조명을 위한 인간 자세와 다중 모드 이미지 융합 - 인간의 이상 행동에 대한 강력한 탐지

Cuong H. Tran [1], 공성곤 [1]
[1] 세종대학교 컴퓨터공학과 석사과정

tdhcuong@sju.ac.kr, skong@sejong.edu

Multimodal Image Fusion with Human Pose for Illumination-Robust Detection of Human Abnormal Behaviors

Cuong H. Tran[1], Seong G. Kong[1]
[1]Dept. of Computer Engineering, Sejong University

요 약

This paper presents multimodal image fusion with human pose for detecting abnormal human behaviors in low illumination conditions. Detecting human behaviors in low illumination conditions is challenging due to its limited visibility of the objects of interest in the scene. Multimodal image fusion simultaneously combines visual information in the visible spectrum and thermal radiation information in the long-wave infrared spectrum. We propose an abnormal event detection scheme based on the multimodal fused image and the human poses using the keypoints to characterize the action of the human body. Our method assumes that human behaviors are well correlated to body keypoints such as shoulders, elbows, wrists, hips. In detail, we extracted the human keypoint coordinates from human targets in multimodal fused videos. The coordinate values are used as inputs to train a multilayer perceptron network to classify human behaviors as normal or abnormal. Our experiment demonstrates a significant result on multimodal imaging dataset. The proposed model can capture the complex distribution pattern for both normal and abnormal behaviors.

1. Introduction

Detecting abnormal human behavior in video surveillance has become a prominent area of interest within the research community. This heightened attention is primarily due to the increasing demand for intelligent systems capable of automatically identifying anomalous events in real-time video streams and surveillance cameras. This demand has been driven by the growing need for enhanced security in public spaces such as airports, train stations, supermarkets, schools, and busy urban streets, where surveillance cameras are employed to monitor human activities and identify deviations from normal behavior. Abnormal behavior detection in video surveillance primarily revolves around pinpointing unusual actions by analyzing both temporal and spatial information in visual recordings. In our study, we propose a robust abnormal human behavior detection framework centered on the analysis of human poses extracted from multimodal image sequences. Pose estimation provides crucial information in the form of keypoints, representing the precise locations of joints on the human body. These keypoints are indispensable for discerning various human activities, as they exhibit strong correlations with specific actions. Leveraging this pose information, we aim to identify abnormal behavior patterns effectively. Furthermore, we demonstrate the effectiveness of incorporating both visible and thermal imaging for abnormal behavior detection, particularly in low-light conditions, such as nighttime surveillance. Thermal cameras have the unique capability to capture invisible heat radiation emitted or reflected by all objects, irrespective of lighting conditions. By fusing data from both thermal and visible cameras, we can significantly enhance the overall performance of abnormal behavior recognition in challenging illumination environments.

2. Related Work

Conventional machine learning approaches used hand-crafted features to extract human appearance and detect spatial-temporal interest points for detecting abnormal behaviors. For example, [1, 2] used HOG and HOF for abnormal event detection. The handcraft-based methods are computationally expensive and not robust to the noise and cluttered background. The authors in [3, 4] built a sparse coding dictionary to record only normal events. The

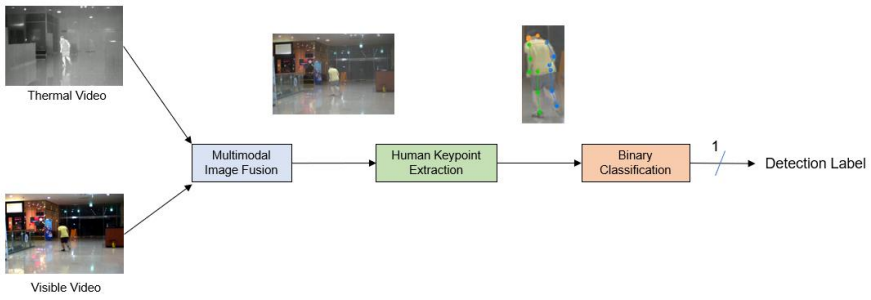


Figure 1. A schematic diagram of the proposed approach for both detection and recognition of human abnormal behaviors.

abnormal events will give a large reconstruction error during the inference. However, the sparse coding process is slow and time-consuming. Deep learning has been shown to be effective for learning representative human abnormal detection and recognition features. One approach for the detection task is to train a deep learning model to reconstruct or predict only normal behavior [5, 6]. Some methods [7, 8] used skeleton features or human joint keypoints to detect human abnormal events. The keypoint features provide the local information for each detected human targets since abnormal behavior detection is highly related to human skeleton and their motion patterns.

