Spatio-Temporal Information for Action Recognition in Thermal Video Using Deep Learning Model

Original Scientific Paper

P. Srihari
School of Computer Science and Engineering,
VIT-AP University, Amaravathi 522237, India
psrihari851@gmail.com

J. Harikiran
School of Computer Science and Engineering,
VIT-AP University, Amaravathi 522237, India
harikiranj@vit.in

Abstract *– Researchers can evaluate numerous information to ensure automated monitoring due to the widespread use of surveillance cameras in smart cities. For the monitoring of violence or abnormal behaviors in smart cities, schools, hospitals, residences, and other observational domains, an enhanced safety and security system is required to prevent any injuries that might result in ecological, economic and social losses. Automatic detection for prompt actions is vital and may help the respective departments effectively. Based on thermal imaging, several researchers have concentrated on object detection, tracking, and action identification. Few studies have simultaneously extracted spatial-temporal information from a thermal image and utilized it to recognize human actions. This research provides a novelty based on frame-level and spatial and temporal features which combines richer context temporal information to address the issue of poor efficiency and less accuracy in detecting abnormal/violent behavior in thermal monitoring devices. The model can locate (bounded box) video frame areas involving different human activities and recognize (classify) the actions. The dataset on human behavior includes videos captured with infrared cameras in both indoor and outdoor environments. The experimental results using the publicly available benchmark datasets reveal the proposed model's efficiency. Our model achieves 98.5% and 94.85% accuracy on IITR Infrared Action Recognition (IITR-IAR) and Thermal Simulated Fall (TSF) datasets, respectively. In addition, the proposed method may be evaluated in more realistic conditions, such as zooming in and out etc.*

Keywords*: Action Recognition, Activity Classification, Complex-Valued Deep Fully Convolutional Network, Deep Learning, Deeplabv3+Net, Fall Detection, Mask R-CNN, Thermal Cameras, Violence.*

1. INTRODUCTION

Human action recognition (HAR) is a method for identifying a continuous action in a video clip that has grasped a lot of attention in recent years because of its various applications, which include gaming

and sports, healthcare, security, autonomous vehicles, automated assisted living systems, human-machine interaction, video surveillance cybernetics analysis. Substantial improvement has been achieved in action detection over the last several decades, and the majority of existing techniques for action classification have been used in visible image clips. [1–3]. In addition, several visible light action datasets, such as UCF101, KTH and HMDB51 have been developed for action detection. HAR in visible light has been widely studied and effectively used in various domains. Still, occlusion, background clutter, shadow and illumination variation continue to be serious obstacles for visible light AR [4-7].

Thermal infrared cameras, rather than visible light cameras, can capture human activities due to the advancement of vision-based technologies. Infrared (IR) action recognition, as contrasted to visible light AR, can address the aforementioned issues [8,9]. IR thermal imagery can capture individuals accurately in low light while they are difficult to see that in visible light recordings. This is particularly useful for nighttime monitoring or human-computer interaction (HCI) under dim illumination. Since the occlusion, background clutter and temperatures of the shadow are very low compared to those of individuals or moving objects in IR movies, these obstacles may be effectively reduced [10,11]. With these

characteristics, infrared AR is useful in a wider range of applications outperforming visible light.

Several current approaches essentially adhere to the process of raw feature extraction, feature encoding, and classifier training [12,13]. In this process, the feature map is widely regarded as the most crucial component, and its capability to represent the fundamental information beneath the video typically influences the efficiency of AR techniques [14,15]. Many well-established feature descriptors, such as STIP, HOG3D, and 3DSIFT, have been developed and utilized in action recognition (AR). Several approaches based on convolutional neural networks (CNN) have been developed for action identification in visible videos, motivated by the effective use of CNNs in other classification of image tasks. The existing two-stream convolutional architecture can effectively integrate motion and appearance data for AR [16, 17]. Yet, the complexity of CNN models and the high computational cost of the back propagation training process reduce the deep network's efficiency in tackling video AR challenges [18]. Furthermore, the existing CNN architectures models are built to analyze 2D raw frames; thus, they cannot be used directly to learn temporal information from input videos.

Recently, the computer vision industry has paid great importance to the timely detection of falls in home video surveillance and violent activities [19,20]. To analyze behavior performance, introduced a dataset of thermal images and videos that replicate violent motions and abnormal activity in outdoor and indoor environments. Recording conditions vary from simple to complicated backgrounds, occlusions and random camera angles. But, researchers faced enormous challenges in accurate and precise identification of these activities in thermal environments due to key factors like diversity in lightening conditions, background complexity and occlusions, HAR is complicated by similar acts performed by several individuals.

