Hu et al. EURASIP Journal on Advances in Signal Processing (2018) 2018:54
https://doi.org/10.1186/s13634-018-0574-4

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor

Xing Hu[1], Yingping Huang[1*], Qianqian Duan[2], Wenyan Ci[3], Jian Dai[1] and Haima Yang[1]

Abstract

Detecting abnormal events in crowded scenes is an important but challenging task in computer vision. Contextual information is useful for discovering salient events in scenes; however, it cannot be characterized well by commonly used pixel-based descriptors, such as the HOG descriptor. In this paper, we propose contextual gradients between two local regions and then construct a histogram of oriented contextual gradient (HOCG) descriptor for abnormal event detection based on the contextual gradients. The HOCG descriptor is a distribution of contextual gradients of sub-regions in different directions, which can effectively characterize the compositional context of events. We conduct extensive experiments on several public datasets and compare the experimental results using state-of-the-art approaches. Qualitative and quantitative analysis of experimental results demonstrate the effectiveness of the proposed HOCG descriptor.

Keywords: Abnormal event detection, Histogram of oriented contextual gradients

1 Introduction

As one of the key technologies in intelligent video sequence, abnormal event detection (AED) has been actively researched in computer vision due to the increasing concern regarding public security and safety [1]. A large number of cameras have been deployed in many public locations, such as campuses, shopping malls, airports, railway stations, subway stations, and plazas. Traditional video surveillance systems rely on a human operator to monitor scenes and find unusual or irregular events by observing monitor screens. However, watching surveillance video is a labor-intensive task. Therefore, significant efforts have been devoted to AED in video surveillance, and great progress has been made in recent years, which can free operators from exhausting and tedious tasks and thereby significantly save on labor costs.

AED in crowded scenes is fairly challenging due to many factors, such as frequent occlusion, heavy noise, clutter and dynamic scenes, complexity and diversity of events, unpredictability, and contextual dependency. The aim of AED is to find unusual or prohibitive events in a scene and essentially identify the patterns that significantly deviate from a predefined normal pattern of models via pattern recognition [2]. The definition of an abnormal event is heavily dependent on how a normal event is modeled and which event description is applied. Therefore, the key component of a successful AED is event description, which is the organization of raw input data into various constructs that represent abstract properties of video data [3].

Traditional pixel-wise descriptors, such as the HOG descriptor, are normally employed to capture the appearance and/or motion of an event. However, these descriptors are unable to capture contextual information that is useful for discovering saliency events in scenes. In contrast to the pixels statistics, contextual information is macro-structure information, which reflects the composition relationship among regions. Boiman et al. [4] first exploited the distributions of cuboids inside a larger ensemble. The authors proposed an inference by composition (IBC) algorithm to compute the joint probability between a database and query ensemble. Although the algorithm was accurate, the computational burden was heavy. Roshtkhari et al. [5] modeled the spatio-temporal composition of small cuboids in a large

* Correspondence: huangyingping@usst.edu.cn
[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
Full list of author information is available at the end of the article

volume using a probabilistic model and detected abnormal events with irregular compositions in real-time. Li et al. [6] exploited the compositional context under a dictionary learning and sparse coding framework. Gupta et al. [7] proposed a probabilistic model that exploits contextual information for visual action analysis to improve object recognition as well as activity recognition. However, these approaches use a learning framework with complicated inference processes rather than an efficient handcrafted descriptor to capture the contextual information.

In this paper, we extend the traditional gradient from pixel- to context-wise and thus propose a novel HOCG descriptor to capture contextual information for event description. Compared with the traditional HOG descriptor, the HOCG descriptor is the distribution of contextual gradients in different directions, which can reflect the compositional relationship among sub-regions within an event. The proposed HOCG descriptor is compact, flexible, and discriminative. We employ an online sparse reconstruction framework to identify abnormal events with high reconstruction costs. We conduct extensive experiments on different public datasets and make extensive comparisons with the HOG as well as other state-of-the-art descriptors to demonstrate the advantages of the proposed HOCG descriptor.

The main contributions of our work are as follows:

1) We extend the gradient computation from pixel- to context-wise. The contextual gradient is more descriptive and flexible than the pixel-wise gradient and is useful in finding salient events in scenes;

2) We construct a HOCG descriptor for event description in AED using the contextual gradients of sub-regions within an event. The HOCG descriptor can efficiently capture contextual information;

3) We conduct extensive experiments on different datasets to validate the effectiveness of the HOCG descriptor for AED.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of related works regarding event descriptors in AED. Section 3 presents a detailed description of our proposed approach, including the principle of the context-wise gradient, construction of the HOCG descriptor, and AED using the HOCG descriptor. Experiments and results analysis are presented in Section 4, and Section 5 concludes the paper.

2 Related works

Event description is one of the main topics of AED research and has a great impact on detection performance. Normally, we must transform raw video into a specific feature space in which abnormal events can become more salient [8]. Using an effective descriptor can capture important information and decrease intra-class variations, which are helpful for achieving a perfect performance. In previous works, the commonly used event descriptors can be approximately classified as follows:

1) Trajectory-wise descriptor

A trajectory-wise descriptor is high-level and robust and can accurately describe the spatial movement of objects. The trajectory of moving objects can be obtained by applying tracking methods, such as a Kalman filter, particle filter, among others. Then, normal trajectories are utilized to build the normal event model. Finally, the abnormal event is identified by measuring the deviation or probability of the testing trajectory with respect to the normal event models. Li et al. [9] learned a dictionary using normal trajectories and then detected abnormal events according to the reconstruction error of their trajectory under the learned dictionary. Aköz et al. [10] proposed a traffic event classification system that learned normal and common traffic flows by clustering vehicle trajectories. Laxhammar et al. [11] proposed a method for online learning and sequential anomaly detection using trajectories. Bera et al. [12] proposed a real-time anomaly detection method in low- to medium-density crowd videos using trajectory-wise behavior learning.

However, trajectory-wise descriptors tend to fail in crowded and complex scenes since object detection and tracking are difficult to implement in crowded scenes. In addition, trajectory-wise descriptors mainly detect objects with unusual routes, while the body action of objects, such as jumping or falling down, cannot be detected. Coşar et al. [13] considered the pros and cons of trajectory-wise descriptors and proposed a unified AED framework by incorporating both trajectory- and pixel-wise analysis.

2) Pixel-wise descriptor

Pixel-wise descriptors can be directly extracted from scenes without requiring object detection and tracking and thus are frequently adopted for analyzing crowd events. HOG and histogram of optical flow (HOF) are two commonly used pixel-wise descriptors used in previous works. Wang et al. [3] applied HOF for AED. Zhao et al. [14] utilized both HOF and HOG to describe the motion and appearance inside a volume. Bertini et al. [15] employed a spatio-temporal HOG to describe both the motion and appearance of a volume. Zhang et al. [16] combined the features of both HOF and gradients for AED. Cong et al. [17] proposed a multiscale histogram of optical flow (MHOF) to describe the motion in a volume at different scales.

In addition to HOG and HOF, there are other types of pixel-wise descriptors that can be used to solve specific problems. Kaltsa et al. [18] proposed a histogram of oriented

swarms (HOS) descriptor based on swarm theory together with the HOG to capture both the motion and appearance in a volume. Colque et al. [19] developed a 3D descriptor for AED called histogram of optical flow orientation and magnitude and entropy (HOFME), which can effectively capture motion (velocity and orientation), appearance, and entropy information. Li et al. [20] modeled the pixels of a volume as a mixture dynamic textures (MDT) to jointly model both the motion and appearance within the volume. Wang et al. [21] proposed a spatio-temporal texture (STT) descriptor for real-time AED, which was constructed by transforming the XY, XT, and YT slices of a volume from a spatio-temporal domain into wavelet space. In [22, 23], spatio-temporal oriented energy (SOE) was exploited, in which a set of energy filters was used to capture a wide range of image dynamics and filter the irrelevant variation. Ribeiro et al. [24] proposed a Rotation-Invariant feature modeling MOtion Coherence (RIMOC) descriptor for violence detection in unstructured scenes, which was able to capture the structure and discriminate motion in a spatio-temporal volume. However, pixel-wise descriptors are unable to capture the contextual information in scenes, which is necessary for AED in some cases, e.g., irregular co-occurrence events.

