough, this change can be recognized. Most of us can easily catch a whispered voice in a quiet room, but in a noisy condition we may not notice someone shouting in our ear. This is the essence of the Weber's law.

Inspired by Weber's law, Chen et al. [21] proposed a local image descriptor named Weber Local Descriptor (WLD) for the task of face recognition. The descriptor consists of two components, i.e., differential excitation (magnitude) and orientation, which are defined in Eqs. 1 and 2 below [21], [22]. In this paper, we use x_c to denote the central pixel (and also the intensity of the central pixel when there is no confusion), and x_i $(i = 0, 1, \ldots, p-1)$ to represent the p neighboring pixels of x_c (and also their intensities when there is no confusion).

WLD Magnitude ξ_m:

$$\xi_m(x_c) = \arctan(\alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}), \tag{1}$$

where the $arctangent$ function is used to prevent the output from being too large and thus partially suppressing the side-effect of noise, and α is a parameter used to adjust the intensity difference between neighboring pixels.

WLD Orientation ξ_o:

$$\xi_o(x_c) = \arctan(\frac{x_1 - x_5}{x_3 - x_7}), \tag{2}$$

where $x_1 - x_5$ and $x_3 - x_7$ indicate the intensity difference of two neighboring pixels of x_c in vertical and horizontal direction, respectively.

According to [21], ξ_m and ξ_o are then linearly quantized into T (in our experiments, T is set to 12) dominant differential magnitudes and orientations, respectively.

WLD first computes the salient micro-patterns by differential excitation, and then builds statistics on these salient patterns along the gradient orientation of the current pixel. In this way, the salient variations within an image can be found to simulate the perception pattern of human beings. Both differential excitation and differential orientation have been proved to be illumination insensitive and computationally efficient [22]. A 2D concatenated histogram based on the differential excitation and orientation can then be constructed to represent each image. As shown in [21] and [23], each row of the 2D WLD histogram corresponds to a dominant differential excitation $\xi_m(x_c)$, and each column corresponds to a dominant orientation $\xi_o(x_c)$. The original WLD histogram [21], [23], [22] denotes the frequency of a certain dominant differential excitation on a certain dominant orientation.

As seen from its definition, WLD is a kind of dense descriptor which is computed for every pixel and depends on the local intensity variation and the central pixel's intensity. This descriptor employs the advantages of both SIFT in terms of computing the histogram using the gradient and its orientation, and LBP in terms of computational efficiency and smaller support regions. Different from SIFT and LBP, WLD depends on both the local intensity variation and the magnitude of the center pixel's intensity. Since WLD is computed around a relatively small region (e.g., 3×3), while SIFT is computed around a relatively large region (e.g., 16×16), the description granularity of WLD is much smaller than that of SIFT. That is to say, WLD is computed in a finer granularity than SIFT. The smaller size of the support regions for WLD enables WLD to capture more local salient patterns. LBP cannot describe variation extent to the center pixel by its differential excitation, while WLD reflects definite orientation information by the statistic of gradient orientations in local regions.

B. Improved WLD (IWLD)

All of the methods described in [21], [23], [22] utilized Eq. 1 to compute Weber magnitudes, and neglected the varying orientations of eight differences. Moreover, when encountered a point on a rather flat area, the ξ_m value in Eq. 1 may become zero, although the area may contain certain levels of textures (see Fig. 3).

In order to address this problem, we propose an improved WLD (IWLD) described as follows.

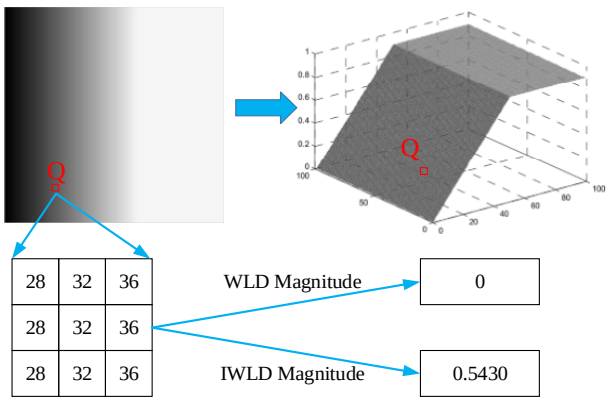


Figure 3: WLD and IWLD response of a point Q within a non-flat area.

Denote $X = (\frac{x_0-x_c}{x_c}, \frac{x_1-x_c}{x_c}, \cdots, \frac{x_{p-1}-x_c}{x_c})^T$ and $I = (1, 1, \cdots, 1)^T$. The WLD magnitude in Eq. 1 can be re-written as:

$$\xi_m(x_c) = \arctan(\alpha < X, I >), \tag{3}$$

where $<,>$ is the inner product operator.

Let us denote the angle between x direction and $x_i - x_c$ (i.e., the vector formed by the subtraction of the central pixel's coordinates from the coordinates of the i-th neighbor of x_c) as θ_i, for $i = 0, 1, \cdots, p-1$. Then, we can use the projections of X's components to x and y directions to compute the effective contributions of the components for the definitions of IWLD magnitudes in x and y directions, in Eqs. 4 and 5, respectively.

IWLD magnitude in x direction:

$$\bar{\xi}_{m_x}(x_c) = \arctan(\alpha < X, J_x >), \tag{4}$$

where $J_x = (\cos\theta_0, \cos\theta_1, \cdots, \cos\theta_{p-1})^T$.

IWLD magnitude in y direction:

$$\bar{\xi}_{m_y}(x_c) = \arctan(\alpha < X, J_y >), \tag{5}$$

where $J_y = (\sin\theta_0, \sin\theta_1, \cdots, \sin\theta_{p-1})^T$.

