Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

Abnormal Event Detection based on Deep Autoencoder fusing optical flow

Meina Qiao[1], Tian Wang *[1], Jiakun Li[1], Ce Li[2], Zhiwei Lin[3], Hichem Snoussi[4]

1. School of Automation Science and Electrical Engineering, Beihang University,Beijing 100191, P. R. China
E-mail: meinaqiao@buaa.edu.cn, wangtian@buaa.edu.cn, lijiakun@buaa.edu.cn

2. College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, P. R. China
E-mail: xjtulice@gmail.com

3. School of Computing, Ulster University, BT37 0QB, United Kingdom
E-mail: z.lin@ulster.ac.uk

4. Institute Charles Delaunay-LM2S-UMR STMR 6279 CNRS, University of Technology of Troyes 10004, France
E-mail: hichem.snoussi@utt.fr

Abstract: As an important research topic in computer vision, abnormal detection has gained more and more attention. In order to detect abnormal events effectively, we propose a novel method using optical flow and deep autoencoder. In our model, optical flow of the original video sequence is calculated and visualized as optical flow image, which is then fed into a deep autoencoder. Then the deep autoencoder extract features from the training samples which are compressed to low dimension vectors. Finally, the normal and abnormal samples gather separately in the coordinate axis. In the evaluation, we show that our approach outperforms the existing methods in different scenes, in terms of accuracy.

Key Words: Abnormal event detection, Deep autoencoder, Optical flow

1 Introduction

Abnormal detection has attracted more and more attention in recent years as it is one of the key components in video surveillance applications. There have been various approaches to the abnormal detection. From the point of categories of abnormal events, a camera parameter independent and perspective distortion invariant approach was proposed to detect two types of abnormal crowd behavior: people gathering and running [1]. For different range of abnormal events, Zhang *et al* proposed an efficient approach to identify both local and long-range motion interactions for activity recognition [2]. Similar to abnormal detection, some of recent work focuses on event summarization and rare event detection together by transforming them into a graph editing framework [3], which is different from the conventional methods. In recent two years, novel approaches have been proposed for this detection. The work in [4] proposed to search for spatiotemporal paths, which correspond to event trajectories in the video space compared to spatiotemporal sliding windows. In [5], an anomaly detector using a joint representation of video appearance and dynamics and globally consistent inference, spanning time, space, and spatial scale, was proposed. By extending the conventional anomaly detection notions such as outlier or distribution drift alone, the study in [6] developed a unified framework for anomaly detection.

In order to improve abnormal event detection, this paper proposes to use deep learning autoencoder so that meaning features can be extracted. As a key part of deep learning, the autoencoder, and its variants, have been applied in many areas for video and image processing [7][8], including dimensionality reduction[11, 12], feature extraction[14, 16, 20], face parsing [15, 19], 3D human pose recognition [13, 17, 18], object detection[9, 10].

In object and abnormal detection areas, autoencoder has been applied to build generic object detectors to learn discriminative and compact features [9]. This semi-supervised model is a generative representation so that the input images can be reconstructed, and at the same time, it is discriminative so that the predictions of image labels are of high accuracy. Xu, *et al* proposed a moving object detection model from dynamic background based on two deep autoencoders, the background extraction network and the background learning network for different purposes [10]. Autoencoder is also used for dimensionality reduction [11] and the study extended the traditional autoencoder to explore the data relationship in order to discover the underlying effective manifold structure. Different from the traditional autoencoder, the proposed model reconstructed a set of instances instead of the input itself and minimized the reconstruction error of each instance for dimensionality reduction [12]. It has been shown that autoencoder approaches are different from other dimensionality reduction methods as the number of the hidden layer nodes is directly related to the dimensionality of the features for the best results.

The rest of the paper is organized as follows. We introduce the method of scene representation: optical flow and autoencoder in Section 2. Section 3 shows the abnormal detection model in detail. We conduct experiments with the datasets of lawn, indoor and plaza in Section 4. Finally, this paper is concluded in Section 5.