3) Context-wise descriptor

A context-wise descriptor is another type of descriptor that is used to capture contextual information and plays a key role in the process of discovering salient events. Contextual information can be classified into a motion context and appearance context according to the descriptor generated based on the motion/appearance feature words. Yang et al. [25] proposed a semantic context descriptor both locally and globally to find rare classes in a scene. Yuan et al. [26] exploited contextual evidence using a structural context descriptor (SCD) to describe the relationship of individuals. Hu et al. [27] proposed a compact and efficient local nearest neighbors distance (LNND) descriptor to incorporate the spatial and temporal contextual information around a video event for AED. In fact, contextual information is an important cue for AED since it reflects the co-occurrence relationships or macro-structural information among semantic descriptors. Meanwhile, the context-wise descriptor is more efficient and flexible than the pixel-wise descriptor for AED since it is computed based on different types of regional features. However, the context-wise descriptor has not attracted as much attention as trajectory- and pixel-wise descriptors.

In our work, the proposed HOCG descriptor is a context-wise descriptor because it reflects the compositional relationship of sub-regions rather than micro-information of pixels within an event. Although previous works [4–7] also exploited the contextual information for event description, all of these works designed a learning framework to learn the contextual descriptor rather than designing an effective handcrafted contextual descriptor.

4) Deep-learned descriptor

In the last decade, much effort has been devoted to learning an effective descriptor via deep learning. Different types of deep neural networks have been designed to learn rich discriminative features, and a strong performance has been achieved in AED. Hasan et al. [28] proposed a convolutional autoencoder framework for reconstructing a scene, and the reconstruction costs were computed for identifying abnormalities in the scene. Sabokrou et al. [29] proposed a deep network cascade for AED. In the first stage, most normal patches were rejected by a small stack of an auto-encode, and a deep convolutional neural network (CNN) was applied to extract the discriminative features for the final decision. Hu et al. [30] proposed a deep incremental slow features analysis (D-IncSFA) network to learn the slow features in a scene. Feng et al. [31] proposed a deep Gaussian mixture model (D-GMM) network to model normal events. Zhou et al. [32] proposed a spatio-temporal CNN to learn the jointed features of both appearance and motion. Although a deep neural network can automatically learn useful descriptors, handcrafted features could still play a dominant role and be widely used in both image and video domains because they can benefit from human ingenuity and prior knowledge as well as enjoy flexibility and computational efficiency without relying on large sets of samples for training.

3 Approaches

In this section, we first introduce the principle of contextual gradient, then present the construction process of the HOCG descriptor for an event, and finally, present the details of AED using the HOCG descriptor under the online sparse reconstruction framework.

3.1 Contextual gradients

Different from traditional gradients that are computed pixel-wise, contextual gradients are computed regional-wise. We define the contextual gradients of the given region R_{ij} in the vertical and horizontal as

$$G_i(i,j) = \operatorname{sign}\left(R_{i+1,j}, R_{i-1,j}\right) \cdot \operatorname{dist}\left(R_{i+1,j}, R_{i-1,j}\right) \quad (1)$$

$$G_j(i,j) = \operatorname{sign}\left(R_{i,j+1}, R_{i,j-1}\right) \cdot \operatorname{dist}\left(R_{i,j+1}, R_{i,j-1}\right) \quad (2)$$

respectively. If the 3D contextual gradient is used, the temporal contextual gradient is also computed:

$$G_\tau(i,j,\tau) = \operatorname{sign}\left(R_{i,j,\tau+1}, R_{i,j,\tau-1}\right) \cdot \operatorname{dist}\left(R_{i,j,\tau+1}, R_{i,j,\tau-1}\right) \tag{3}$$

where dist(, ·,) is the distance measure between a pair of regions and sing(, ·,) returns the sign of the contextual gradient. Figure 1 shows visualizations of gradient maps horizontally, vertically, and temporally as well as a gradient magnitude map of scenes for pixel- and context-wise gradients.

Unlike the sign of a pixel gradient, which can be directly determined by a value comparison, we could not directly judge the sign of the gradient between two regions. To solve this problem, we utilized the saliency value of the region to determine the sign of the gradient. Specifically, we first computed the saliency value for each region in the scene and then determined the sign of the contextual gradients by comparing their saliency values given by.

$$\operatorname{sign}\left(R_i, R_j\right) = \begin{cases} 1 & if\ S_{R_i} > S_{R_j} \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

where S_{R_j} refers to the local saliency value of R_j. Saliency is one of the most popular concepts for computational visual attention modeling and can be quantitatively measured by the center-surround difference, information maximization, incremental coding length and site entropy rate, among others. For each region R_i, we used the context-aware method [33] to compute its saliency value. The saliency value of a given region as the center-surround difference measured by the distance between the features of the center and its K nearest neighbors in the surrounding regions is given by

$$S_{R_i} = \sum_{k=1}^{K} \operatorname{dist}_{fea}(R_i, R_k) \tag{5}$$

where f_i refers to the regional features extracted from region R_i and $\operatorname{dist}_{fea}(\cdot,\cdot)$ refers to the distance measure in