Then, we define the IWLD magnitude as:

$$\bar{\xi}_m(x_c) = \sqrt{(\arctan\alpha < X, J_x >)^2 + (\arctan\alpha < X, J_y >)^2}, \tag{6}$$

and the IWLD Orientation is defined as

$$\bar{\xi}_o(x_c) = \arctan(\frac{\bar{\xi}_{m_y}(x_c)}{\bar{\xi}_{m_x}(x_c)}). \tag{7}$$

Comparing the IWLD with the original WLD, we can see that the IWLD represents local patterns more effectively and accurately in the patch. As shown in Fig. 3, when a point falls in a non-flat area, Eq. 6 gives a non-zero response.

C. IWLD Histogram

Collecting the values of Weber Magnitudes and Orientations in IWLD on an image region, we can build an IWLD histogram for the region. However, the resultant IWLD histogram is not rotation-invariant, and is sensitive to partial occlusion and deformation. Aiming to address these two problems, we propose to rebuild the IWLD histogram by aggregating the IWLD histograms of neighboring regions and also aligning these histograms to their dominant orientation, as below:

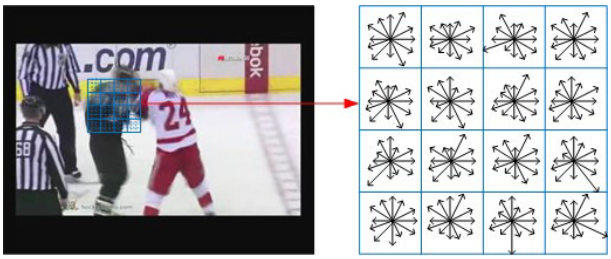


Figure 4: Grid aggregation for IWLD feature descriptors. Pixels in a neighborhood are grouped into 4×4 grids, each containing $3 \times 3 = 9$ pixels. An orientation histogram with 12 bins is formed for each grid resulting into a 192-element vector for the neighborhood.

1) The IWLD magnitude and orientation are calculated according to Eqs. 6 and 7 for every pixel in a region of a Gaussian-blurred image F.
2) The IWLD orientation is quantified into 12 dominant bins by using the non-linear quantization method in [23], with each bin covering 30 degrees. An orientation histogram with 12 bins is then formed.
3) An aggregated histogram of IWLD magnitudes and orientations from all neighboring regions is formed as the feature representation of a local appearance. This representation makes the IWLD descriptor be tolerant to partial occlusion and deformation.
4) When an interest point is detected, the corresponding dominant orientation can be determined. The locations of all IWLD magnitudes in the neighboring regions are then rotated according to the dominant orientation to achieve the rotation invariance.
5) Pixels in the neighboring region are normalized into 144 (16×9) elements, which are grouped into 16 (4×4) grids of 9 pixels each as shown in Fig. 4. Each grid has its own IWLD orientation histogram describing the orientation of the sub-region. This results in an IWLD feature vector of 192 dimensions ($4 \times 4 \times 12 = 192$).

Fig. 4 illustrates the idea of the IWLD histogram grid aggregation. Pixels in a neighborhood are grouped into 4×4 blocks, each containing 3×3 pixels. By constructing the IWLD feature vector in this way, we can obtain a more discriminative descriptor of $16 \times 12 = 192$ dimensions in total.

D. Motion IWLD (MoIWLD)

The IWLD feature descriptor only describes the properties of still images and carries no motion information of a video. Therefore, the detected candidate points are distinctive in appearance only, but are independent of the motions or actions in a video. Clearly, motion information is essential for extracting interest points to eliminate irrelevant information for action recognition. In our previously proposed MoWLD algorithm [24], we adopted the widely used optical flow approach to detect the movement within an image region. A

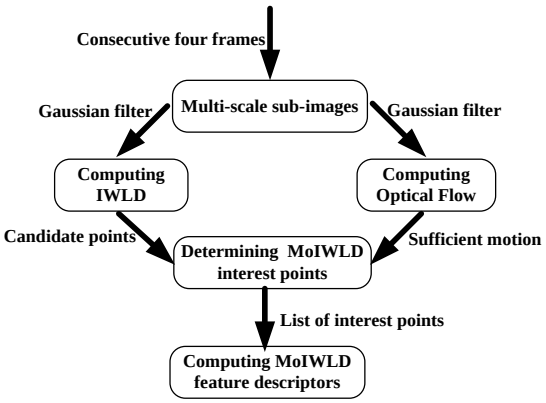


Figure 5: The flowchart of our MoIWLD algorithm. Four consecutive frames are input to compute the IWLD and optical flow. Candidate points with sufficient motion are determined as the MoIWLD interest points, for which MoIWLD features are extracted.

local extreme among all IWLD feature points can only become an interest point if it has sufficient motion in optical flow field.

The optical flow approach detects the movement of a region by calculating the temporal differences of the region in two consecutive frames. Compared to video cuboids or volumes, optical flow explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance changes over a time period. Explicitly measuring motion is beneficial for recognizing actions. In our work, to add motion information into the IWLD, we apply the same aggregation idea as that for generating the IWLD descriptor to integrate the optical flow of every grid in a region into the IWLD descriptor and form the motion IWLD (MoIWLD).

Surveillance videos are typically captured by stationary cameras, so the direction of a movement generated by violent actions is typically irregular and variable, and it can be used to distinguish the violent actions from normal actions. The dominant orientation feature is a main difference between IWLD and an optical flow. The magnitude and direction of a movement can be used to construct an optical flow histogram similar to the construction of an IWLD histogram. The orientations of optical flows in each grid are normalized into 12 directions and an optical flow histogram of 12 bins is constructed for each grid. For a 4×4 grid neighborhood, this results in an aggregated optical flow histogram of a dimension of $4 \times 4 \times 12 = 192$.