To the aforementioned issues, this paper focus on vision-based approaches for automatically detecting violence and abnormal activities in thermal cameras. The framework is designed based on frame-level and extraction of both temporal and spatial information. Additionally, to perform experiments, IITR-IAR and Thermal Simulated Fall datasets are considered as two benchmark datasets to compare the proposed method with existing methods. The simulation results show that the proposed method is helpful for the operators who supervise in several monitoring services since the safety and security of our daily life is so serious in society. The proposed framework's modules are visualized to demonstrate their efficiency. The contributions of the research paper are as follows:

- A unique deep learning technique is proposed for multimodal activity detection in thermal video.
- The proposed method can localize regions and recognize the multiple individual activities.
- The proposed framework can recognize more than 30 behaviors based on spatiotemporal information.
- Contrast limited adaptive histogram equalization (CLAHE) is implemented to enhance the quality of low-resolution images.
- To propose Gaussian – Adaptive Bilateral Filter (GABF) for removing the noise from the thermal image and handling the occlusion problem.
- For obtaining ROI proposed a deep learning method Mask-RCNN with the backbone as Densenet-41.
- For extracting the spatial and temporal information DeepLabv3+Net is adopted.
- Finally, the activity of the human is classified by the complex-valued fully convolutional network. For performance evaluation, implemented on IITR-IAR and TSF datasets.

This paper is organized as follows: Section 2 presents a comprehensive literature review. Section 3 presents the challenges and overviews of the proposed architecture and its main components. Section 4 reports the experiments conducted and analyzes the results. In Section 5, the conclusion and future work are presented.

2. LITERATURE SURVEY

During the last decade, various approaches for detecting violence or abnormal in different applications have been developed. This section examines efforts in the area of abnormality identification and fall detection that are conceptually similar to the present work research.

Manssor et al. [21] developed a deep person detector that recognizes nighttime pedestrians from TIR pictures. To accomplish the aim, Tiny-convolutional yolov3's layers are contrast-enforced at the channel level. A network design employing PDM-Net and TIE-Net with an Up-Sampling layer was demonstrated. The TIE-Net optimizes and processes TIR images by eliminating information loss between initial convolution layers. Darknet-53 and PDL-Net make up PDM-Net. The TIR image's features are retrieved by Darknet-53 and classified by PDL-Net.

Imran et al. [22] developed a four-stream network based on BiLSTM and CNN models to integrate global and local motion data. SSDIs may improve action identification accuracy coupled with SDFDI dense optical flow-based pictures. CNN uses ResNet, whereas RNN uses BiLSTM. Finally, two CNN streams are trained using a single SSDI and a single SDFDI to collect global Spatiotemporal knowledge. To collect local temporal and spatial information, the clip is divided into eight segments and eight SSDIs and SDFDIs are created. These SSDIs and SDFDIs train two CNN-BiLSTM streams. Late fusion combines all four streams to analyze the class label.

Krito et al. [23] employed CNN models for RGB pictures to recognize people in thermal photos. They examined the performance of state-of-the-art object detectors and retrained them using a dataset of thermal images collected from movies simulating unauthorized movements along the border and in protected regions. Nighttime videos are captured in fog, rain and clear weather at varied ranges, and with various movement patterns. YOLOv3 was quicker than other detectors and had equivalent performance, therefore it was employed in investigations.

Batchuluun et al. [24] suggested a technique to extract joints and skeletal information by transforming a 1-channel thermal picture into a 3-channel thermal image and utilizing it as input for Joint-GAN. Using the collected joints and skeletal information, CNN-LSTM was used to recognize human activities. To test the suggested approach, 3-channel and 1-channel thermal pictures were compared.

Ding et al. [25] designed a system for identifying airport-ground activities based on an infrared video stream. First, they create a seven-class action dataset based on airport-goer behavior. Considering the limitations of infrared surveillance photos, they used a tracking technique to separate the target from the background and constructed short-term Spatio-temporal feature vectors. Finally, they built an action classification network with two LSTM layers to extract long-term spatial and temporal data and a fully connected classifier.

Hei et al. [26] introduced an infrared AR framework called REWS to reweight training set samples to overcome limited IR action data. They first translate IR action video data to low-dimensional feature space and then calculate the score of the training data set samples based on the similarity measures between the training and testing set feature data. Each sample of a training set has its weight. The weighted training sets are then used to train a support vector machine (SVM) to detect infrared actions.