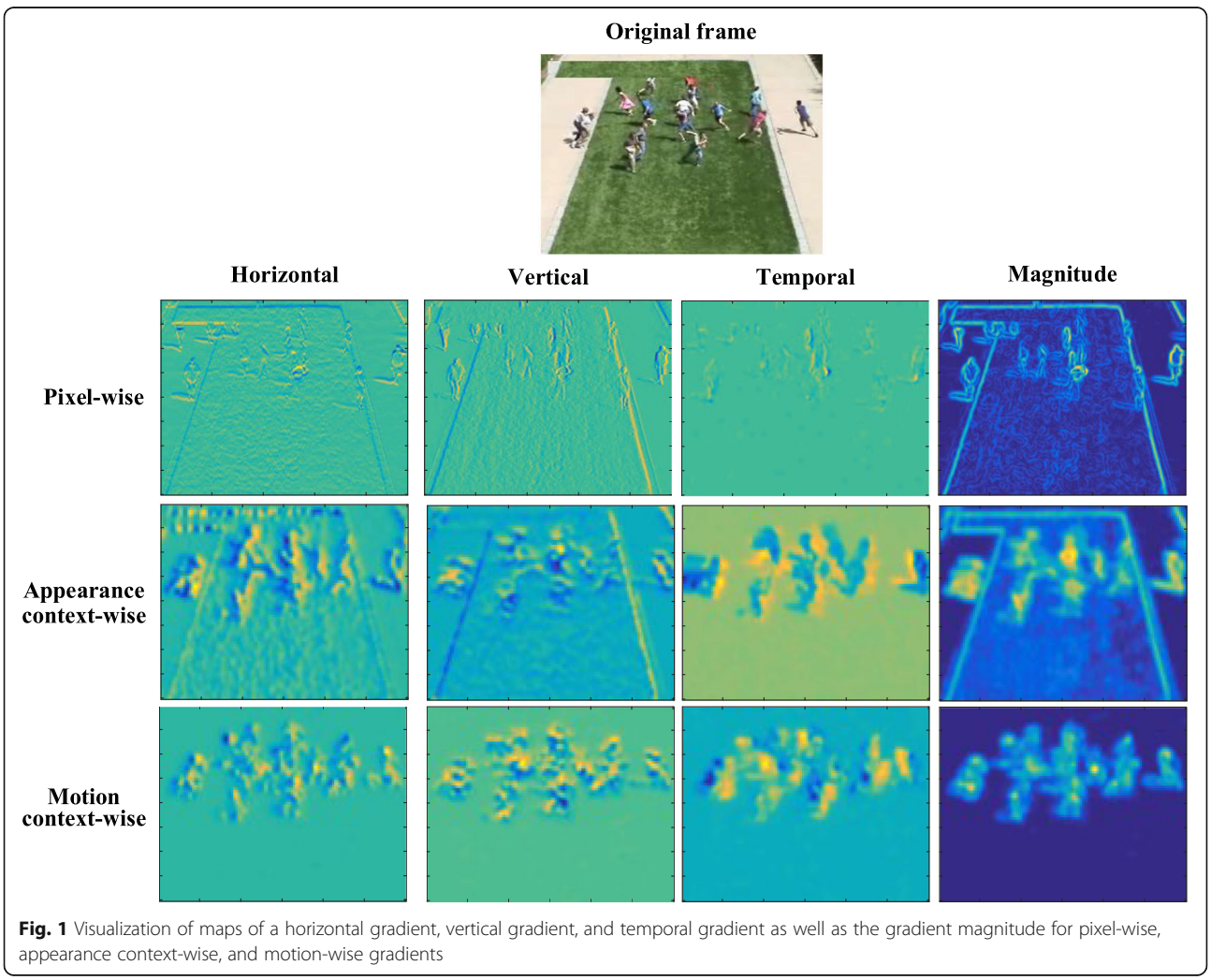

Fig. 1 Visualization of maps of a horizontal gradient, vertical gradient, and temporal gradient as well as the gradient magnitude for pixel-wise, appearance context-wise, and motion-wise gradients

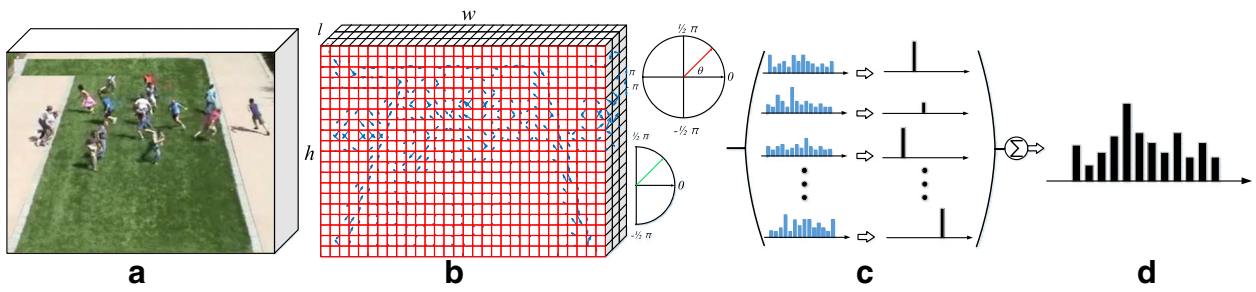

Fig. 2 Illustration of the processes for constructing a HOCG descriptor for an event. **a** A video clip (an event), (**b**) directions of contextual gradients of sub-regions, (**c**) quantizing descriptors of sub-regions, and (**d**) HOCG descriptor

the feature space. On the one hand, to reduce computational complexity, we only use the immediate eight surrounding neighbors of the center region. On the other hand, to reduce the influence of noise, we select the four nearest neighbors in the feature space from the eight neighbors. The contextual gradient can be computed based on different types of features, such as the gray values of raw pixels, HOG, HOF, and gradient central moments (GCM) [34]. We adopted the commonly used Euclidean distance as the distance measure between two features, defined as

$$\mathrm{dist}_{fea}(R_i, R_k) = \left\| f_i - f_k \right\|_2 \tag{6}$$

Other robust distance measurements, such as earth movers' distance (EMD) [35], can also be adopted to improve the robustness.

3.2 Histogram of oriented contextual gradient descriptor construction

In our work, a video event is a spatio-temporal volume and contextual gradients are computed for each small sub-region within the event. Based on the proposed contextual gradient, we construct a histogram for each event by quantizing each regional descriptor into a specific direction bin with respect to its contextual gradients. Given that a volume $V_{mnt}$with size of $w \times h \times l$ consists a set of non-overlapping sub-regions $\{R_{ijt}\}$, with each having a size of $p \times q \times r$, three contextual gradients (i.e., horizontal, vertical, and temporal contextual gradients) are computed for each sub-region, where w, h, and l are divisible by p, q, and r, respectively. Figure 2 illustrates the process of the HOCG descriptor construction. The contextual gradient magnitude ϕ_{ijt} is computed as

$$\phi_{ijt} = \sqrt{G_i^2 + G_j^2} \tag{7}$$

and the spatial directional anglesθ_{ijt} is computed as

$$\theta_{ijt} = \tan^{-1} G_i / G_j, \theta_{ijt} \in [-\pi, \pi] \tag{8}$$

If the 3D HOCG descriptor is used, the magnitude ψ_{ijt} should be computed using three spatial and temporal gradients that are given by

$$\phi_{ijt} = \sqrt{G_i^2 + G_j^2 + G_t^2} \tag{9}$$

and the temporal directional angles ϕ_{ijt} also should be computed

$$\phi_{ijt} = \tan^{-1} \frac{G_t}{\sqrt{G_i^2 + G_j^2}}, \phi_{ijt} \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right] \tag{10}$$

Using the computed magnitude and directional angle of all of the sub-regions in the volume, we construct a histogram with B_s bins, which means that 360° of the spatial direction range is quantized into B_s directions; the angle range of each direction is $360^{\circ}/B_s$. If the 3D HOCG descriptor is used, we need to further quantize 180° of the temporal direction into B_t directions with the angle range of each direction at $180^{\circ}/B_t$ and construct a histogram with $B_s B_t$ bins. Given a sub-region R_{ijt}, we quantize the region into the bth bin. For 2D HOCG,

$$b = \left\lceil B_s \theta_{ijt} / 2\pi \right\rceil. \tag{11}$$

For 3D HOCG,

$$b = B_s \left\lfloor B_t \phi_{ijt} / \pi \right\rfloor + \left\lceil B_s \theta_{ijt} / 2\pi \right\rceil \tag{12}$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the operations of rounding down and rounding up, respectively. Then, we assign the cuboid R_{ijt} with a B-dimensional vector $u_{ijt} = 0$, $u_{ijt} \in \mathbb{R}^B$, in which all elements are zeros except for $u(b) = 1$ weighted by its magnitude ϕ_{ijt}. Finally, we obtain the HOGR descriptor of the volume by accumulating all of the vectors of cuboids in the volume.

Algorithm 1 $\boldsymbol{HORG(V_{mnt})}$

V_{mnt} is a volume indicating an event occurred at the spatio-tempral location (m, n, t)
Initial $w_{mnt} = 0,\ w_{mnt} \in \mathbb{R}^B$
for each sub-region $R_{ij\tau}$
  Extract feature descriptor $f_{ij\tau}$;
  Compute saliency value $S_{R_{ij\tau}}$ using (5);
  Compute G_i, G_j and G_τ using (1), (2), and (3), respectively;
  Compute $\varphi_{ij\tau}$, $\theta_{ij\tau}$, and $\phi_{ij\tau}$ using (7), (8), and (10), respectively;
  Assign $u_{ij\tau}$ to $R_{ij\tau}$;
  $w_{mnt} = w_{mnt} + u_{ij\tau}$;
end for
return w_{mnt}

$$w_{mnt} = \sum\nolimits_{R_{ij\tau} \in V_{mnt}} u_{ij\tau} \tag{13}$$

Algorithm 1 shows the algorithms of the HOCG descriptor construction.