2 Scene Representation

Scene representations, also called features of images of a video [22], have great impact on abnormal event detection. For accurate detection, we combine the *optical flow* and *deep autoencoder* to learn features. First of all, the optical flow of

This work is partially supported by the ANR AutoFerm project and the Platform CAPSEC funded by Rgion Champagne-Ardenne and FEDER, the Fundamental Research Funds for the Central Universities (YWF-14-RSC-102), Aeronautical Science Foundation of China (2016ZC51022), and the National Natural Science Foundation of China (U1435220, 61503017, 61365003), Gansu Province Basic Research Innovation Group Project (1506RJIA031).

an original image is calculated and visualized as optical flow image. Then the optical flow image is used as input to a deep autoencoder for training and testing.

2.1 Optical flow image

Optical flow, proposed by Gibson in 1950, has been used to represent the motion information of objects between adjacent frames [21, 23, 24], and is of great importance in motion image analysis. Optical flow is the instantaneous velocity of pixel of moving objects projected to observing plane. The idea of optical flow for moving objects detection is to assign a velocity vector to each pixel in the whole image so as to forming the motion field of the image. On one hand, if no moving objects exist in the image, then the optical flow of the whole image area is continuous. On the other hand, when there is relative motion between the foreground and the background, the velocity vectors of moving object and the background are obviously different. As a result, the moving objects can be detected between adjacent frames.

There are two types of optical flow methods used widely: the first one is Lucas–Kanade (LK), a local optical flow method and the other one is Horn–Schunck (HS), a whole optical flow method. The latter HS method is employed in this paper. HS is developed based on two hypothesizes: brightness constancy constraint and the optical flow field of the whole image are smooth, which can be formulated as:

$$E_s = \min\{(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 + (\frac{\partial v}{\partial x})^2 + (\frac{\partial v}{\partial y})^2\} \quad (1)$$

where $u = dx/dt$, $v = dy/dt$ express the velocity vector along X axis and Y axis.

In addition, according to the optical flow fundamental constraint equation that the gray value of pixel before moving is the same as the one after moving, just as brightness constancy, we can get the formula as follows:

$$E_c = \iint (I_x u + I_y v + I_t)^2 dxdy \quad (2)$$

where I is the gray value of the pixel point at time t, I_x , I_y and I_t are the derivatives of the gray value at x, y and t.

So the solution to the optical flow field can be transformed into the solution of the following problem:

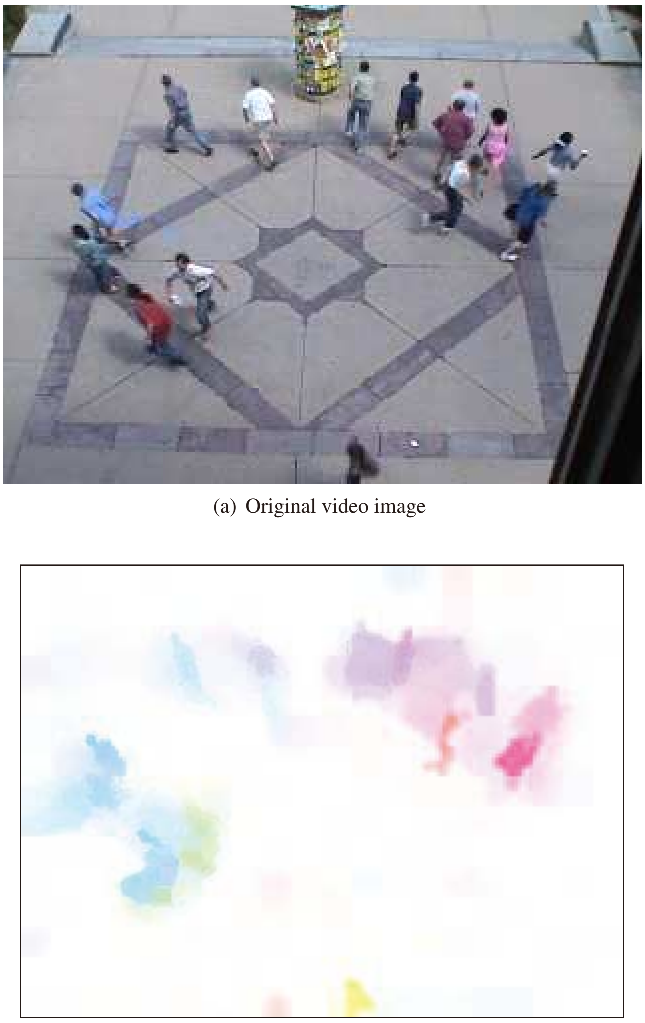
$$\min\iint\{(I_x u+I_y v+I_t)^2+[(\frac{\partial u}{\partial x})^2+(\frac{\partial u}{\partial y})^2+(\frac{\partial v}{\partial x})^2+(\frac{\partial v}{\partial y})^2]^2\}dxdy \quad (3)$$