De Boissiere et al. [27] suggested that a pre-trained 3D CNN is employed as an IR module for extracting visual data from recordings. Using a multilayer perceptron, both feature vectors are then combined and used cooperatively. RoI from IR videos are cropped using 2D skeletal coordinates. This work integrates infrared and skeletal data. They assessed their system using the biggest dataset for HAR from depth cameras, the NTU RGB+D dataset. They conduct substantial ablation research.

3. METHODOLOGY

Recognizing human activities in thermal cameras is a crucial task. Figure 1 depicts the architecture of the proposed method. In this part, some of the research obstacles are discussed initially. Then, a DL-based approach is presented for the task of localization and action recognition.

3.1. CHALLENGES

Our framework highlights the issue of human behavior recognition, which involves the investigation of a few distinct actions. As a result, the approach will handle the following major issues:

- Intra-class and Inter-class Variations. Changes in internal or external stimuli cause people to react in different ways. For a particular movement, such as "walking," a person might exhibit multiple stances, speeds, and even occluded body parts while doing the same activity. In other words, a single motion may include numerous distinct movement patterns. Also, high levels of occlusion posed by other people or scenario objects make it difficult to view the whole movement area. There are parallels between activities such as "walking" and "running." These actions with tiny changes provide difficulty for deep learning systems as well.

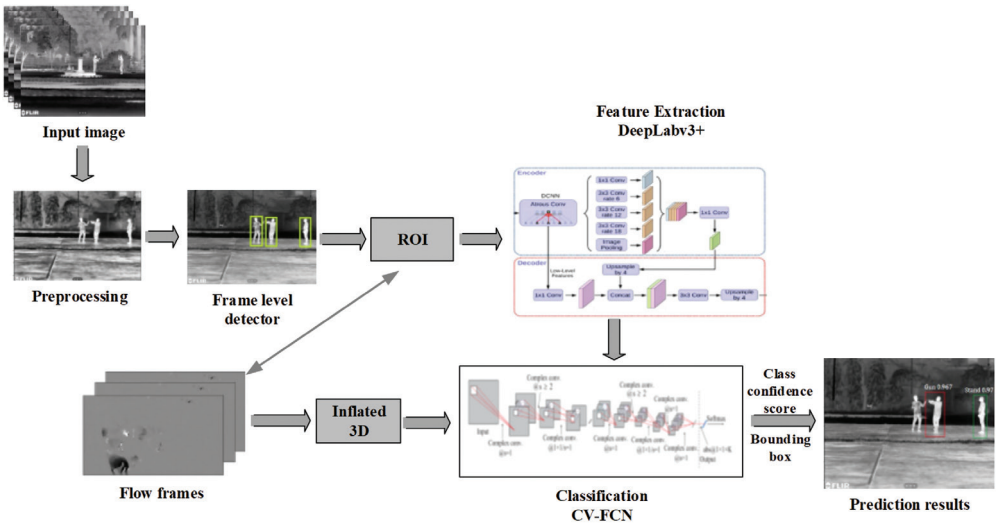


Fig. 1. Proposed framework

- Background and Time-Period. The proposed method seeks to identify various behaviors that occur at various times and the whole day. Therefore, the recordings were done at varying moments, including daytime and nighttime views. Even if the background in these videos may be the same, the intensity of light varies and the activities are distinct. Despite the difficulty of evaluating these conditions, the data collected from these tests would allow for more accurate monitoring of human activity.
- Early fall detection. A reliable fall detection system is required to limit the impacts of falling. The fall detection issue is considered an abnormal detection problem since falling is an irregular behavior. Even though many existing training models are focused on detecting falls, they are nevertheless susceptible to false positives. To eliminate false positives and achieve high accuracy, developing a reliable fall detection system is the major issue.

3.2. DATA ACQUISITION

To collect the IITR-IAR dataset, a FLIR T1020 camera is used. Its focal length is 8.4 mm and it shoots video at 1024 × 768 in a formate MPEG-4 at 30 fps. This camera has a spectral range of 7.5–14 μm. The temperature range is between −15°C and 50°C, and the field of vision (FOV) is 12°.

The second dataset (Thermal Simulated Fall) is comprised of videos acquired from FLIR ONE thermal camera placed on an Android mobile in a single-view room environment. The frame rates of the videos are either 15 or 25 fps, which was determined by analyzing the attributes of every video.