3. Methodology

Fig. 1 demonstrates our approach for detecting abnormal human behavior. Given the input thermal and visible videos, the first step is fusing both of them to produce the single fused video. The fused video is very robust in very low light conditions since it can provide an enhanced spectral range to visible eyes by capturing the high contrast between the environment and objects' temperatures. We employed the HRNet [9] to extract human keypoints from each human target on the scene. The keypoints features comprised of (x, y) coordinates for 17 keypoints in human joints following COCO format. Given the 17-keypoint coordinates and its corresponding labels, we can treat abnormal behavior detection problem as a supervised classification where 17-keypoint coordinates are our features to detect the target labels (0 or 1). The keypoint features and their corresponding labels are used to train a Multiplayer Perceptron Network (MLP) to classify a 17-keypoints skeleton as normal or abnormal one.

3.1 Multimodal Image Fusion

Fig. 2 illustrates the general process of multimodal image fusion for acquiring the fused video. Since the resolutions of the visible and thermal videos are different, the first step is to resize the spatial coordinates so that both videos share the same resolution. We upscaled the thermal video (640 × 512) to match the resolution of the visible video (1024 × 768) and used it as a reference image for the registration step. Image registration aligns both images, and it contains two sub-steps: offset remover and finetuning registration. The offset remover is actually the hardware correction to make two images close to each other. This step will choose a fixed transformation matrix based on the gap between the two cameras and also the distance between the observed objects and the cameras. This matrix will be used for initially warping the visible video to obtain the first result of registered RGB. We applied the ECC image alignment algorithm [10] for the finetuning process on RGB images. The algorithm uses the similarity measurement called Enhanced Correlation Coefficient to estimate the parameters of the geometric transformation matrix between the visible and thermal image in order to maximize the ECC score. ECC algorithm is robust to lighting contrast and distortion, which is suitable for different input domains such as thermal and RGB images. The final step is applying the averaging technique to fuse both thermal and visible images. We also stabilize the fused video by smoothing the optical flow between each frame.

3.2 Human Keypoint Extraction

We employed HRNet [9] as our pose detector. HRNet receives the bounding box location of each human target, then processes it through its high-to-low architecture. Its architecture focused on maintaining the high-resolution representation through multiple stages. Its original implementation had four stages, starting from the high-resolution convolution stream, and gradually adding high-to-low resolution convolution streams at the next stage. HRNet outputs the coordinates of 17-keypoints having the format of (x, y) where x-coordinate is related to width of the image and y-coordinate is related to the height of the image. Since the bounding box can have different resolutions, (x, y) can have a wide range of values. This makes the training process harder, and the output prediction will be highly influenced by the large values. So, we normalized all the values of into the range of (0,1). All the 17-keypoint coordinate values will be converted to a 34-dimension vector. This vector is treated as one skeleton instance to train the Multiplayer Perceptron Network along with the corresponding labels of abnormal