According to the characteristics of the velocity vector of each pixel, the dynamic analysis of the image can be carried out. When there are moving objects in the image, there is relative motion between the object and the background. Figure 1 is an instance of optical flow image.

2.2 Autoencoder

Autoencoder is an unsupervised model whose output is to approximate the input. Figure 2 shows autoencoder.

In Figure 2, the left part of the structure is encoder, with input $X = \{x_1, x_2, \cdots x_n\}$ and output $Z = \{z_1, z_2, \cdots z_m\}$ ($m \ll n$). The encoder is designed to represent the input vector by a compressed vector whose dimension is much smaller than the original dimension, like PCA, in order to find the main components to represent the input. The second part is the decoder, whose input is $Z = \{z_1, z_2, \cdots z_m\}$ and whose output is $Y = \{y_1, y_2, \cdots y_n\}$. The objective is to minimize the loss of $||X - Y||_2$.



(a) Original video image

(b) Optical flow image

Fig. 1: Optical flow image. (a) The image in the original video frame. (b) The corresponding optical flow image

The whole deep autoencoder includes the input layer X, the output layer Y, and the hidden layers. The deep autoencoder is a neural network that is a repetition of the input with error as small as possible. That is to say, different layers are different representations of the input. In other words, each hidden layer is a kind of feature extracted from the input.

Since the connection between the layers is fully connected, the structure of the left part can be formulated as:

$$Z = f(wX + b) \quad (4)$$

where f is an activation function of the neuron, which can be sigmoid function, hyperbolic tangent, rectified linear unit (ReLU) and so on. In this paper, we use ReLU. This is just the formula of the encoder, the role of which is dimension reduction of the input.

The process of the right part is formulated as:

$$Y = f(w'Z + b') \quad (5)$$

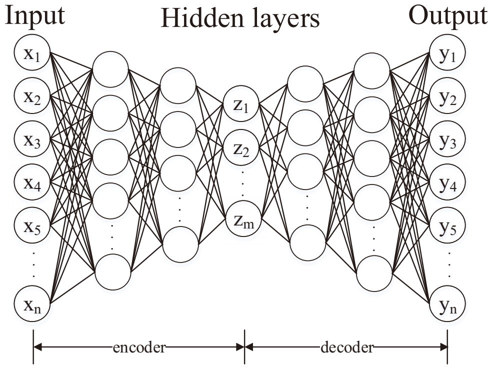


Fig. 2: Structure of deep autoencoder

The formula above is just the explanation of the decoder, the role of which is reconstruction from the compressed vector. So the whole structure of the deep autoencoder is formulated as:

$$Y = f(w'(f(wX + b)) + b') \tag{6}$$

3 Abnormal detection model

The abnormal detection model is divided into 2 parts: the training of deep autoencoder and the test for abnormal detection, which is outlined in Figure 3.

3.1 Training of deep autoencoder

In this section, a deep autoencoder is trained to reconstruct the input with minimal loss so that we can employ the trained encoder to extract features from the input.

Before training, optical flow features are extracted from the original video sequence images and visualized as optical flow image. Then the raw pixel of the optical flow image from the top left to bottom right are connected in series into vectors, which is used as the input of the deep autoencoder.

Since the size of the original image and the optical flow image is 240×320, the number of the input neurons is 76800. The dimension of the compressed vector is 3, namely the output of the encoder. Therefore, the structure of the deep autoencoder is, the encoder compressed the dimension of 76800 to 3, then the decoder reconstructs the 76800 neurons from 3 compressed neurons. In conclusion, the compressed 3 neurons contain all of the feature information of the 76800 neurons, the transformation of optical flow image, original video sequence image.

To gain better performance, we use ReLU as activation function, which is defined as: $f(x) = \max(0, x)$ It is more in line with the principle of neuron signal excitation than used widely sigmoid or hyperbolic tangent.