Furthermore, to create a more robust descriptor that contains both temporal and contextual information, we also integrate the IWLD and optical flow histograms of three previous frames into the descriptor. Therefore, all four sets of the aggregated IWLD and optical flow histograms on each Gaussian filtered image are concatenated to form the MoIWLD descriptor (as shown in Fig. 5), which now has 1536 ($4 \times 2 \times 192 = 1536$) dimensions.

E. Multi-Scale MoIWLD

The IWLD features described above are extracted from small patches (i.e., grids) of the same size (i.e., sclae) of 3×3 pixels. However, sub-images containing various human structures may have different sizes. so we conduct a multi-scale image analysis to extract more discriminative and robust features of different human local structures. We adopt the multi-scale feature analysis approach [21] to develop a multi-scale IWLD for each image on a Gaussian pyramid (see Fig. 5). The multi-scale IWLD provides local salient patterns around a point (e.g., the central pixel in Fig. 6) at various scales. It is computed on a set of square, symmetric neighborhoods consisting of various number of pixels, as shown in Fig. 6.

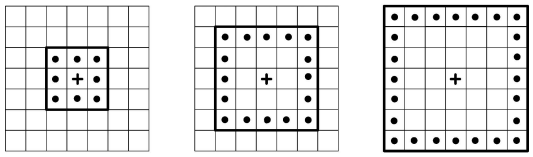


Figure 6: A set of square, symmetric neighborhoods consisting of various numbers of pixels.

Furthermore, multi-scale optical flows are also computed on each Gaussian filtered image. Multiple-scale optical flows are calculated according to the IWLD scales. A local extreme of the multi-scale IWLD feature points can only become an interest point if it also has sufficient motion in terms of multi-scale optical flows in the Gaussian pyramid. We assume that a complicated action can be represented by the combination of a reasonable number of interest points. Therefore, we do not assign strong constraints to spatio-temporal interest points. As long as the candidate interest points have made movements with distances larger than a minimum value, they can be deemed as MoIWLD interest points. The extracted MoIWLD interest points are invariant to scale and rotation in the spatial domain but they are not invariant to scale in the temporal domain. The multi-scale MoIWLD selects distinctive interest points with sufficient motions from which humans can 'see' the actions happened at the corresponding points and machines can learn an action model.

Since the multi-scale MoIWLD is based on IWLD and optical flow, it has the advantages described as follows. Instead of combining the histograms obtained from IWLD and optical flows, we build a single feature descriptor through the 'early fusion' that concatenates both histograms into one vector. This single future descriptor captures, at the same time, the appearance and motion information that are essential for classifying actions. Moreover, the MoIWLD descriptor captures local appearance using an aggregated histogram of gradients in neighboring regions, so it is tolerant to partial occlusion and deformation. Furthermore, when an interest point is detected, a dominant orientation is calculated and all gradients in the neighborhood are rotated according to the dominant orientation. Therefore, the multi-scale MoIWLD is rotation-invariant.

F. KDE-based Feature Selection

To improve both performance and computational efficiency, we employ a nonparametric density estimation method to

select the most representative features from the extracted MoIWLD features.

Nonparametric density estimation, an important tool for statistical analysis of data, is an alternative to parametric approaches, in which one specifies a model up to a small number of parameters and then estimates the parameters via the likelihood principle. Being a classic non-parametric density estimation method, the Kernel Density Estimation (KDE)-based feature selection method [30], [31] depends only on data samples to get an estimation and does not need a prior knowledge of the data distribution. As long as there are enough samples, KDE can give a satisfactory approximation of an underlying distribution no matter whether it is regular or irregular, and single modal or multi-modal. Therefore, we employ KDE to find the discriminative features which have multi-modal distributions.

Suppose $x_1, x_2, \ldots, x_N$ are N independent and identically distributed (i.i.d.) data of a one-dimensional random variable x. KDE infers the probability density function (PDF) of x by centering a kernel function $K(x)$ at each data point x_i as:

$$f_h(x) = \frac{1}{hN} \sum_{i=1}^{N} K(\frac{x - x_i}{h}), \tag{8}$$

where h is a smoothing parameter, named as bandwidth, which can be adaptively chosen using the method proposed in [31].

In order to reduce the dimension of the MoIWLD feature space, we use KDE to obtain a smooth probability density function based on our training data. However, the common Gaussian kernel density estimator [30] lacks local adaptivity, and this often results in a higher sensitivity to outliers. Therefore, an adaptive kernel, like the one discussed in [31], is chosen to improve local adaptivity and reduce bias.

If the probability density function of a feature is bimodal or multimodal, this feature is considered to be more discriminative than those of only a single mode. We estimate the probability density function (PDF) of each feature on the original 1536 MoIWLD features. According to the number of modes, we sort the 1536 MoIWLD features in descending order. Finally, the first 550 features are selected to form the reduced MoIWLD, which is more effective than the original ones (as shown in Fig. 7).

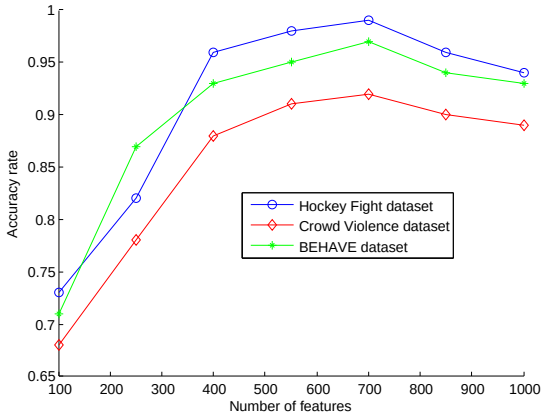


Figure 7: The variation of the performance with the change of number of features on three benchmark datasets.