3.3 Abnormal event detection

AED can be classified as global AED (GAED) and local AED (LAED) [17]. GAED aims to detect an abnormal event caused by the group that occurs in the whole scene, such as a suddenly scattered crowd. Additionally, AED aims to detect an abnormal event caused by individuals and that occurs in a local region of the scene. For GAED, the video sequence is first divided into a set of temporal clips and each clip is considered as a global event. For local AED, each clip is further divided into a set of local volume and each volume is considered as a local event. Figure 3 illustrates the process of dividing the video sequence for global AED and local AED.

Due to the unpredictability of abnormal events, most previous approaches only learn normal event models in an unsupervised or semi-supervised manner, and abnormal events are considered to be patterns that significantly deviate from the created normal event models. In this work, we employ the

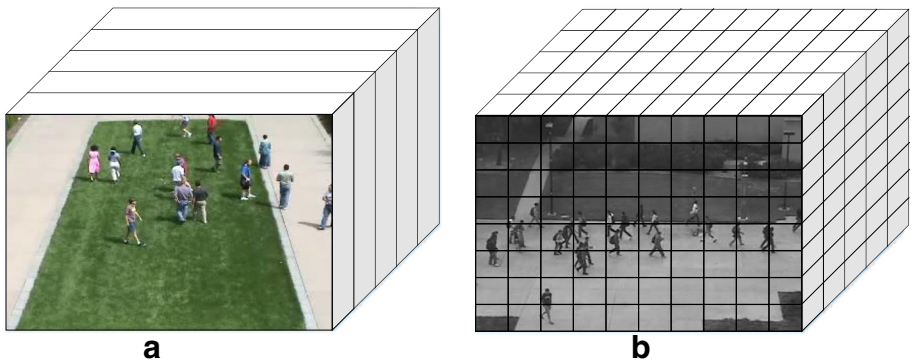

Fig. 3 Illustration of the division of a video sequence. The sequence is divided into (**a**) a set of temporal clips for global AED and (**b**) a set of spatio-temporal cuboids for local AED

online dictionary learning and sparse reconstruction framework for AED in which the abnormal event is identified as its sparse reconstruction cost (SRC) higher than a specific threshold.

Given an event with HOCG descriptors w_{mnt}, its SRC can be computed as

$$C_{mnt} = \frac{1}{2}\left\|D_{mn,t-1}\alpha_{mnt} - w_{mnt}\right\|_2^2 + \lambda\|w_{mnt}\|_1 \quad (14)$$

where $D_{mn,\ t-1} \in \mathbb{R}^{B\times S}$ is the online dictionary updated at $t-1$, $D_{mn,\ t-1}$is continuously updated to D_{mnt} at each time t using w_{mnt}, λ is the regularization parameter, and $\alpha_{mnt} \in \mathbb{R}^S$ is the sparse coefficient obtained by sparse coding under $D_{mn,\ t-1}$.

Dictionary learning is a representation learning method that aims at finding a sparse representation of the input data in the form of a linear combination of atoms in the dictionary. The dictionary can be learned in either an offline or online manner. Offline learning must process all training samples at one time, while online learning only draws one input or a small batch of inputs at any time t. Consequently, both the computational complexities and memory requirements of the online method are significantly lower than those of the offline method. Meanwhile, the online learning method has better adaptability than the offline method in practice. Thus, our work adopts the online dictionary learning method for AED, which is followed by two steps: sparse coding and dictionary updating.

3.4 Sparse coding

Given a fixed dictionary $D_{mn,\ t-1}$and a HOCG descriptor w_{mnt}, the sparse coefficient $\alpha_{mnt} \in \mathbb{R}^S$ can be obtained by optimizing

$$\alpha_{mnt} = \arg\min_{\alpha_{mnt}} \frac{1}{2}\left\|D_{mn,t-1}\alpha_{mnt} - w_{mnt}\right\|_2^2 + \lambda\|w_{mnt}\|_1 \quad (15)$$

This sparse approximation problem can be efficiently solved using orthogonal matching pursuit (OMP), which is a greedy forward selection algorithm.

3.5 Dictionary update

At each time t, the optimal dictionary can be obtained by optimization

$$D_{mnt} = \arg\min_{D_{mnt}} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}w_{mnt} - D_{mnt}\alpha_{mnt}{}_2^2\right) \quad (16)$$

For more details regarding online dictionary learning and sparse coding, we refer to [36]. Finally, we labeled the event as normal or abnormal based on a threshold δ

$$\text{Label}(V_{mnt}) = \begin{cases} \text{Normal} & C_{mnt} \le \delta \\ \text{Abnormal} & \text{otherwise} \end{cases} \quad (17)$$

where the threshold δ can be chosen experimentally when the approach achieves the best performance. The

Algorithm 2 Online AED Using HOCG descriptor

Input: HOCG descriptors, regularization parameter λ, threshold δ, initial dictionary D_{mn0}
Repeat
  for each spatial location (m, n)
  Compute sparse coefficient $\alpha_{mnt}$of w_{mnt} using (15), compute SRC value C_{mnt} of w_{mnt} using (14);
  if $C_{mnt} > \delta$ then
    Label event V_{mnt} as abnormal;
  end if
  Update $D_{mn,t-1}$ to D_{mnt}using the dictionary updating algorithm in [34];
  end for
until reach the end of video sequence