In order to reconstruct the input as accurate as possible, the weights ought to be modified by some algorithm, which is achieved by Error Back Propagation (BP) in this paper. BP algorithm combined with an optimization method such as gradient descent has been widely used in multi-layer neural network training. It consists of 2 parts: one is the forward propagation of the signal through the network, the other is the back propagation of the error. Since BP algorithm requires the determined output for the calculation of loss function, it is usually regarded as supervised algorithm. However, since the output of the deep autoencoder equals to its input, that is to say, the output is input. The BP algorithm can also be used in deep autoencoder.

Apart from the BP algorithm, appropriate reconstruction error such as cross entropy loss and mean square loss need to be determined and should be minimized. For facilitative calculation, mean square loss (MSE) is used in this paper, which is defined as:

$$L(X, Y) = \|X - Y\|^2 = \|X - f(w'(f(wX + b)) + b')\|^2$$

So the deep autoencoder is trained by BP algorithm until convergence, and the deep autoencoder can reconstruct the input with minimized error by then.

3.2 Test for abnormal detection

After the training of deep autoencoder, the encoder has gained a very comprehensive representation of the high-dimension input by low dimension. For the process of testing, encoder, the part of the trained deep autoencoder, is adopted to represent the input original video sequence. The treatment to the input is the same as training. What is different is that, the image in the test is the image for training and testing, not training, and only the encoder network is used.

Therefore, the testing images are transformed into optical flow images and connected in series into vectors as the input of encoder. The trained encoder compress the vectors into 3 vectors to be the representation of the image, respectively. Then the 3 vectors are used as the value of X, Y, Z coordinates of the point in space. In the end, we can draw the points in three-dimension coordinate axis of the corresponding images both normal and abnormal.

For abnormal detection, since we can only get the normal scenes in training, we employ the training image to find out the range of the coordinates of the points corresponding to the normal samples. We can use a cube to draw the range, which is shown in Figure 3. And the normal and abnormal points gather in different areas judging from the results of experiments. The trained normal (the green square) and tested normal samples (the blue circle) almost gather together while the tested abnormal samples (the red upward-pointing triangle) have no intersection with both trained normal samples and tested normal samples. So we can detect the abnormal samples from the normal ones. In other words, the features, that is each 3 dimension vectors extracted from the original video images, are comprehensive and representative.

4 Experimental results

We conduct experiments to evaluate the precision of the proposed algorithm. In our model, optical flow of the original video image sequence is calculated and visualized as optical flow image, which is then as the input of the deep autoencoder. Then the deep autoencoder is trained and the weights are optimized using Error Back Propagation (BP) algorithm. Finally, the image can be compressed to 3 vectors using encoder, a part of trained deep autoencoder. The 3 vectors can be drawn in 3-dimension coordinate axis. Since the normal and abnormal samples gather separately, we can distinguish between normal and abnormal ranges in 3-dimension coordinate axis, obviously.

To verify the accuracy of the model for abnormal detection, we make experiments on the datasets of indoor, plaza and lawn. The results prove that the algorithm is accurate enough for abnormal detection.