3.3. DATA ANNOTATION

Based on the following criteria, created the annotation approach to identify 38 distinct kinds of actions: Normal, Violent and Abnormal activities.

Normal activities: This activity includes clap, crouch, hop, run, walk, wave1, wave2, drop, recording video, selfie, throw, handshake, hug, passing object, sit, eat, drink, making the bed, Transfers from wheelchair to chair, Removing and putting on shoes, changing clothes, getting into and out of bed, sleep, read, cough, sneeze, bend.

Violent activities: This activity includes pointing a gun, chase, fight, kick, punch, push

Abnormal activities: This activity includes slow falls, fast falls such as falling from chain, falling while walking, falling when changing from chair to bed etc.

3.4. PREPROCESSING

The thermal video taken under different conditions is given as the input to the proposed method. The Gaussian–Adaptive Bilateral Filter is used to reduce noise from the input images as part of the pre-processing. These procedures will be described in detail below.

3.4.1. Selecting the Input and Frame splitting

In video format, a dataset for HAR based on thermal images is gathered. Preprocessing is performed in the IITR-IAR Dataset after video capture. Because the video is 30 frames per second, skipping every 12 frames in the collected dataset to gain one frame for training and testing the detection model. The attributes of the input videos are as follows:

Type	: VLC media file (.avi)
Dimensions	: 1024 x 768
Frame rate	: 30 frames per second

The video input is divided into frames. The frame rate is measured in frames per second. Frame rate is the number of visible frames per second.

The properties of the Thermal Simulated Fall input videos are,

Type	: VLC media file (.avi)
Dimensions	: 640 × 480
Frame rate	: 25 frames per second

It includes 9 non-fall scenario video portions and 35 fall scenario video portions. Other pre-processing methods are described in the following subsections.

3.4.2. Contrast limited adaptive histogram equalization

CLAHE is used to enhance the low image quality. This method improves image resolution by reducing amplification and clipping certain histogram thresholds. It is used to precisely improve the inverse of picture intensity. It is a superior approach for picture enhancement in comparison to adaptive histogram equalization (AHE) and other histogram equalization techniques due to its exceptional performance.

3.4.3. Gaussian – Adaptive Bilateral Filter

GABF is used to effectively eliminate the image's noise and provide improved edge preservation and smoothing. The proposed method may significantly increase infrared image quality. Input image I and Guidance g differ from the bilateral filter and is stated in eqn. (1):

$$f(x) = \sum_y W^g_{x,y}(g) I_y \tag{1}$$

Where, the center position of the input image is represented as I, $W^g_{x,y}$ is given as:

$$W^g_{x,y(g)} = \frac{1}{kx} \exp\left(-\frac{\|x-y\|^2}{\sigma_s^2}\right) \tag{2}$$

Where the normalizing factor is k_y. In Eqn. (2), the Gaussian spatial kernel is represented by second term

$\exp\left(-\frac{\|x-y\|^2}{\sigma_s^2}\right)$, the sizes of the window is denoted by. The GAB kernel is expressed as:

$$W_{x,y}^{gab}(I,g^{-}) = \frac{1}{kx}\exp\left(-\frac{\|x-y\|^2}{\sigma_s^2}\right)\exp\left(-\frac{\|I_x - g^{-}\|^2}{\sigma_r^2}\right) \quad (3)$$

Where g- is obtained by eqn. (1). In eqn. (3), the third term $\exp\left(-\frac{\|I_x - g^{-}\|^2}{\sigma_r^2}\right)$ is the range kernel. Variation in intensities is de by σ_r. The final output f(x) of the GABF can be expressed as,

$$f(x) = \sum_y W_{x,y}^{gabf}(I,\bar{g})I_y \quad (4)$$

The entire process of our preprocessing output is shown in figure 2.