IV. Sparse Representation and Dictionary Learning

Once we have obtained action feature representation, action classification can then be performed by employing pattern recognition technologies. Wright et al. [27] proposed a general classification scheme based on sparse representation and applied it for robust face recognition. However, how to learn a discriminative dictionary for both sparse data representation and classification is still a challenging problem.

Note that, the conventional Sparse Representation based Classification (SRC) scheme does not consider the relationship between the sub-dictionaries of two classes. It is used to minimize reconstruction error instead of classification error. In addition, the discriminative information in training samples cannot be sufficiently exploited by such a naively supervised dictionary learning (DL) method.

As an improvement, in our proposed model (detailed as below), each sample is locally approximated by a linear combination of its nearby samples and a classification error is introduced. The proposed sparse representation model selects the samples that can best represent their own action classes as the bases of their class-specific dictionaries. Specifically, we combine the reconstruction error with a classification error to form a unified objective function. The learned dictionary encourages the signals from the same class to have similar sparse codes and those from different classes to have dissimilar sparse codes to achieve accurate classification results.

A. Original SRC Model

In [27], Wright et al. proposed a general sparse representation based classification (SRC) scheme, in which 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 $D = [A_1, A_2, \cdots, A_K]$ 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: $\hat{x} = \arg\min_x \{\|y - Dx\|_2^2 + \gamma \|x\|_1\}$, where γ is a scalar constant.

(3) Label the test sample y by: $Label(y) = \arg\min_i \{e_i\}$, where $e_i = \left\|y - A_i \hat{\alpha}^i\right\|_2^2$, 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 [27] showed that sparse representation is naturally discriminative.

B. Proposed SRC Model

In our proposed model, two terms, i.e., the representation-constrained term and the coefficient adjustment term, are

introduced and described below to ensure that the learned dictionary is sufficiently discriminative. The representation-constrained term is utilized to enforce 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 adjustment term is utilized to enforce the class-specific sub-dictionary to have a poor capability when reconstructing a query image using training samples with different class labels. Based on these two terms, a classification scheme is then developed to exploit the discriminative information.

The samples in different action classes have different MoIWLD descriptors, so we will focus on class-specific dictionary learning (DL). In a 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. By representing a test sample over the learned dictionary D, the representation residual associated with each class can be employed to classify it, as in the SRC method.

Given training MoIWLD 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 MoIWLD 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 by $D = [d_1, d_2, \cdots, d_k] \in \mathbb{R}^{n \times k}$ ($k > n$ and $k \ll N$). We propose the following modified sparse model:

$$\begin{aligned} \langle D, W, Z \rangle = \arg \min_{D,W,Z} \{ &\|A - DZ\|_F^2 + \lambda_1 \|Z\|_1 + \\ &\lambda_2 \|Z - m\|_F^2 + \gamma_1 \|WZ - B\|_F^2 + \gamma_2 \|W\|_F^2 \}, \\ &s.t. \|d_c\|_2 \leq 1, \forall c \in \{1, 2, \cdots, k\} \end{aligned} \tag{9}$$

where $Z = [Z_1, Z_2, \cdots, Z_i] \in \mathbb{R}^{k \times N}$ 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.

The first two terms on the right hand side of Eq. 9 can be seen as the basic model of class-specific DL. Different from the conventional sparse model SRC in [27], in our model, the representation-constrained 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. 9.

1) Representation-constrained term: Z_i denotes the sparse coefficients of A_i over dictionary D, so $A_i \approx DZ_i$. Since Z_i is associated with class i, it is naturally expected that Z_i can be well represented by only m_i because m_i denotes the mean vector of Z_i in class i. Therefore, there should exist a Z_i such that $\|Z_i - m_i\|_F^2$ is small. This term can control the reconstruction error of coefficients Z_i. On the other hand, W denotes the matrix of classifier parameters and B records the class labels, so $\gamma_1 \|WZ - B\|_F^2$ represents the classification error, and minimizing this term is to minimize the classification error.

Overall, minimizing the representation-constrained term defined by $\phi = \lambda_2 \|Z - m\|_F^2 + \gamma_1 \|WZ - B\|_F^2$ minimizes both the reconstruction error and classification error.

2) Coefficient adjustment term: Given a test sample, Wright et al. [27] suggested that its sparse representation could be found by an SRC scheme, and in the sparse coefficients recovered by SRC, the largest coefficients were associated with the training samples that had the same class label as the test sample. It implies that the test sample can be approximated by a weighted linear combination of its own training samples with these largest coefficients. Likewise, in our proposed class-specific dictionary learning, because W denotes the set of classifier parameters, minimizing $\|W\|_F^2$ is for the class-specific DL to reach the minimum classification error.

On the other hand, minimizing the coefficient adjustment term and representation-constrained term is efficient for classification, it allows feature sharing among the 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 different class over D are associated with the corresponding different sub-dictionary. Therefore, the coefficient adjustment term enforces a label consistency constraint on the sparse codes.

C. Supervised Class-Specific Dictionary Learning

Although the objective function in Eq. 9 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. 9 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. 9 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. 9 can be replaced by:

$$\begin{aligned} \|A - DZ\|_F^2 + \lambda_1 \|Z\|_1 + \lambda_2 \|Z - m\|_F^2 \\ + \gamma_1 \|WZ - B\|_F^2 \end{aligned} \tag{10}$$

and Eq. 10 can be rewritten as:

$$\begin{aligned} \langle Z_i \rangle = \arg \min_{Z_i} \{ &\|A_i - DZ_i\|_2^2 + \lambda_1 \|Z_i\|_2^2 + \\ &\lambda_2 \|Z_i - m_i\|_2^2 + \gamma_1 \|WZ_i - b_i\|_2^2 \}. \end{aligned} \tag{11}$$

By solving Eq. 11, we have:

$$Z_i = \{D^T D + (\lambda_1 + \lambda_2)I + \gamma_1 W^T W\}^{-1}(D^T A_i + \lambda_2 m_i + \gamma_1 W^T b_i). \quad (12)$$

Updating D**:** When Z and W are fixed, Eq. 9 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. 9 can be replaced by:

$$\langle D \rangle = \arg\min_D \|A - DZ\|_F^2, \quad s.t. \|d_c\|_2 = 1, \forall c \in \{1, 2, \cdots, k\}. \quad (13)$$

The subproblem in Eq. 13 can be solved effectively by the Lagrange dual method [64].