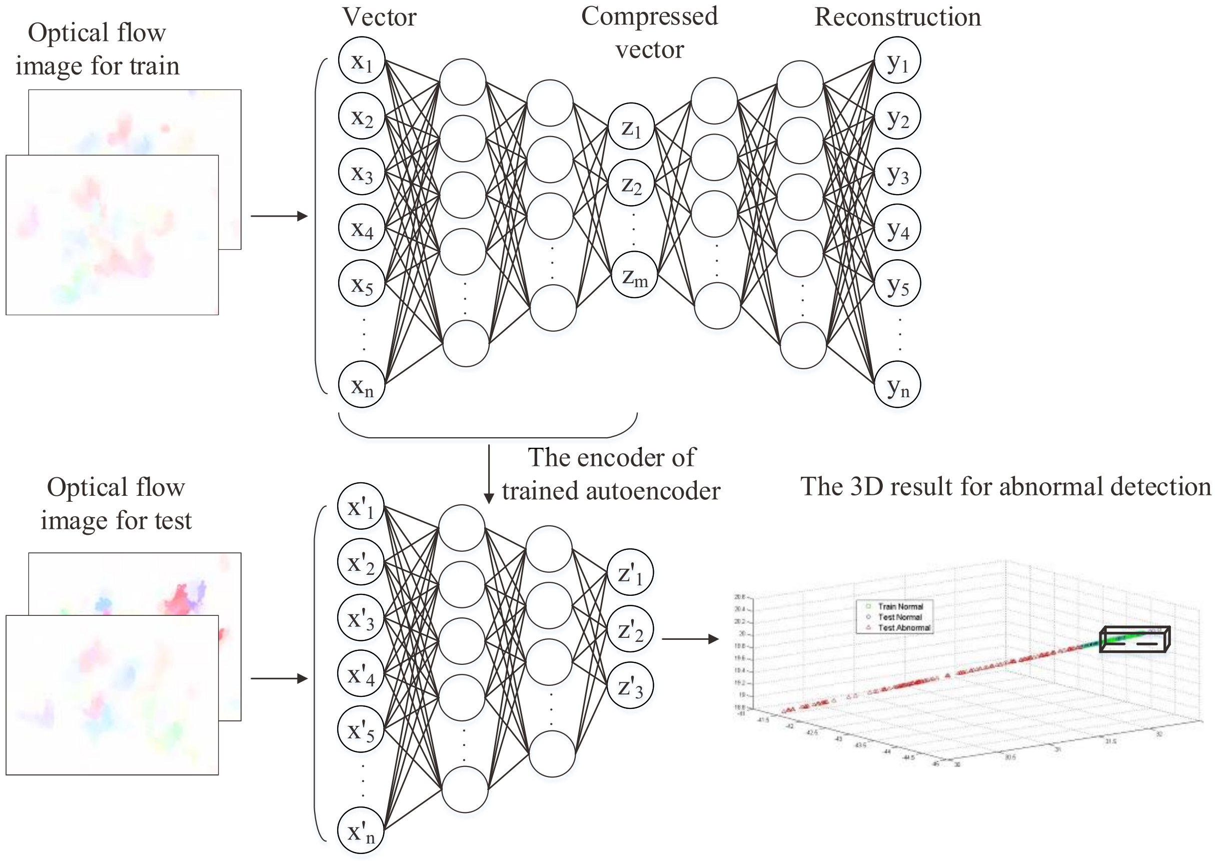


Fig. 3: The flowchart of abnormal detection.

Algorithm 1: Abnormal event detection based on deep autoencoder fusing optical flow

```
Data: Video image sequences $\{X_i\}_{i=1}^{N+1}$, where $X_i$ is a
      $240 \times 320$ image
Result: Abnormal sequences detection
Calculate the optical flow of the original video image $X_i$
  sequences then visualized as optical flow images $O_i$:
  $\{X_1, X_2, \dots, X_{N+1}\} \rightarrow \{O_1, O_2, \dots, O_N\}$.
Use $\{O_1, O_2, \dots, O_N\}$ to train the deep autoencoder by BP
  algorithm until convergence.
Calculate optical flow images $\{(O^{'})_j^t\}_{j=1}^m$ for the test images
  $\{Y_j\}_{j=1}^{m+1}$, and use $\{(O^{'})_j^t\}_{j=1}^m$ as input to the trained encoder
  which is extracted from the trained autoencoder in training
  process to obtain 3-dimension vectors $Z_j$, where $Z_j$ are the
  features of the test images.
Visualize the output of the encoder, 3 vectors in 3-D
  dimensional coordinates. Draw the range of the coordinates of
  the points of the trained normal images using a cube. And
  then visualize the test images the same way as trained images.
The abnormal test images are the points out of the cube.
```

4.1 Experiment on indoor scene

In the scene of indoor, the first 492 frames are normal, used for train and the following 275, 452, 105 are abnormal, normal, abnormal frames, respectively. One of the representation of normal and abnormal frame are shown in figure 4.In our experiment, we first use the 492 frames for training the autoencoder. Then the three part frames are as the input of the encoder of the trained autoencoder, which can be compressed to 3 vectors. Finally, the 3 vectors of every frame are calculated and visualized in three-dimension coordinate axis, which is shown in Figure 4. We use a cube to represent the range of 492 frame as the normal area. As a result, the area out of the cube is regarded as abnormal, the task for abnormal detection is achieved in this way.