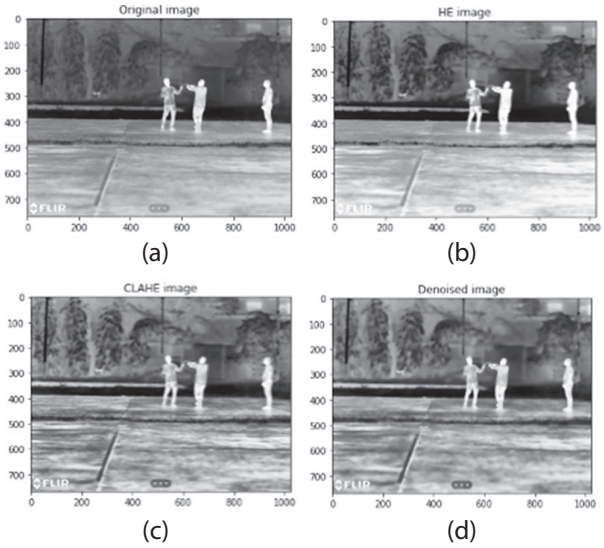


Fig. 2. Preprocessing results: (a) Original, (b) Enhance IR after HE, (c) CLAHE, (d) GABF

3.4.4. Handling Occlusion

Even though target monitoring has a wide variety of applications, it, however, poses challenges. Object or self-occlusion is the second most prevalent difficulty in target tracking applications, behind the noise. In some frames, the person is fully occluded by other things or people in the area. In addition, issues like changes in lighting, size variation, background clutter, etc. have a detrimental impact on the robustness of object tracking. For this reason, monitoring is based on a Gaussian – Adaptive Bilateral filter. These filters must be updated at every frame, even frames in which the target is obscured, therefore it is likely that background information is utilized to update the filter, which reduces the filters' discriminative power.

3.5. PROPOSED FRAMEWORK

The proposed technique recognizes human actions in thermal videos. Consequently, our objective is to address the issue of activity identification of various persons in a scenario. To this goal, we offer a paradigm influenced by recent development in deep learning-based Spatio-temporal recognition. The system consists of three distinct components: a) ROI detection frame-based, b) temporal-context extraction of feature and c) Spatial and temporal recognition.

3.5.1. ROI frame-level detector

This section is motivated by recent developments in object recognition [28]. Here, the DL method Mask-RCNN is employed for both person recognition and segmentation of pixels. It is the expansion of Faster RCNN that can localize and classify along with segmentation. It has two parts: (1) Convolutional backbone part: Convolutional backbone extracts image features; (2) Head part: It does bounding-box identification and mask prediction. A network for extracting key points must have a large number of convolution layers to learn more accurate and discriminative features. This research has employed Denenet-41 with Mask-RCNN. The backbone network Densenet-41 is utilized to obtain the important features. However, DenseNet-41 contains 24 channels on the first convolution layer and the size of the kernel is 3 × 3. In addition, numerous hidden layers within each dense block are adjusted for the computational complexity. The RPN network is fed the feature map generated from feature extraction to obtain RoIs. Figure 6 depicts an example of the ROIs extracted from the preprocessed thermal images.

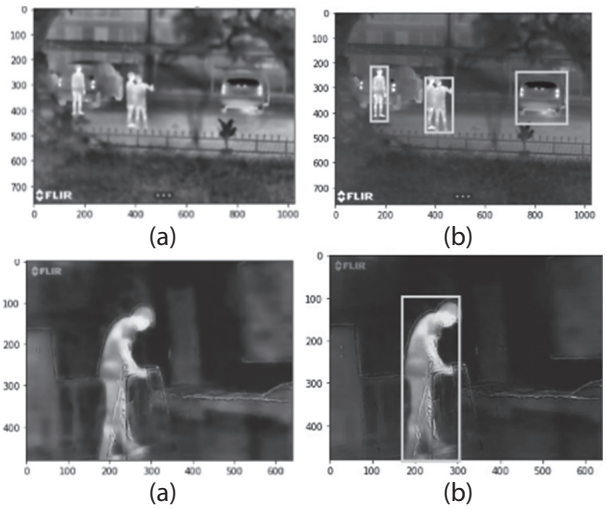


Fig. 3. (a) Preprocessed image, (b) ROI obtained

3.5.2. Temporal context-based feature extraction

To acquire temporal-context features of the targets in the image, employed the representation of different frames. The network is used for extracting temporal attributes in this sort of convolution. As a result, in the action recognition process, proposed DeepLabv3+Net as the backbone network for feature extraction. An encoding and decoding module comprises the architecture of deeplabv3+Net. The encoding module's goal is to extract temporal characteris-

tics from video frames in each video sequence step by step or layer by layer. Feature maps produced by deep layers acquire more abstract as layers are added, and they incorporate deeper semantic information that is useful to the pixel categorization. However, due to the stride of convolution, the temporal resolution of the feature maps reduces. This implies that local information such as borders and other features are removed from the feature map. As a result, a decoding module must be included. To construct a new feature map, the decoding module combines high- and low-resolution feature maps from shallow and deep layers. This new feature map comprises the borders of the action performers which are included in the feature set, as well as the semantics information that is useful for categorization. These characteristics are then fed into the classifier.