Updating W**:** When D and Z are fixed, Eq. 9 can be replaced by:

$$\arg\min_W \{\gamma_1 \|WZ - B\|_F^2 + \gamma_2 \|W\|_F^2\} \quad (14)$$

Let $W = [W_1, W_2, \cdots, W_i]$, Eq. 14 can be rewritten as:

$$\arg\min_{w_i} \{\gamma_1 (\|W_i Z_i - b_i\|_2^2 + \frac{\gamma_2}{\gamma_1} \|W_i\|_2^2)\} \quad (15)$$

In Eq. 15, $\|W_i Z_i - b_i\|_2^2 + \frac{\gamma_2}{\gamma_1} \|W_i\|_2^2$ can be further rewritten as:

$$\|W_i Z_i - b_i\|_2^2 + \rho \|W_i\|_2^2 \quad (16)$$

Denote $\rho = \frac{\gamma_2}{\gamma_1}$. Obviously, Eq. 16 can be solved using the least square method. Thus we can get the following solution:

$$W_i = b_i Z_i^T (Z_i Z_i^T + \frac{\gamma_2}{\gamma_1} I)^{-1} \quad (17)$$

Thus, based on the above equations, we can get the optimized values of all parameters for Eq. 9.

D. 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:

$$\hat{\alpha} = \arg\min_\alpha \{\|y - D\alpha\|_F^2 + \gamma \|\alpha\|_2\} \quad (18)$$

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 big. 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:

$$l = W\hat{\alpha}. \quad (19)$$

Lastly, we simply use the linear predictive classifier to estimate the label i of vector l, and the label index corresponds to the largest element of the vector l.

V. Experiments and Results

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, including the BoW based methods, the RVD violence detection method in [14], the Appearance and Motion DeepNet (AMDN) method in [48], the Violent Flow (ViF) method in [13], the method in [27] and our recently published method in [24]. To evaluate the classification accuracy, we employ the 5-fold cross validation test on each dataset. Results are reported with mean prediction accuracy (ACC)$\pm$ standard deviation (SD), as well as the area under the ROC curve (AUC). Furthermore, in our proposed model, there are two stages, i.e., dictionary learning stage and classification stage. In the dictionary learning stage, we set $\lambda_1 = 0.005, \lambda_2 = 3, \gamma_1 = 1, \gamma_2 = 0.1$; and in classification stage, we set $\gamma = 0.01$. Next, we first briefly introduce the three benchmark datasets, and then present experimental results with discussion.

A. Datasets

Experiments of our method were conducted on three challenging benchmark datasets, i.e., the Hockey Fight dataset [13], the BEHAVE dataset [65], and the Crowd Violence dataset [13].

The Hockey Fight dataset contains 1000 video clips of actions from hockey games of the National Hockey League (NHL), of which 500 are manually labeled as fight and others are labeled as non-fight. Each clip consists of 50 frames (with resolution of 360×288 pixels).

The BEHAVE dataset contains more than $200,000$ frames (with 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 and manually label them as violence or non-violence. Each clip consists of at least 100 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, presenting 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 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 behaviors and the other half presents non-violent crowd behaviors.

B. Results and Discussion

The proposed SRC vs. the original SRC.

In order to demonstrate the performance of the proposed sparse classification model proposed in Section IV, we first compare our method with the original SRC classification algorithm on the three databases. The average results shown in Fig. 8 demonstrate that our proposed model achieves a higher classification rate than the original SRC algorithm.

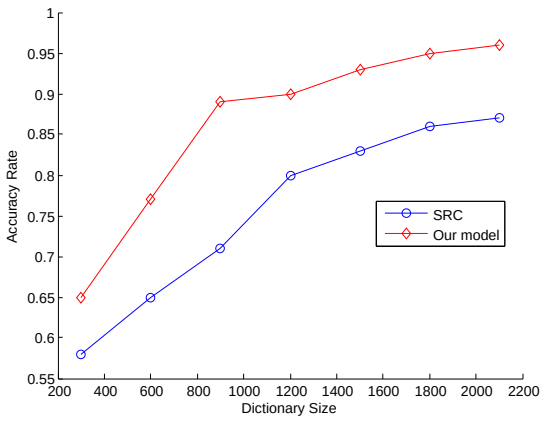


Figure 8: Performance comparison of our sparse classification model with the original SRC classification model on benchmark datasets.

Table I: Comparison on average classification time (in ms) on the three benchmarking datasets.

Method	Hockey Fight	BEHAVE	Crowd Violence
SRC	100.19	190.81	94.49
Proposed method	4.67	8.38	4.15

Also, the comparison result on their average computation times for classifying images is shown in Table I. As our algorithm learns a single overcomplete dictionary and an optimal linear classifier jointly, compared with the SRC model, on average it is around 25 times more efficient.

Results on the Hockey Fight dataset.

We empirically demonstrate the superior performance of our ideas of MoWLD and MoIWLD over existing popular feature descriptors, including HOG, HOF, and MoSIFT [2]. Table II shows the results on the Hockey Fight dataset, where "HOG+BoW", "HOF+BoW" and "MoSIFT+BoW" refer to HOG, HOF and MoSIFT when paired with the BoW model.

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. MoIWLD has further improved the results, so MoIWLD is more discriminative and effective.