4.2 Experiment on plaza scene

In the scene of plaza, the first 543 frames are normal, used for train and the following 114, 570, 85 are abnormal, normal, abnormal frames, respectively. One of the representation of normal and abnormal frame are shown in figure 5. In our experiment, we first use the 543 frames for training the autoencoder, updating the weights by BP until convergence. Then the experiment is made the same way as indoor scene. The result of the 3 vectors of every frame and the cube of normal range are visualized as Figure 5.

4.3 Experiment on lawn scene

In the scene of lawn, the first 480 frames are normal, used for train and the following 136, 673, 142 are abnormal, normal, abnormal frames, respectively. One of the representation of normal and abnormal frame are shown in figure 6. In our experiment, we first use the 480 frames for training the autoencoder, updating the weights by BP until convergence. Then the experiment is made the same way as above experiments. The result of the 3 vectors of every frame and the cube of normal range are visualized as Figure 6.

The ROC curves of these 3 datasets are shown in Figure

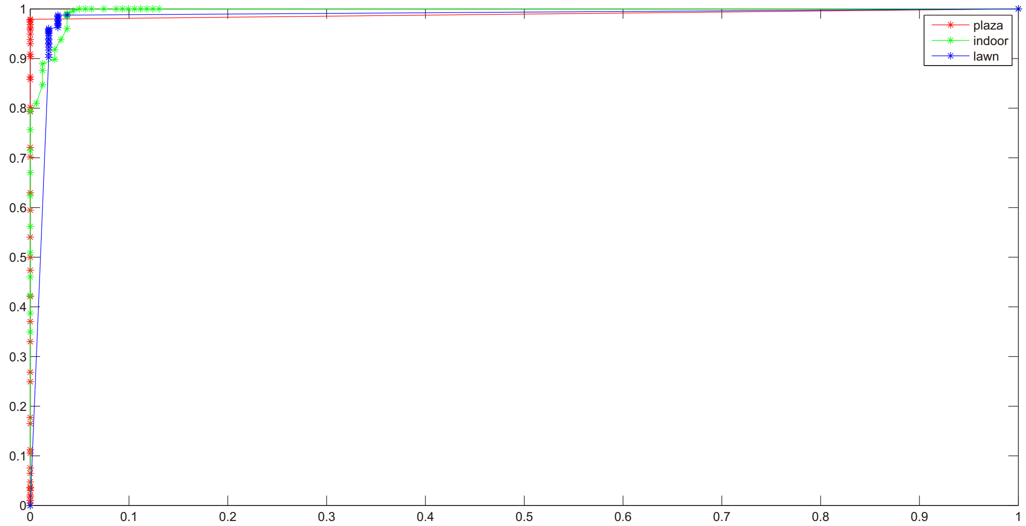
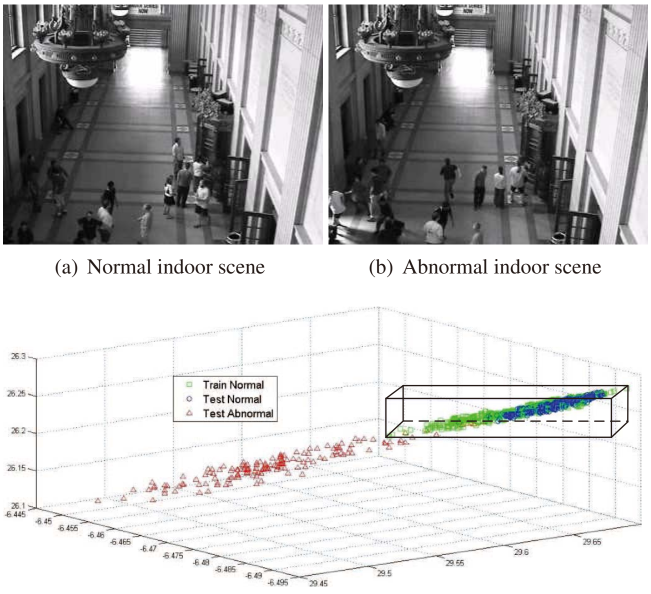


Fig. 7: ROC curves of the datasets.