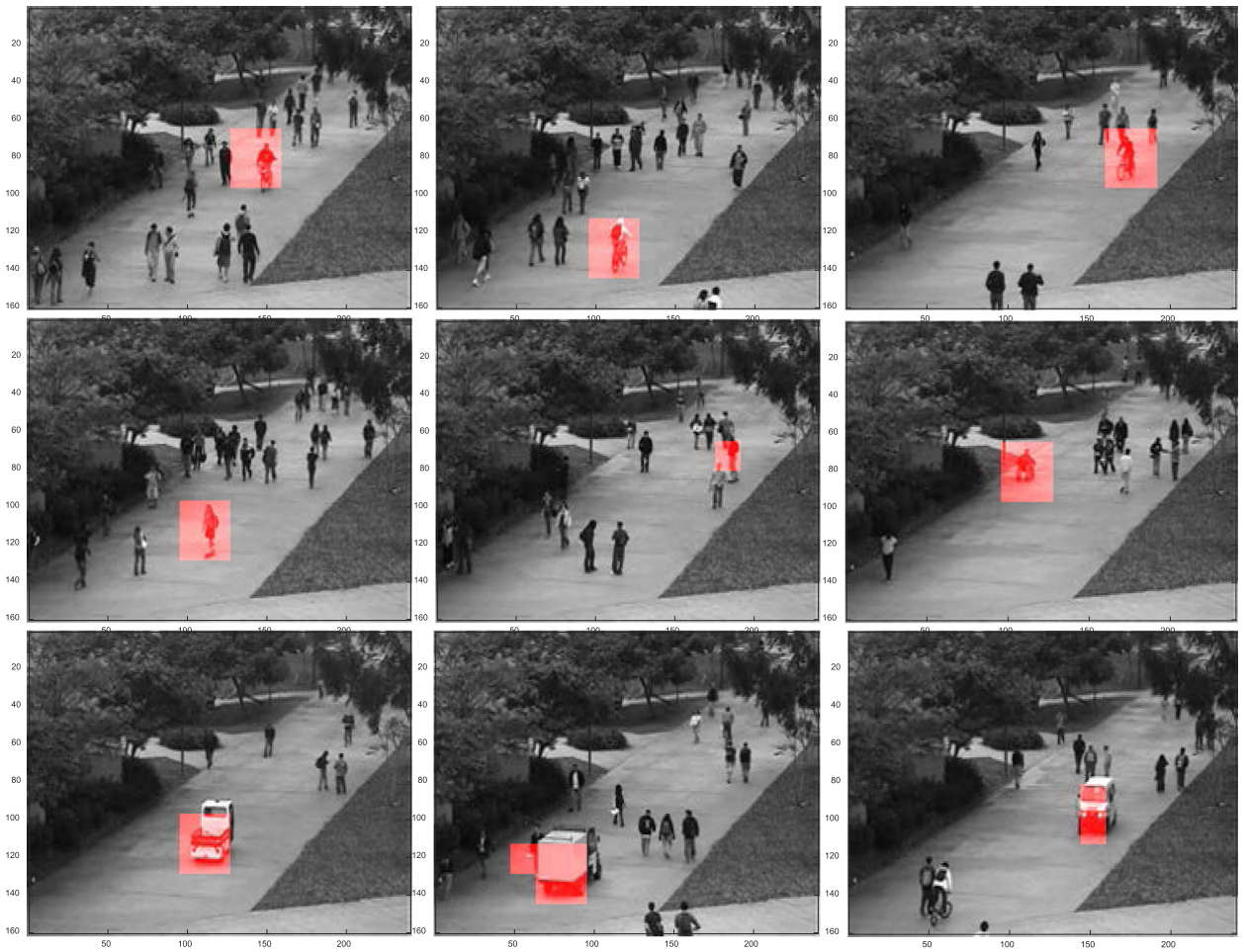

Fig. 4 Examples of abnormal event detection in the Ped1 dataset

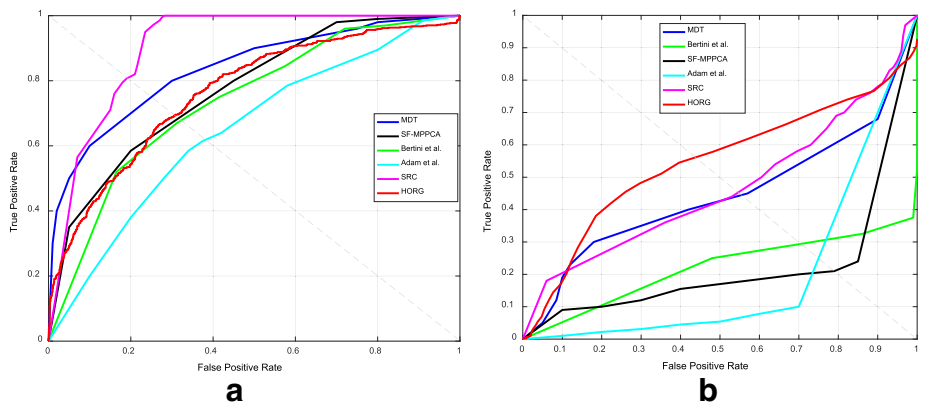

Fig. 5 Quantitative comparison of the detection results in the Ped1 dataset using ROC curves for the (**a**) frame-level criterion and (**b**) pixel-level criterion

Table 1 Summary of the quantitative performance and comparison with state-of-the-art descriptors in the Ped1 dataset

Approaches	EER (%)	DR (%)
SRC [41]	19	46
MDT [41]	25.4	45
SF-MPPCA [34]	31	21
Bertini et al. [15]	33	29
HOFME [19]	33.1	–
HOOF [19]	36.4	–
Adam et al. [42]	38	25
MBH [19]	43.4	–
HOG3D [19]	50	–
HOCG	30.6	57

Table 2 Performance comparisons between the HOCG descriptors and its based regional features

Approaches	AUC(%)	Improvement(%)	Dimension
Gray	63.20 71.94	8.74	64
HOCG			8
SF [42]	71.48 75.74	4.26	30
HOCG			8
GCM [34]	75.32 76.44	1.12	48
HOCG			8
3D gradient	62.62 74.04	11.42	64
HOCG			8

steps of AED with online dictionary learning and sparse coding are shown in Algorithm 2.

4 Results and discussion

We conduct experiments on different public datasets to evaluate the performances of AED approaches using the HOCG descriptor. The public datasets are UCSD [37], UMN [38], PETS2009 [39], and Avenue datasets [40]. All of the experiments are performed on a PC with a dual-core 2.5 GHz Intel CPU and 4 GB of RAM using MATLAB R2016a implementation. We use the UMN and PETS 2009 datasets for global AED, where abnormal events occur in most of the parts of scenes. We use the UCSD and Avenue datasets for local AED, where abnormal events occur in a relatively small local region.

4.1 UCSD dataset

4.1.1 Results

The UCSD dataset consists of the Ped1 and Ped2 subsets, which are taken from the UCSD campus by stationary monocular cameras. The density of the crowd varies from sparse to very crowded. The only normality in the scene is pedestrians walking on the walkway. The abnormalities include bikers, skaters, and vehicles crossing the walkway. We adopt the Ped1 subset for experiments since it provides complete ground truth for evaluating performance. The Ped1 dataset contains 34 training and 36 testing clips, in which each clip contains 200 frames with a resolution of 158 × 238 pixels. We resize the resolution to 160 × 240. The training set contains 34 clips of normal event, and the testing set contains 36 testing clips. The sequence is first divided into a set of volumes with a size of 16 × 16 × 5, in which each volume is considered as an event. Then, the volume is further divided into a set of cuboids with a size of 4 × 4 × 5. We extract the slow features proposed in our previous works [34] from each cuboid as the regional feature, which is robust and discriminative. We construct a 2D HOCG descriptor for each event, i.e., the spatial direction range is quantized into 8 directions with each direction being 45°. The dimensionality of the HOCG descriptor is 8. In contrast to the

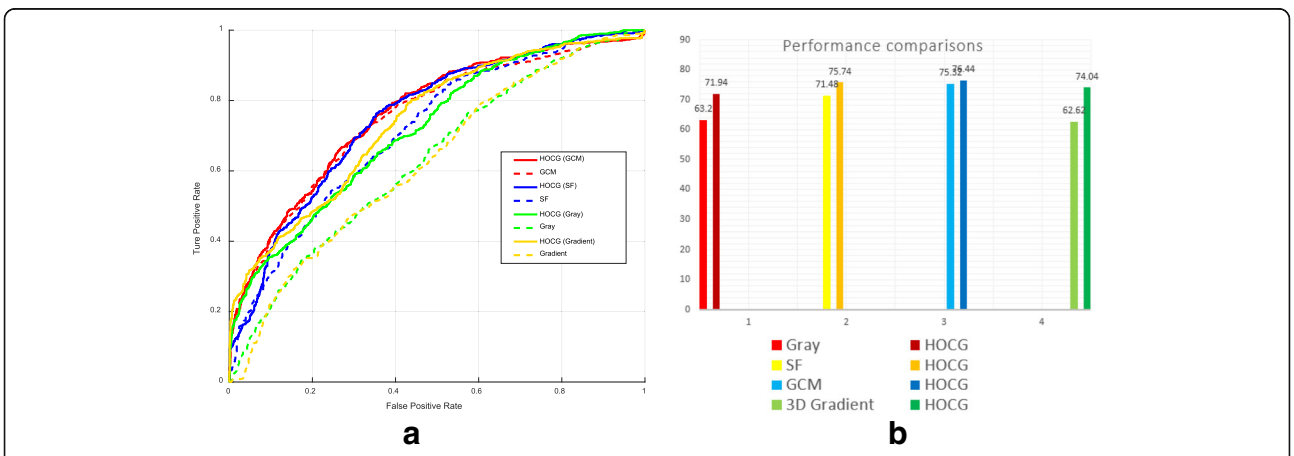

Fig. 6 Performance comparisons between the HOCG descriptor and its based features via (**a**) ROC curves and (**b**) bar diagrams

Table 3 Comparisons of the computational efficiency of HOCG descriptors with different parameters

Size of frame	Size of volume	Sizes of regions	Speeds (FPS)
160 × 240	16 × 16 × 5	2 × 2 × 5	30.6
		4 × 4 × 5	73.9
240 × 320		2 × 2 × 5	20.8
		4 × 4 × 5	44.9

20-dimensional regional features, the dimensionality is reduced significantly. The number of atoms for the dictionary is set to 20, and $\lambda = 0.5$.