Furthermore, to assess the impact of dictionary size on classification accuracy, we have run the experiments with dictionaries of different sizes. As shown in Table II, with the increase of the dictionary size, the performance begins to rise and then stays stable. This phenomenon indicates that selecting an appropriate dictionary size is significant to both accuracy and computational efficiency.

Table III shows the results obtained by the proposed approach after adopting the BoW approach and the proposed sparse classification model into the MoIWLD approach. The dictionary size is fixed to 1800 in this experiment. We compare the results with not only HOG, HOF, HNF, MoSIFT, MoWLD and MoIWLD paired with BoW, but also ViF, RVD and AMDN. The RVD method [14] adopts a Gaussian Model of Optical Flow (GMOF) to extract candidate violence regions, which has reduced many noise disturbances, so its performance is better than the BoW-based approaches. The AMDN approach [48] utilizes optical flow as the input image feature, and there exist many redundant and interference features, so its performance is not very good. It can be seen from Table III that using our proposed MoIWLD descriptor together with the proposed sparse classification model has performed the best. Furthermore, our approach outperforms the competing SRC algorithm and MoWLD+Sparce Coding with the same size of dictionary. This is due to the fact that our proposed supervised class-specific dictionary learning framework incorporates representation-constrained and coefficient adjustment terms resulting in the highest recognition rate.

Table III: Detection results on the Hockey Fight dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [2]	88.77±0.73%	0.9123
HOF+BoW [2]	86.07±0.59%	0.8843
HNF+BoW [2]	89.27±0.79%	0.9294
ViF [13]	90.07±0.99%	0.9429
MoSIFT+BoW [2]	88.8±0.75%	0.9052
MoWLD+BoW	89.28±0.93%	0.9112
MoIWLD+BoW	91.8±1.03%	0.9412
RVD [14]	92.1±1.01%	0.9496
AMDN [48]	89.7±1.13%	0.9198
SRC [27]	94.4±1.07%	0.9623
MoWLD+Sparse Coding [24]	93.8±1.08%	0.9618
Proposed method	96.8±1.04%	0.9808

Results on the BEHAVE dataset. We compare our approaches with the state-of-the-art approaches implemented by us on the BEHAVE dataset, where 20 clips of this dataset are randomly picked for training. Table IV presents the results obtained with the above mentioned methods on this dataset. The dictionary size is fixed to 1800 in this set of experiments. As it can be seen from the table, our proposed SRC based method outperforms other approaches. This again demonstrates that the proposed approach is significantly superior in performance to all other approaches. The performance of the RVD method [14] and the AMDN [48] on this dataset is consistent with their performance on the Hockey Fight dataset. Furthermore, our MoIWLD combined with our proposed sparse classification method outperforms SRC methods. It validates that the representation-constrained term and coefficient adjustment term can improve the discriminative ability of sparse representation model. The results on this dataset demonstrate that our algorithm is also 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. 9).

Results on the Crowd Violence dataset. This dataset is more challenging than the other two datasets because it contains many crowded scenes. The set contains 246 clips divided into five splits, each containing 123 violent and 123 non-violent scenes. Table V presents the results obtained using different methods mentioned above on this dataset. The dictionary size is again fixed to 1800 in this set of experiments.

Table II: Accuracy comparison of various feature representation paired with BoW for violence detection on the Hockey Fight dataset.

Vocabulary	HOG+BoW [2]	HOF+BoW [2]	MoSIFT+BoW [2]	MoWLD+BoW	MoIWLD+BoW
300 words	90.8%	87.2%	90.4%	91.3%	92.2
600 words	91.4%	87.4%	90.5%	91.7%	92.8
900 words	91.6%	88.5%	90.9%	91.9%	93.2
1200 words	91.4%	88.3%	90.6%	91.8%	93.7
1500 words	91.2%	88.2%	90.6%	91.5%	93.9
1800 words	91.2%	88.4%	90.5%	91.1%	94.1
2100 words	90.7%	87.9%	90.4%	91.2%	94.3

Table IV: Detection results on the BEHAVE dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [2]	58.97±0.34%	0.6394
HOF+BoW [2]	60.03±0.28%	0.5923
HNF+BoW [2]	58.24±0.31%	0.6113
ViF [2]	83.62±0.19%	0.8632
MoSIFT+BoW [2]	62.78±0.23%	0.6679
MoWLD+BoW	81.65±0.18%	0.8324
MoIWLD+BoW	81.98±0.15%	0.8415
RVD [14]	85.29±0.16%	0.8878
AMDN [48]	84.22±0.17%	0.8562
SRC [27]	82.7±0.14%	0.8538
MoWLD+Sparse Coding [24]	87.07±0.13%	0.8928
Proposed method	88.83±0.11%	0.9108

Table V: Detection results on the Crowd Violence dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [2]	57.98±0.37%	0.6252
HOF+BoW [2]	58.71±0.12%	0.5931
HNF+BoW [2]	57.05±0.32%	0.6154
ViF [2]	82.13±0.21%	0.8595
MoSIFT+BoW [2]	57.09±0.37%	0.6073
MoWLD+BoW	88.16±0.19%	0.9028
MoIWLD+BoW	88.78±0.19%	0.9109
RVD [14]	82.89±0.19%	0.8559
AMDN [48]	84.72±0.17%	0.8891
SRC [27]	89.6±0.18%	0.9288
MoWLD+Sparse Coding [24]	89.38±0.13%	0.9398
Proposed method	93.19±0.12%	0.9508

Figure 9: Examples of false alarms on the BEHAVE dataset.

Figure 10: Examples of false alarms on the Crowd Violence dataset.