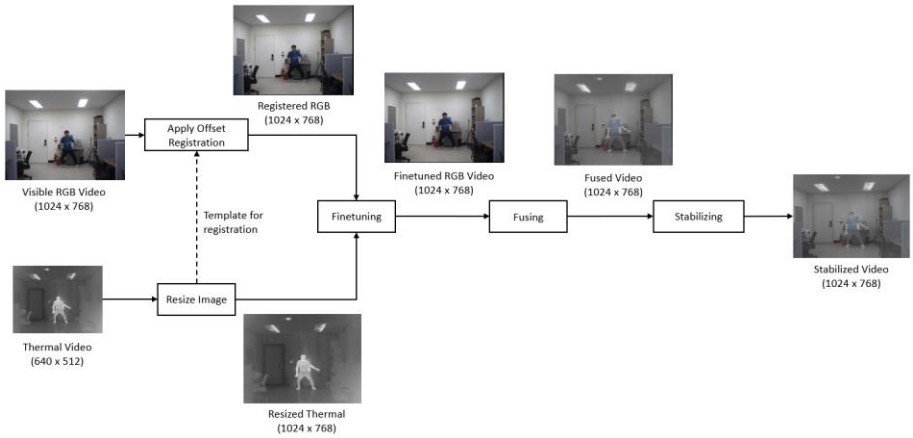


Figure 2. Multimodal image fusion process for acquiring the multimodal fused video.

class as 1 and normal class as 0.

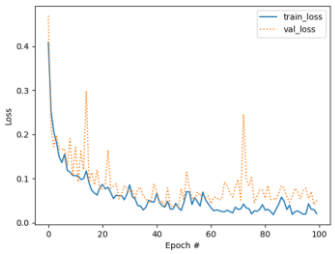
3.3 Multilayer Perceptron Training

We have devised a straightforward Multi-Layer Perceptron (MLP) network tailored for binary classification. The architecture of this MLP model comprises four linear layers, sequentially configured as 34 → 128 → 64 → 16 → 2. To enhance the stability of the training process, batch normalization layers have been strategically incorporated between each of these linear layers. Additionally, in the output layer, we've introduced a dropout layer to mitigate the risk of overfitting. Our model takes as input a dataset consisting of 17 keypoints' data. Each keypoint is associated with two values (x and y) representing its coordinates. Consequently, a single human skeleton's data is represented as a 17x2 = 34-dimensional feature vector, serving as a singular input. These 34-dimensional vectors are directed into the MLP's input layer, which is comprised of 34 units. Subsequently, the input layer is connected to the second layer, which houses 128 units. The second layer, in turn, connects to the third layer, which contains 64 units. The third layer is subsequently linked to the second-to-last layer, accommodating 16 units. Lastly, the output layer is configured with 2 units, each corresponding to one of the two classes: 0 denoting the normal class and 1 representing the abnormal class.

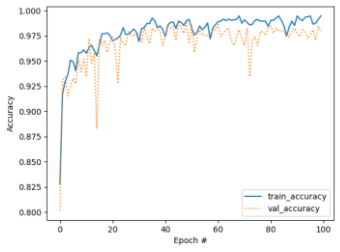
4. Experiment Results

We contrasted a multimodal image dataset, named MIHAB, an abbreviation for "Multimodal Imaging for Human known as MIHAB, which stands for "Multimodal Imaging for Human Abnormal Behavior." MIHAB encompasses images captured using both thermal and visible cameras, offering a diverse range of visual data sources. Our dataset has been meticulously designed to focus on five distinct categories of abnormal human behavior, which include: fighting (Fi), running (Rn), riding a bike or scooter (Bi), carrying a suspicious object (Oc), and leaving a suspicious object (Ou), as illustrated in Fig. 3.

We initially extracted the 17 key points data to train the MLP for binary classification. The training set has a total of 2320 skeleton instances, while the evaluation dataset has 580 skeleton instance. After 100 epochs of training, we can achieve the accuracy of 99% on the training set and 97% on the validation set. Fig. 4 shows the learning curves for the MLP training on training set and validation set. Looking at the loss diagram, the train loss is reducing steadily. It means that our MLP is able to learn how to classify the data. The train loss starts with a value of over 0.4 at the first epoch then it's reducing continuously to just 0.02 at epoch 100. The validation loss fluctuates during the training but overall its value decreases to 0.08 at epoch 100. At the end, we can see that there is just a small amount of overfitting since the gap between the train loss curve and validation is small.