Figure 4 shows examples of detection results; it is seen that different types of abnormalities, such as bicycles, skaters, and vehicles, can be detected and localized more accurately. To carry out a quantitative evaluation, we compare the performances of the HORG descriptor with several state-of-the-art descriptors, such as MDT [20], MOHF [17], and HOMFE [19] among others. Figure 5 shows the ROC (receiver operating characteristic) curves of the detection results from our approach as well as from other comparison approaches. The performances are evaluated by the equal error rate (EER) and detection rate (DR), which are reported for frame- and pixel-level evaluations, respectively. The lower the EER value, the better the performance that can be achieved, while the DR value is the opposite. The EER value is the ratio of misclassified frames at which FPR = TPR. The DR is at the pixel level, in which a frame is considered to be a detection if and only if 40% of truly abnormal pixels are identified; otherwise, it is considered to be a false positive. Compared to the frame-level criterion, the pixel-level criterion is more rigorous. Table 1 lists both the EER and DR values of our approaches and comparison approaches. Table 1 shows that our performances are better than the approaches of SF-MPPCA [41] by Bertini et al. [15] and Adam et al. [42] as well as HOFME [19], HOG3D [19], MBH [19], and HOOF [19]. Figure 6 shows a performance comparison between the HORG descriptor and the direct use of regional feature descriptors. Table 2 lists the AUC values of both the HORG descriptors and their based regional features.

4.1.2 Discussion

Although our performances are lower than those of the approaches of MDT [41] and SRC for the frame level evaluation, our performances outperform all of the comparison approaches for the pixel level evaluation, which is stricter than the frame level evaluation. As various types of regional features can be embedded in the HORG descriptor, we also demonstrate the performance improvement of using HORG descriptor as well as the reduction of dimensionality. We utilize four types of

Fig. 7 Examples of detection results for global abnormal event detection from the UNM dataset

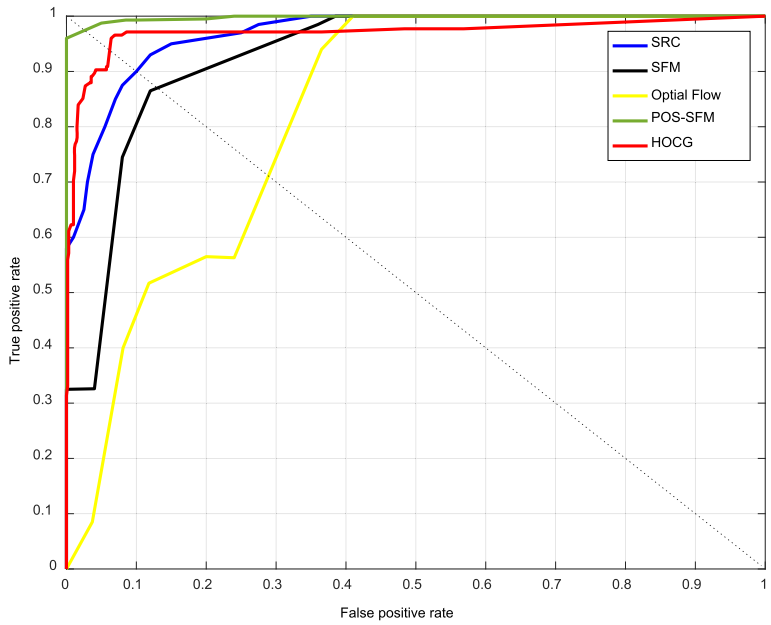

Fig. 8 Quantitative comparison of the detection results from the UMN dataset using ROC curves

regional features, i.e., the gray value, 3D gradient, GCM [34], and slow features (SF) [43] descriptors. The comparisons demonstrate that after constructing the HORG descriptors, not only the performance of AED is improved but also the dimensionality is also reduced.

To demonstrate the computational efficiency of the HORG descriptor, we recorded the speed (frames per second, FPS) of the construction of the HORG descriptor with different sizes of regions, different dimensionalities of regional features as well as different resolution of image. Table 3 lists the speed of the HORG descriptor with different parameters. It is interesting to note that the HORG descriptor can be constructed in real time.

4.2 UMN dataset

The UMN dataset contains three crowd escaping scenes in both indoor and outdoor environments. The normal events depict people wandering in groups, while the abnormal events depict a crowd escaping quickly. The dataset contains 11 sequences that are captured in three different scenes (lawn, indoor, and plaza) with a resolution of 240×320. The total frame number in the UMN dataset is 7740. The color frames are converted to gray scale; then, the size of each frame is resized to a resolution of 160×240. The video sequence is divided into a set of clips with a size of $160 \times 240 \times 5$, where each clip is considered as a global event. Each clip is further divided into a set of cuboids with sizes of $5 \times 5 \times 5$. We extract the GCM descriptor from the cuboids as regional features and then construct a HORG descriptor for each clip. The remaining parameters are the same as the setting in the experiment of the UCSD Ped1 dataset. Figure 7 shows some examples of the detection results of global abnormal events. Quantitative evaluation and a comparison with the state-of-the-art approaches [17, 43–46] are shown by ROC curves in Fig. 8 and the area under the curve (AUC) in Table 4. The performance of our approaches comparable to that of the state-of-the-art approaches is shown.

4.3 PETS2009

We continue to evaluate our approaches for global AED by conducting experiments on the sequences of S3\High Level\Time 14–33 from the PETS 2009 dataset. The four sequences depict an event with running, gathering, and dispersing of the same crowd from different views. The normal event is the crowd walking or merging at

Table 4 Summary of the quantitative performance and a comparison with state-of-the-art approaches on the UMN dataset

Approaches	AUC
Optical flow [43]	0.84
Social force [43]	0.96
MHOF [17]	0.978
Chaotic [44]	0.99
Interaction potential [45]	0.992
PSO-SF [46]	0.996
HOCG	0.993

Fig. 9 Examples of detection results for global abnormality detection from the PETS2009 dataset

normal speed, and the crowd running or dispersing suddenly is abnormal. The frame resolution is resized to 160×240, and the other parameters are the same as in the experiment on the UMN dataset. Some examples as well as ROC curves of the detection results for AED are shown in Fig. 9 and Fig. 10, respectively. These figures show that our approaches can well detect the global abnormal event of people quickly dispersing and achieve high performance. To evaluate the performances of our approaches, we compared our approaches with the approaches of PSO-SF [46], LBP-TOP [47], optical flow [47], and DBM [48] in Table 5. The comparison demonstrates that our performances were better than or comparable to those of the state-of-the-art approaches.