(a) Normal indoor scene (b) Abnormal indoor scene

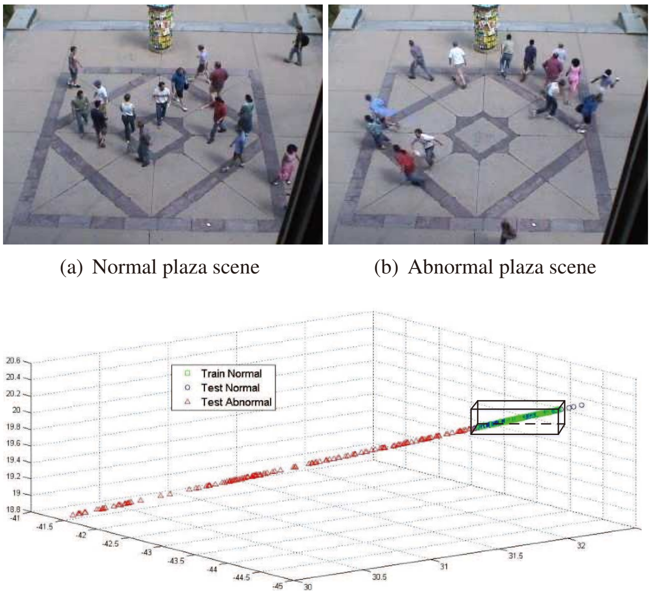
(c) Visualization of 3 vectors

Fig. 4: Example of scenes and the results of indoor scene dataset.

7, and the comparison between our algorithm and other algorithms are shown in Table 1. From the results of experiments on indoor, plaza, and lawn datasets, the model proposed can distinguish the abnormal scenes from the normal scenes with high accuracies.

5 Conclusions

In this paper, we proposed a novel algorithm for abnormal event detection, by combining optical flow and autoencoder. Optical flow of the original video image sequence is firstly calculated and visualized as optical flow image, which is then used as the input of a deep autoencoder, so that the features can be extracted with the deep autoencoder. Finally, the image can be compressed to 3 vectors using the trained encoder and the 3 vectors are drawn in 3-dimension coordinate axis, by which we can distinguish abnormal area from normal area. From the point of the result, our model is proved to be effective and accurate.



(a) Normal plaza scene (b) Abnormal plaza scene

(c) Visualization of 3 vectors

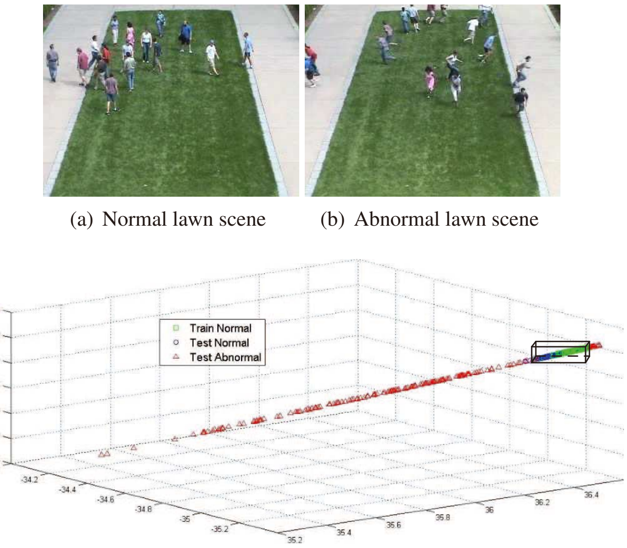
Fig. 5: Example of scenes and the results of plaza scene dataset.

References

[1] Xiong G, Cheng J, Wu X, et al. An energy model approach to people counting for abnormal crowd behavior detection, in *Neurocomputing*, 2012, 83: 121-135.

[2] Zhang Y, Liu X, Chang M C, et al. Spatio-temporal phrases for activity recognition, in *European Conference on Computer Vision(ECCV)*, 2012: 707-721.

[3] Kwon J, Lee K M, A unified framework for event summariza-



(a) Normal lawn scene (b) Abnormal lawn scene

(c) Visualization of 3 vectors

Fig. 6: Examples of scenes and the results of lawn scene dataset.