3.5.3. Spatio-temporal action prediction

This section aims to enhance the framework's Spatio-temporal features. Spatio-temporal characteristics strive to capture video motion patterns and give additional temporal information across frames. In fact, the frame-level detector employs Mask RCNN to extract Region of Interests. Using CV-FCNN of *k successive* frames (k = 10), we got the features from the temporal context. The motion characteristics are extracted by optical flow. Using the interim frames, the optical flow is calculated. The Spatio-temporal characteristics and ROIs are integrated to provide a more accurate representation of the scenes and their objects. Particularly the complex feature vector of size 1 × K, where K is the total number of classes, may be derived from the final convolution layer. As a result, at the feature map level, average pooling is employed to merge the ROI extracted component with the optical flow portion. The magnitude of the complex feature vector is computed at the output layer for bounding box prediction and confidence scores, and the softmax function is then applied to the final prediction.

3.5.4. Magnitude Operation

According to [29], the final convolution output in the hidden layer has complex values. Therefore, targeting is often a real-valued label. By transforming either label or output, the complex-valued output is compared with the real-valued label. In the output layer of CV-FCNN, the magnitude operation is performed to transform complex values into real numbers prior to softmax classification. Assume the last hidden layer convolution produces a 1 × K complex feature vector, then the magnitude is calculated by,

$$O_k^L = \sqrt{\left(\Re\left(O_k^{L-1}\right)\right)^2 + \left(\Im\left(O_k^{L-1}\right)\right)^2} \tag{5}$$

Where complex feature vector for kth component is represented as O_k^{L-1}.

3.5.5. Softmax Classification

In multiclass classification, Softmax is always used. It can transfer each real-valued vector component to 0–1, and the total equals 1, meeting the probability criteria. In the output layer of CV-FCNN, the softmax function is employed for action recognition. From eqn. (6), a real-valued vector with a size 1 × K can be obtained. Then the output of softmax classification is formulated by,

$$p_k = \frac{\exp\left(O_k^L\right)}{\sum_{m=1}^{K} \exp\left(O_m^L\right)} \tag{6}$$

Where, k^{th} class probability for one complicated visual or training sample is represented by p_k.

Training the classification model end-to-end attempts to eliminate loss functions as in Eqn. (7). Cross-entropy is used as a loss function.

$$Loss = -\sum_{k=1}^{k} q_k \, In \, p_k \tag{7}$$

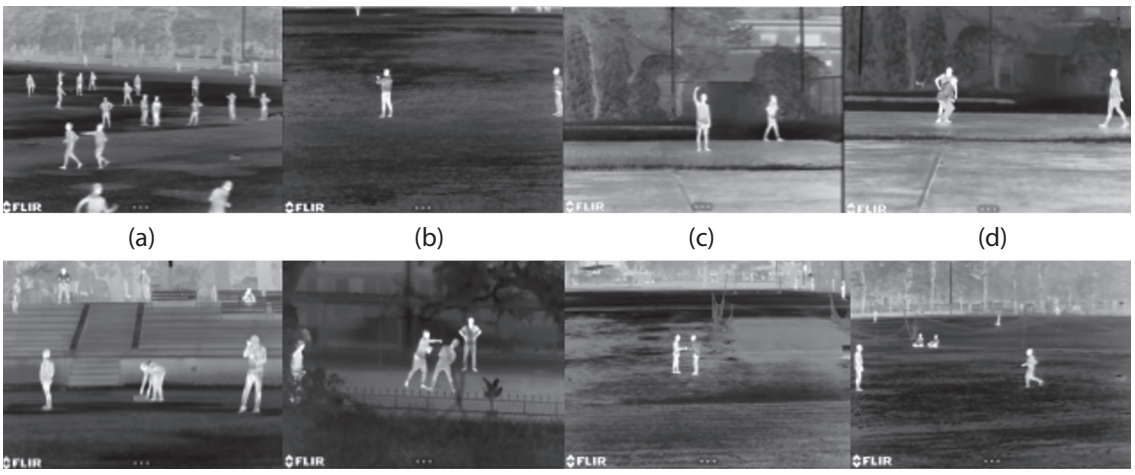
Where, the true classification result of one training sample is represented as q_k. That is if k is the training sample label is, then q_k is equal to 1; otherwise, q_k is equal to 0.