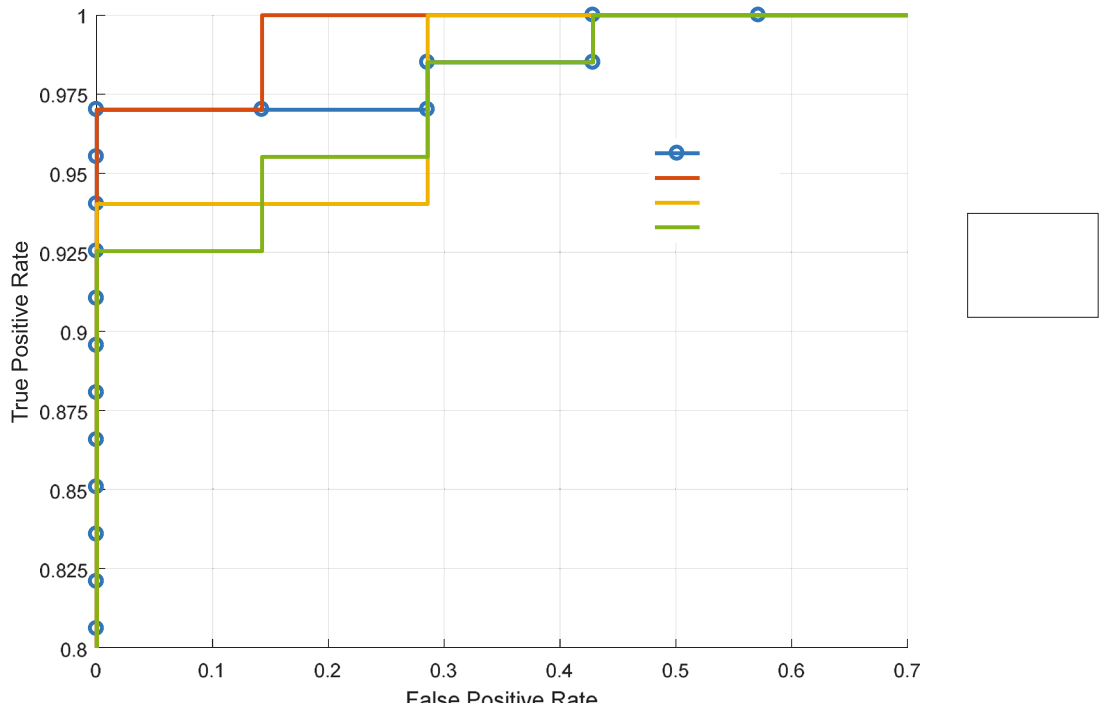

Fig. 10 Quantitative comparison of the detection results from the PETS 2009 dataset using ROC curves

Table 5 Summary of the quantitative performance and a comparison with the state-of-the-art approaches from the PETS2009 dataset

Approaches	View1	View2	View3	View4
PSO-SF [46]	94.14	–	–	–
TOP-LBP [47]	97.72	92.46	95.26	81.86
Optical flow [47]	98.01	–	–	–
DBM [48]	99.39	87.84	97.79	–
HOCG	98.93	99.57	98.72	98.08

4.4 Avenue dataset

The Avenue dataset was recently captured by researchers to evaluate the performance of the AED approach; this dataset contains 16 training videos with 15,328 frames and 21 testing videos with 15,297 frames. It is worth noting that the Avenue dataset is captured by a camera with a horizontal view, unlike the vertical view in UCSD, UMN, and PETS 2009 datasets. Therefore, we not only have to detect abnormal motion events but also must detect abnormal body actions, such as dancing, and throwing. The only normal event in the dataset is people working in front of the camera; the abnormal events are various and include unusual actions (running, throwing, dancing), waling in the wrong direction, and unusual objects. We resize the size of all frames to 160×240 and then divide the sequence into a set of volumes with a size of $16 \times 16 \times 16$, with 50% overlapping the spatial neighboring volumes. Each volume is further divided into a set of cuboids with a size of $2 \times 2 \times 16$. The GCM descriptor is extracted from each cuboid as the regional feature. The remaining parameters are identical to the setting in the experiments on the UCSD dataset. Figure 11 shows some examples of detection results; it is seen that different types of abnormalities, such as running, throwing, and loitering, can be accurately detected and localized. To quantitatively evaluate the performance of the HORG descriptor, Fig. 12 plots ROC curves of the detection results and Table 6 lists the AUC values of both the HORG descriptor and state-of-the-art comparison approaches [28, 31, 32, 49–52]. The comparisons demonstrate that the performance of the HORG descriptor outperforms the comparison approaches.

Fig. 11 Examples of detection results on the Avenue dataset

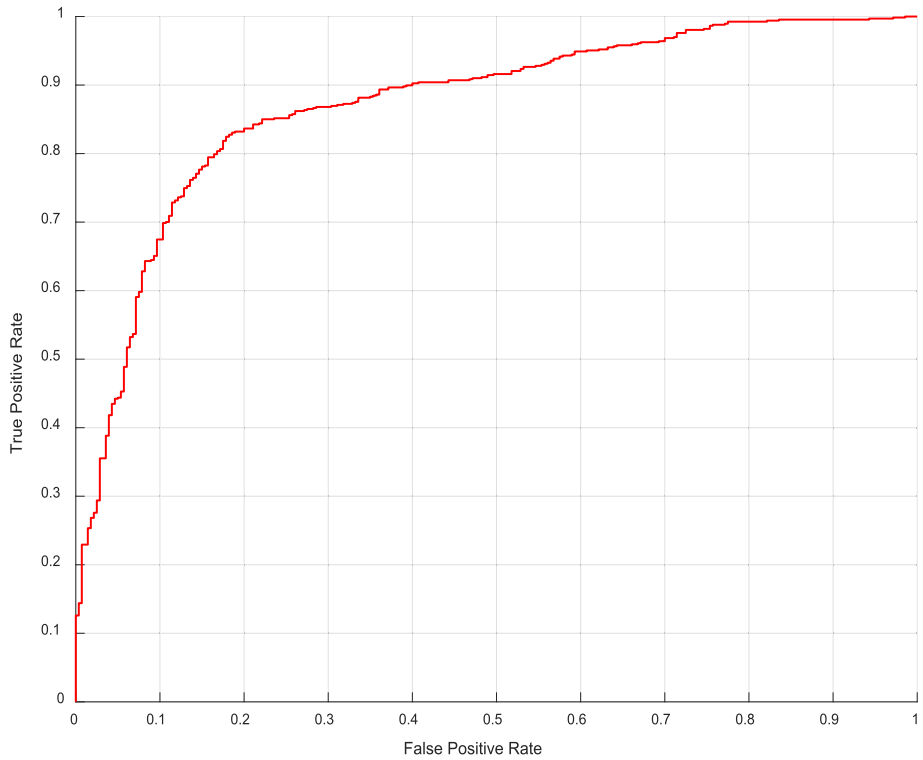

Fig. 12 ROC curves of the detection results on the Avenue dataset

5 Conclusions

In this paper, we extended the gradient from pixel- to context-wise and then constructed a HOCG descriptor using contextual gradients for AED. The HOCG descriptor is simple, compact, flexible, discriminative, and can efficiently capture the contextual information of an event. We conducted extensive experiments on different challenging public datasets to demonstrate the effectiveness of context-wise gradients. Quantitative and qualitative analyses of the experimental results showed that the HOCG descriptor outperformed the traditional pixel-wise HOG descriptor in AED and was comparable to the state-of-the-art approaches without using complicated modeling approaches. In future works, on the one hand, we will explore other applications of the HOCG descriptor, such as human action recognition and crowd activity recognition. On the other hand, we will investigate how well the compositional context of events under the deep learning framework is captured.

Table 6 Summary of the quantitative performance and comparison with the state-of-the-art approaches on the Avenue datasets

Approaches	AUC (%)
SCL [49]	70
Conv-AE [28]	70.2
3D Gradient	72.3
CLRS [50]	73.2
SHT [51]	74.2
Deep-GMM [31]	75.4
ST-AE [32]	80.3
GGMT [52]	87.7
HOCG	87.19

Acknowledgements
This work was jointly supported in part by the National Natural Science Foundation of China (Grant No. 61374197), Shanghai Natural Science Foundation (17ZR1443500), Shanghai Sailing Program, and Talent Program of Shanghai University of Engineering Science.

Funding
National Natural Science Foundation of China (Grant No. 61374197);
Shanghai natural science foundation (17ZR1443500);
Shanghai Sailing Program;
Talent Program of Shanghai University of Engineering Science.

Availability of data and materials
All data and material are available.

Authors' contributions
YH initiated the project. XH designed the algorithms. QD and JD conceived, designed, and performed the experiments. WC and HY analyzed the data. XH wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China. [2]School of Electronics and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China. [3]School of Electric Power Engineering, Nanjing Normal University Taizhou Colledge, Taizhou 225300, China.