Table 1: Area under ROC curve

Methods	Area under ROC		
	Lawn	Indorr	Plaza
Social Force [25]	0.96		
Optical Flow [25]	0.84		
NN [26]	0.93		
SRC [26]	0.995	0.975	0.964
STCOG [27]	0.9362	0.7759	0.9661
Ours	0.9833	0.9956	0.9895

tion and rare event detection, in *IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*, 2012

[4] Tran D, Yuan J, Forsyth D. Video event detection: From subvolume localization to spatiotemporal path search, in *IEEE transactions on pattern analysis and machine intelligence*, 2014, 36(2): 404-416.

[5] Li W, Mahadevan V, Vasconcelos N, Anomaly detection and localization in crowded scenes, in *IEEE transactions on pattern analysis and machine intelligence*, 2014, 36(1): 18-32

[6] Kittler. Josef, Christmas. William, De Campos. Teofilo, Windridge. David, Yan. Fei, Illingworth. John and Osman. Magda, Domain anomaly detection in machine perception: A system architecture and taxonomy,in *IEEE transactions on pattern analysis and machine intelligence*,36, 845-859,2014

[7] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives, in *IEEE transactions on pattern analysis and machine intelligence*, 2013, 35(8): 1798-1828.

[8] Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: A review, in *Neurocomputing*, 2016, 187: 27-48.

[9] Yang Y, Shu G, Shah M. Semi-supervised learning of feature hierarchies for object detection in a video, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 1650-1657.

[10] Xu P, Ye M, Li X, et al. Dynamic background learning through deep auto-encoder networks, in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014: 107-116.

[11] Wang W, Huang Y, Wang Y, et al. Generalized autoencoder: A neural network framework for dimensionality reduction, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) Workshops*, 2014: 490-497.

[12] Wang Y, Yao H, Zhao S. Auto-encoder based dimensionality reduction, in *Neurocomputing*, 2016, 184: 232-242.

[13] Tekin B, Katircioglu I, Salzmann M, et al. Structured prediction of 3D human pose with deep neural networks, in *arXiv preprint*, arXiv:1605.05180, 2016.

[14] Lee D, Lee J, Kim K E. Multi-View Automatic Lip-Reading using Neural Network, in *ACCV 2016 Workshop on Multi-view Lip-reading Challenges*, 2016.

[15] Luo P, Wang X, Tang X, Hierarchical face parsing via deep learning, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, 2480-2487.

[16] Carneiro G, Nascimento J C. The use of on-line co-training to reduce the training set size in pattern recognition methods: Application to left ventricle segmentation in ultrasound, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012: 948-955.

[17] Park S, Hwang J, Kwak N. 3D human pose estimation using convolutional neural networks with 2D pose information, in *European Conference on Computer Vision(ECCV) 2016 Workshops*, 2016: 156-169.

[18] Zhou X, Sun X, Zhang W, et al. Deep kinematic pose regression, in *European Conference on Computer Vision(ECCV)*, 2016: 186-201.

[19] Zhang J, Shan S, Kan M, et al. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, in *European Conference on Computer Vision(ECCV)*, 2014: 1-16.

[20] Kan M, Shan S, Chen X. Bi-shifting auto-encoder for unsupervised domain adaptation, in *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 3846-3854.

[21] Wang T, Snoussi H. Detection of abnormal visual events via global optical flow orientation histogram, in *IEEE Transactions on Information Forensics and Security*, 2014, 9(6): 988-998.

[22] Qiao M, Wang T, Dong Y, et al. Real time Object Tracking based on Local Texture Feature with Correlation Filter, in *IEEE International Conference on Digital Signal Processing (DSP)*, 2016: 482-486.

[23] Zhao Y, Shi H, Chen X, et al. An overview of object detection and tracking, in *IEEE International Conference on Information and Automation*, 2015: 280-286.

[24] Portz T, Zhang L, Jiang H. Optical flow in the presence of spatially-varying motion blur, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012: 1752-1759.

[25] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model, in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009: 935-942.

[26] Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,2011: 3449-3456.

[27] Shi Y, Gao Y, Wang R. Real-time abnormal event detection in complicated scenes, in *IEEE International Conference on Pattern Recognition (ICPR)*, 2010: 3653-3656.