4. EXPERIMENTAL SETUP AND RESULTS

In this part, we discuss the research observations and assessments of the objective of human action recognition using frame-level and spatiotemporal information. The proposed system is capable of operating in both the image and video fields. To evaluate the performance of the proposed framework, provided with qualitative and quantitative outcomes. This technique will be evaluated using two commonly used action datasets (IITR-IAR and TSF) to assess HAR methods. The implemented model's efficacy is further shown by the results.

4.1. DATASETS

As seen in Figure 4, the IITR-IAR dataset contains 21 action types that may be generally categorized into three groups. They are (1) individual actions: two hand wave (wave2), one hand wave (wave1), walking (walk), running (run), hopping (hop), crouching (crouch) and clapping (clap). (2) person-object interaction: throwing object (throw), clicking selfie (selfie), recording video (video), picking up an object (pickup), carrying/pointing a gun (gun), dropping object (drop) and (3) person-person interaction: pushing (push), punching (punch), passing object (pass), kicking (kick), hugging (hug), handshaking (handshake), fighting (fight) and chasing (chase). For each action class, 70 videos containing 35 distinct individuals aged 8 to 37 have been gathered. As shown in Figure 5, a total of 44 videos are gathered, with 35 videos including a fall in addition to typical ADL and 9 videos simply containing ADL. There are several empty frames in the dataset, i.e. scenarios in which no human is present. It also includes shots of individuals attempting to enter from the left and right.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. IITR-IAR dataset a) chase, b) clap, c) Taking selfie, d) crouch, e) drop object, f) fight, g) handshaking, h) run

Fig. 5. Sample images from the Thermal ADL dataset

4.2. ANALYSIS METRICS

We evaluated the proposed model using the most standard metrics, including Accuracy, Precision, Recall, and F1-score. These are mathematically in equations (8) to (11)

$$\Pr ecision = \frac{TP}{TP + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - score = 2 \times \frac{\Pr ecision \times \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \tag{12}$$

True positive (TP) refers to violent action accurately detected, whereas false positive (FP) refers to nonviolent or normal activity wrongly forecasted as violent. True negative (TN) represents accurately determined nonviolent activity statistics. False negatives (FN) are violent events misclassified as nonviolence.

4.3. IMPLEMENTATION AND PARAMETER SETTINGS

Before training the model, appropriate adjustments of parameters are necessary. Here, chose 128 as the default sliding window size and 25 as the default stride for all experimental datasets, which is a tradeoff between classification performances. Furthermore, Adam is utilized to update CV-FCN parameters with a momentum of 0.9. The learning rate η is 0.0001. The mini-batch size is fixed at 30. Until the model converges, the training epoch number is 200. In addition, dropout regularization is used to reduce dimensionality. Experiments were performed on a personal workstation with an Intel Core i7 8th edition CPU, 4 GB GPU driver providing CUDA, 64 bits processor, RAM capacity of 8 GB, and Windows 10 operating system. The deep architecture has been implemented in Python.

We initially built the frame-level detector to recognize human actions using a single utilizing the keyframes taken from the clips. A frame is used as the detector's input in this portion. These qualities help identify activities by integrating contextual information in the boxes. Then temporal context attributes

(DeepLabV3+ and CV-FCN) are added and motion information (optical flow) to the frame-level framework as an initial point. These characteristics are designed to capture video motion patterns and give additional temporal information across frames.

4.4. TRAINING AND TESTING ON IITR-IAR DATASET

The proposed IITR-IAR dataset includes 21 action types. Each image is taken at a 1024×768 resolution. We chose 45 videos from each class at random as training images and the remaining 25 for testing. All of the tests are carried out multiple times.

4.4.1. Performance analysis of proposed method on IITR-IAR dataset

We provide demonstrations of qualitative outcomes for many sorts of situations in videos captured at various times. Adopting the frame-level and Spatio-temporal frameworks, these outcomes are produced. Figure 6 illustrates various sample outcomes of HAR in sequential video frames. In other instances, aggressive activities such as fighting, shooting, pursuing, and kicking are also seen. Individuals, on the other hand, exhibit normal activities such as sitting, standing, taking selfies, and sometimes walking at night. It is also observed that it is complicated to distinguish partial activities. It is correlated to the object's size. Also noted is that the integration of Spatio-temporal data enables the combination of motion and appearance characteristics across frames. Consequently, the proposed method distinguishes between normal and violent behavior.