Received: 16 January 2018 Accepted: 30 July 2018
Published online: 24 August 2018

References

1. M Paul, SME Haque, S Chakraborty, Human detection in surveillance videos and its applications-a review. EURASIP. J. Adv. Signal. Process **2013**(1), 176 (2013)
2. V Chandola, A Banerjee, V Kumar, Anomaly detection: a survey. ACM Comput. Surveys **41**(3), 15 (2009)
3. T Wang, H Snoussi, Detection of abnormal visual events via global optical flow orientation histogram. IEEE Trans. on Inf. Foren. Sec **9**(6), 988–998 (2014)
4. O Boiman, M Irani, Detecting irregularities in images and in video. Int. J. Comput. Vis. **74**(1), 17–31 (2007)
5. MJ Roshtkhari, MD Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Comput. Vis. Image Understand. **117**(10), 1436–1452 (2013)
6. N Li, X Wu, D Xu, H Guo, W Feng, Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing **155**, 309–319 (2015)
7. A Gupta, LS Davis, *Proc. IEEE Conf. on CVPR. Objects in action: an approach for combining action understanding and object perception* (2007), pp. 1–8
8. V Saligrama, J Konrad, PM Jodoin, Video anomaly identification. IEEE Signal Process. Magaz. **27**(5), 18–33 (2010)
9. C Li, Z Han, Q Ye, J Jiao, Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. Neurocomputing **119**, 94–100 (2013)
10. Ö Aköz, ME Karsligil, Traffic event classification at intersections based on the severity of abnormality. Mach. Vision Appl **25**(3), 613–632 (2014)
11. R Laxhammar, G Falkman, Online learning and sequential anomaly detection in trajectories. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1158–1173 (2014)
12. A Bera, S Kim, D Manocha, *Proc. IEEE-CVPRW. Realtime anomaly detection using trajectory-wise crowd behavior learning* (2016), pp. 50–57
13. S Coşar, G Donatiello, V Bogorny, C Garate, LO Alvares, F Brémond, Toward abnormal trajectory and event detection in video surveillance. IEEE Trans. Circuits Syst. Video Technol. **27**(3), 683–695 (2017)
14. B Zhao, L Fei-Fei, EP Xing, *Proc. IEEE-CVPR. Online Detection of Unusual Events in Videos via Dynamic Sparse Coding* (2011), pp. 3313–3320
15. M Bertini, A Del Bimbo, L Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization. Comput. Vis. Image Und. **116**(3), 320–329 (2012)
16. Y Zhang, H Lu, L Zhang, X Ruan, Combining motion and appearance cues for anomaly detection. Pattern Recogn. **51**, 443–452 (2016)
17. Y Cong, J Yuan, J Liu, Abnormal event detection in crowded scenes using sparse representation. Pattern Recogn. **46**(7), 1851–1864 (2013)
18. V Kaltsa, A Briassouli, I Kompatsiaris, LJ Hadjileontiadis, MG Strintzis, Swarm intelligence for detecting interesting events in crowded environments. IEEE Trans. Image Process. **24**(7), 2153–2166 (2015)
19. RVHM Colque, C Caetano, MTLD Andrade, WR Schwartz, Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Trans. Circ. Syst. Vid **27**(3), 683–695 (2017)
20. W Li, V Mahadevan, N Vasconcelos, Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell **36**(1), 18–32 (2014)
21. J Wang, Z Xu, Spatio-temporal texture modelling for real-time crowd anomaly detection. Comput. Vis. Image Und **144**, 177–187 (2016)
22. A Zaharescu, R Wildes, *Proc. ECCV. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing* (2010), pp. 563–576
23. E Ali, TH Mehrtash, D Farhad, CL Brian, Novelty detection in human tracking based on spatiotemporal oriented energies. Pattern Recogn **48**, 812–826 (2015)
24. PC Ribeiro, R Audigier, QC Pham, RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. Comput. Vis. Image Und **144**, 121–143 (2016)
25. J Yang, B Price, S Cohen, MH Yang, *Proc. IEEE-CVPR. Context Driven Scene Parsing with Attention to Rare Classes* (2014), pp. 3294–3301
26. Y Yuan, J Fang, Q Wang, Online anomaly detection in crowd scenes via structure analysis. IEEE Trans. Cybern **45**(3), 548–561 (2015)
27. X Hu, S Hu, X Zhang, H Zhang, L Zhang, Anomaly Detection Based on Local Nearest Neighbor Distance Descriptor in Crowded Scenes. Sci. World. J **2014**(6), 632575 (2014)
28. M Hasan, J Choi, J Neumann, AK Roy-Chowdhury, LS Davis, *Proc. IEEE-CVPR, Learning temporal regularity in video sequences* (2016), pp. 733–742
29. M Sabokrou, M Fayyaz, M Fathy, R Klette, Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans. Image Process **26**(4), 1992–2004 (2017)
30. X Hu, S Hu, Y Huang, H Zhang, H Wu, Video anomaly detection using deep incremental slow feature analysis network. IET Comput. Vis **10**(4), 258–265 (2016)
31. Y Feng, Y Yuan, X Lu, Learning deep event models for crowd anomaly detection. Neurocomputing **219**(5), 548–556 (2017)
32. S Zhou, W Shen, D Zeng, M Fang, Y Wei, Z Zhang, Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process-Image Communication **47**, 358–368 (2016)
33. S Goferman, L Zelnikmanor, A Tal, Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2012)
34. X Hu, S Hu, J Xie, S Zheng, Robust and efficient anomaly detection using heterogeneous representations. J. Electron. Imaging **24**(3), (033021)1–(03302112 (2015)
35. Y Rubner, C Tomasi, L Guibas, The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
36. C Lu, J Shi, J Jia, *Proc. IEEE-CVPR, Online Robust Dictionary Learning* (2013), pp. 415–422
37. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm, Accessed 18 Jul 2014
38. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi, Accessed 26 Oct 2012
39. http://www.cvg.reading.ac.uk/PETS2009/a.html, Accessed 6 Jun 2011
40. http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html, Accessed 10 May 2016
41. V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos,Proc. IEEE-CVPR, Anomaly Detection in Crowded Scenes (2010), pp. 1975-1981
42. A Adam, E Rivlin, I Shimshoni, D Reinitz, Robust realtime unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 555–560 (2008).
43. R Mehran, A Oyama, M Shah, *Proc. IEEE-CVPR, Abnormal crowd behavior detection using social force model* (2009), pp. 935–942
44. S Wu, BE Moore, M Shah, *Proc. IEEE-CVPR, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes* (2010), pp. 2054–2060
45. X Cui, Q Liu, M Gao, DN Metaxas, *Proc. IEEE-CVPR, Abnormal detection using interaction energy potentials* (2011), pp. 3161–3167
46. R Raghavendra, A Del Bue, M Cristani, V Murino, *Proc. IEEE-CVPRW, Optimizing interaction force for global anomaly detection in crowded scenes* (2011), pp. 136–143
47. J Xu, S Denman, C Fookes, S Sridharan, *Proc. IEEE-AVSS, Unusual Scene Detection Using Distributed Behaviour Model and Sparse Representation* (2012), pp. 48–53
48. J Xu, S Denman, S Sridharan, C Fookes, R Rana, *Proc. 2011 ACM-MM, Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes* (2011), pp. 25–30
49. C Lu, J Shi, J Jia, *Proc. IEEE-ICCV, Abnormal event detection at 150 FPS in MATLAB* (2013), pp. 2720–2727
50. Z Zhang, X Mei, B Xiao, Abnormal event detection via compact low-rank sparse learning. IEEE Intell. Syst, **31**(1), 29–36 (2016)
51. Y Yuan, Y Feng, X Lu, Statistical hypothesis detector for abnormal event detection in crowded scenes. IEEE Trans. Cybern **47**(11), 3597–3608 (2016)
52. S Li, Y Yang, C Liu, Anomaly detection based on two global grid motion templates, Signal Process-Image, (2017)