In this dataset, due to more crowded scenes, the detection rate of RVD method decreases. On the contrary, the performance of AMDN method is still very stable. However, because of the introduction of optical flow noise, AMDN's performance is not very good. Our proposed sparse classification-based approach still outperforms other approaches. MoIWLD descriptor is still significantly superior in performance to HOG, HOF, HNF, RVD, AMDN and MoWLD. It confirms that our proposed MoIWLD is a more effective descriptor for describing action feature. Consistent with the results on the previous two datasets, our MoIWLD combined with the proposed sparse classification method 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. 10) 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.

VI. Conclusion

This paper has targeted on effective spatio-temporal interest point detection and feature representation for detecting violent behaviours in a real video scene. Following our recently proposed Motion Weber Local Descriptor (MoWLD), two major improvements have been made. First, considering the shortcomings of the well-known WLD, we have proposed the improved WLD, i.e., IWLD, and then proposed to extend the IWLD by adding a temporal component to the appearance descriptor to obtain MoIWLD. MoIWLD implicitly captures local motion information together low-level image appearance information. Secondly, a modified sparse model has been proposed to learn a dictionary for classification. In the proposed

sparse model, the representation-constrained term and the coefficient incoherence term have been introduced to ensure the learned dictionary to obtain a powerful discriminative ability. With this learned dictionary, both the representation residual and the representation coefficients are discriminative. Based on the proposed SRC, we have presented a corresponding classification scheme. Experimental results on three benchmark datasets have demonstrated the superiority of the proposed approach over the state-of-the-art approaches.

ACKNOWLEDGMENT

This research was partly supported by NSFC, China (No: 61273258, 61375048, 61170109).

REFERENCES

[1] L R Huesmann, J Moise-Titus, C L Podolski, and L D Eron, "Longitudinal relations between childrens exposure to tv violence and their aggressive and violent behavior in young adulthood," *Developmental Psychology*, vol. 39, no. 2, pp. 201–221, 2003.

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

[3] J. K Aggarwal and M. S Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[4] O. P. Popoola and Kejun Wang, "Video-based abnormal human behavior recognition - a review," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 865–878, 2012.

[5] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *European Conference on Computer Vision (ECCV), 2008*. Springer, 2008, pp. 548–561.

[6] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1975–1981.

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

[8] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," *Computer Vision and Pattern Recognition (CVPR), 1992 IEEE Conference on*, pp. 379–385, 1992.

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

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

[11] A. Datta, Shah M., and Da Vitoria Lobo N., "Person-on-person violence detection in video data," *Proceedings of IEEE International Conference on Image Processing (ICIP2002)*, pp. 433–438, 2002.

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

[13] T. Hassner, Itcher Y., and O. Kliper-Gross, "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)*, pp. 1–6, 2012.

[14] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications*, pp. 1–23, 2015.

[15] F. D. M. de Souza, G. C. Chavez, E. A. do Valle, and A. de A. Araujo, "Violence detection in video using spatio-temporal features," in *SIBGRAPI 2010, Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2010, pp. 224–230.

[16] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, "Action recognition via structured codebook construction," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 546–555, 2014.

[17] D. Lowe, "Distinctive image features from scale invariant key points," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3517–3524, 2010.

[19] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," in *Tech. rep, Carnegie Mellon University*. Carnegie Mellon University, 2009, pp. 1–10.

[20] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *Proc. Brit. Mach. Vis. Conf*, pp. 1–11, 2009.

[21] J. Chen, S. Shan, Ch He, G. Zhao, X. Chen, and W. Gao, "Wld: A robust local image descriptor," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1705–1720, 2010.

[22] B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on weber's law with application to face recognition," in *Signal Processing Letters, IEEE*. IEEE, 2011, vol. 18, pp. 462–465.

[23] S. Li, D. Gong, and Y. Yuan, "Face recognition using weber local descriptors," in *Neurocomputing*. Elsevier, 2013, vol. 122, pp. 272–283.

[24] T. Zhang, W. Jia, J. Yang, and X. He, "Mowld: a robust motion image descriptor for violence detection," *Multimedia Tools and Applications*, pp. 1–20, 2015.

[25] N. Gkalelis, A. Tefas, and I. Pitas, "Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition," *IEEE Trans. Circuits Syst. Video Technol*, pp. 1511–1521, 2008.

[26] A. Iosifidis, A. Tefas, and A. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Netw. Learning Syst*, pp. 412–424, 2012.

[27] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis Mach. Intell*, pp. 210–227, 2009.

[28] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[29] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," *10th Asian Conference on Computer Vision, ACCV2010*, pp. 660–671, 2011.

[30] Z. I. Botev, J. F. Grotowski, and D. P Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[31] X. Geng, C. Yu, and G. Hu, "Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting," *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 112–117, 2012.

[32] T. Ojala, M. Pietikfinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[33] V. Sarvesh and A. Anupam, "A survey on activity recognition and behavior understanding in surveillance video," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[34] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered hmms," *Multimedia, IEEE Transactions on*, vol. 8, no. 3, pp. 509–520, 2006.

[35] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov model," in *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, pp. 955–960. IEEE, 2005.

[36] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa, "Propagation networks for recognition of partially ordered sequential action," *Computer Vision and Pattern Recognition (CVPR), 2004 IEEE Conference on*, pp. 862–869, 2004.

[37] P Dai, H Di, L Dong, L Tao, and G Xu, "Group interaction analysis in dynamic context," in *Systems, Man, and Cybernetics, IEEE Transactions on*. IEEE, 2008, pp. 275–282.

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

[39] AF Bobick and JW Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[40] D. B. Marco, B.and Alberto and S. Lorenzo, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320–329, 2012.

[41] S. Baysal and P. Duygulu, "A line based pose representation for human action recognition," *Signal Processing: Image Communication*, vol. 28, no. 5, pp. 458–471, 2013.