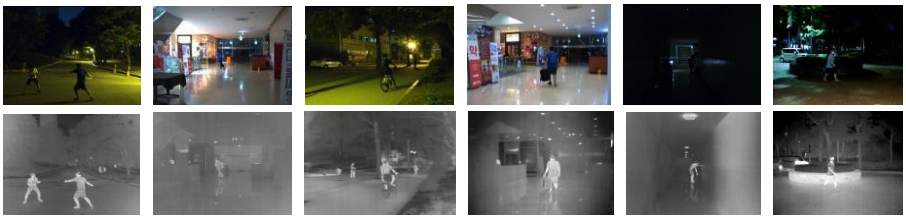
(a)



(b)

Figure 4. Learning curve of MLP model training on keypoint dataset. (a) Loss values, (b) Accuracy values.

Fig. 5 shows the testing results in the normal, running and fighting videos. The red bounding box indicates the abnormal behaviors along with the tracker ID and type of abnormal events. The green bounding box indicates normal behavior,



(a) *(b)* *(c)* *(d)* *(e)* *(f)*

Figure 3. Samples images in the MIHAB dataset. Top row: visible samples, Bottom row: thermal samples. (a) Fighting (Fi), (b) Running (Rn), (c) Biker (Bk), (d) Suspicious Object Carrier (Oc), (e) Bag Left Unattended (Ou), (f) Normal walking pedestrian (No)

while the white bounding box is unidentified behavior due to low resolution and occlusion.

Figure 5. Experiment results on testing videos.

Table 1 shows the evaluation metrics of our approach for human abnormal behavior detection in MIHAD dataset. The test set includes 8 videos from different scenarios such as fighting, running, bag left unattended. Our results showcase an impressive accuracy of 0.82 and a robust F1-Score of 0.83, underscoring the high effectiveness and reliability of our methodology.

Table 1. The evaluation metrics for human abnormal behavior detection in MIHAB Dataset

	Multimodal image fusion with human pose keypoints approach
Accuracy	0.82
Precision	0.89
Recall	0.78
F1-Score	0.83

5. Conclusion

This paper presents an innovative approach for detecting human abnormal behaviors in low-light environments. We achieve this by harnessing the power of multimodal image fusion and human pose keypoints. Our methodology involves the precise localization of each human subject, followed by the application of HRNet to estimate their pose keypoints. Subsequently, we employ an MLP trained in a supervised manner to discern between abnormal and normal instances. Notably, our MLP model exhibits an impressive accuracy rate of 98% during the evaluation phase. Furthermore, we demonstrate the advantages of incorporating pose estimation in our approach, as evidenced by our results on test videos. In our forthcoming research endeavors, we intend to explore the integration of both appearance features and pose features. This combined approach aims to further enhance the accuracy of abnormal behavior detection on the MIHAB database, thereby advancing the state of the art in this domain.

ACKNOWLEDGMENT

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT) under Grant 2019-0-00231, and in part by the Development of artificial Intelligence-Based Video Security Technology and Systems for Public Infrastructure Safety.

References

[1] N. Navneet and B. Triggs, “Histograms of oriented gradients for human detection,” Computer Vision and Pattern Recognition Conference, vol. 1, pp. 886-893, 2005.

[2] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” European Conference on Computer Vision, pp. 428-441, 2006.

[3] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," Conference on Computer Vision and Pattern Recognition, pp. 3313-3320, 2011.

[4] Y. Cong, J. Yuan and J. Liu, “Sparse reconstruction cost for abnormal event detection,” Conference on Computer Vision and Pattern Recognition, pp. 3449-3456, 2011.

[5] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 733-742, 2016.

[6] W. Liu, W. Luo, D. Lian and S. Gao, “Future frame prediction for anomaly detection–a new baseline,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545, 2018.

[7] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, “Learning regularity in skeleton trajectories for anomaly detection in videos”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11996–12004, 2019.

[8] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, “Multi-timescale trajectory prediction for abnormal human activity detection”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2626–2634, 2020.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5693-5703, 2019.

[10] G. D. Evangelidis and E. Z. Psarakis, "Parametric Image Alignment using Enhanced Correlation Coefficient Maximization", IEEE Transaction on Pattern Analysis & Machine Intelligence, Vol. 30, No. 10, pp. 1858-1865, 2008.