[42] B. Saghafi and D. Rajan, "Human action recognition using pose-based discriminant embedding," *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 96–111, 2012.

[43] A Oikonomopoulos, I Patras, M Pantic, and N Paragios, "Trajectory-based representation of human actions," *Artificial Intelligence for Human Computing*, vol. 44, no. 51, pp. 133–154, 2007.

[44] S Vishwakarma, A Sapre, and A Agrawal, "Action recognition using cuboids of interest points," in *In: IEEE Int. Conf. on Signal Processing, Communications and Computing (ICSPCC)*. 2011, pp. 1–6, IEEE.

[45] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Transactions on Multimedia*, pp. 572–581, 2013.

[46] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph leaning with partial tags and multiple features for image and video annotation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4371–4379, 2015.

[47] J. R. Mehrsan and L. Martin, "Online dominant and anomalous behavior detection in videos," *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2609–2616, 2013.

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

[49] J Mairal, F Bach, and J Ponce, "Task-driven dictionary learning," *IEEE Trans.Pattern Anal. Mach. Intell*, vol. 34, no. 4, pp. 791–804, 2012.

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

[51] H Wang, C Yuan, W Hu, and C Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, vol. 45, no. 11, pp. 3902–3911, 2012.

[52] Z. L Jiang, Z Lin, and L.S Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans.Pattern Anal. Mach. Intell*, vol. 35, no. 11, pp. 791–804, 2013.

[53] J Mairal, F Bach, J Ponce, G Sapiro, and A Zisserman, "Supervised dictionary learning," in *NIPS*. IEEE, 2009, pp. 792–800.

[54] S Bengio, F Pereira, Y Singer, and D Strelow, "Group sparse coding," in *NIPS*. IEEE, 2009, pp. 791–804.

[55] M Aharon, M Elad, and A Bruckstein, "An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE T. SP*, vol. 54, no. 11, pp. 5311–5322, 2006.

[56] I Ramirez, P Sprechmann, and G Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared feature," in *Proc. IEEE Int. Conf. CVPR*. IEEE, 2010, number 11, pp. 3602–3611.

[57] M Yang, L Zhang, X. C Feng, and D Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. ICCV*. IEEE, 2011, number 17, pp. 654–662.

[58] J Mairal, F Bach, J Ponce, G Sapiro, and A Zissserman, "Learning discriminative dictionaries for local image analysis," in *Proc. CVPR*. IEEE, 2008, number 17, pp. 1233–1240.

[59] A Castrodad and G Sapiro, "Sparse modeling of human actions from motion imagery," *Int'l Journal of Computer Vision*, , no. 100, pp. 1–15, 2008.

[60] M Yang, L Zhang, X. C Feng, and D Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int'l Journal of Computer Vision*, , no. 100, pp. 1–15, 2015.

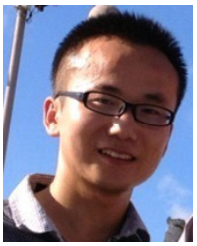
[61] N Zhou and J.P Fan, "Learning inter-related visual dictionary for object recognition," in *Proc. CVPR*. IEEE, 2012, number 17, pp. 3490–3497.

[62] S Kong and D.H Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Proc. ECCV*. Springer, 2012, number 100, pp. 186–199.

[63] L Shen, S.H Wang, G Sun, S.Q Jiang, and Q.M Huang, "Multi-level discriminative dictionary learning towards hierarchical visual categorization," in *Proc. CVPR*. IEEE, 2013, number 16, pp. 1320–1327.

[64] S Cai, W Zuo, and L Zhang, "Support vector guided dictionary learning," Springer, 2014, pp. 183–202.

[65] E. Andrade and R Fisher, "Modelling crowd scenes for event detection," in *In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006, vol. 01, pp. 175–178.



Tao Zhang received his Bachelor degree from Henan Polytechnic University, China in 2008. He is currently a PhD candidate at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His major research interests include visual surveillance, object detection, and pattern analysis.

Wenjing Jia is currently a Lecturer at the Faculty of Engineering and IT and a core research member at the Global Big Data Technologies Centre (GBDTC) at University of Technology Sydney (UTS), Australia. She received her Bachelor degree in Communications Engineering from Changchun Institute of Posts and Telecommunications (now part of the Jilin University) in 1999, her Master degree in Communications and Information System from Fuzhou University in 2002, and her PhD degree in Computing Sciences from UTS in 2007. Her research interests are mainly in the areas of image processing/analysis, computer vision and pattern recognition. Till the date, she has published over 70 quality journal articles and conference papers.

Xiangjian He received the Bachelor of Science degree in Mathematics from Xiamen University in 1982, the Master of Science degree in Applied Mathematics from Fuzhou University in 1986, the Master of Science degree in Information Technology from the Flinders University of South Australia in 1995, and the PhD degree in Computing Sciences from the University of Technology Sydney, Australia in 1999. Currently, he is a full professor and the Director of Computer Vision and Pattern Recognition Laboratory at the Global Big Data Technologies Centre (GBDTC) and a co-leader of the Network Security research team at the Centre for Real-time Information Networks (CRIN) at the University of Technology Sydney. He has many high quality publications and has received various research grants including four national Research Grants awarded by Australian Research Council (ARC) as a Chief Investigator. He is an IEEE Senior Member and an IEEE Signal Processing Society Student Committee member. He has served as a guest editor for various international journals.

Jie Yang received his PhD from the Department of Computer Science, Hamburg University, Germany in 1994. Currently, he is a professor at the Institute of Image Processing and Pattern recognition, Shanghai Jiao Tong University, China. He has led many research projects (e.g., National Science Foundation, 863 National High Tech. Plan), had one book published in Germany, and authored more than 200 journal papers. His major research interests are object detection and recognition, data fusion and data mining, and medical image processing.