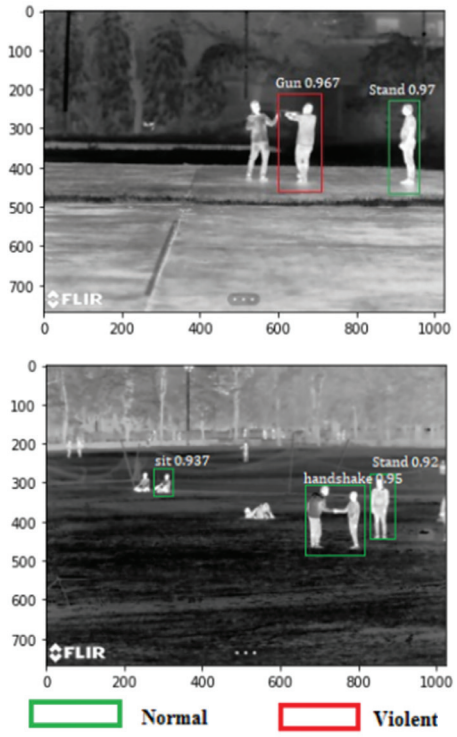


Fig. 6. Recognition results of normal and violent actions in an infrared video

Figures 7 and 8 depict the confusion matrices for the two distinct cases. It is evident from the data that the majority of complicated action examples are accurately identified. The minimum recognition accuracy (93%) is attained for routine outdoor activities that record actions. While the highest accuracy (96%) is attained for maximum normal actions and 97 percent for violent acts in an outdoor scenario. Only the punch and push classes are unclear. This is because both acts entail two people approaching one another and then utilizing either two hands (for a push) or a single hand (for a punch) to achieve the desired action. Overall, we reach a baseline average precision of 98.5% throughout the whole dataset. Figure 9 demonstrates that the ROC curves for both the normal and abnormal categories have desirable qualities and that the classification impact and generalizability are rather robust.

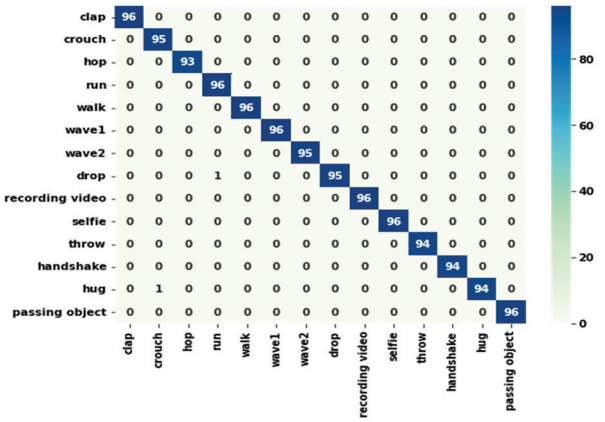


Fig. 7. Confusion matrix for Non-Violent activities

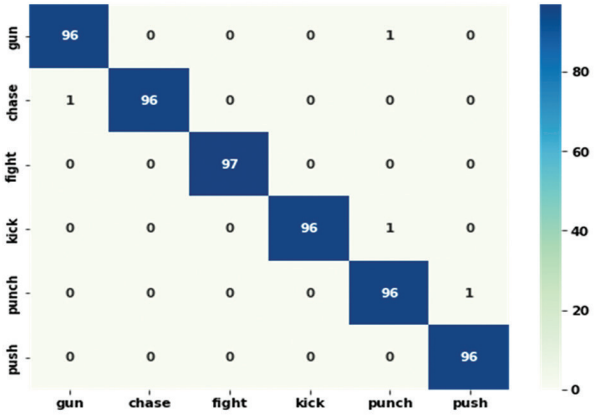


Fig. 8. Confusion matrix for violent activities

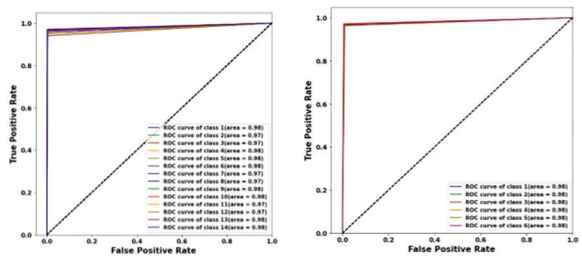


Fig. 9. ROC for normal activities and violent activities

4.4.2. Comparison with existing methods

To examine the role of Spatio-temporal characteristics in the network, we compared the outcomes of our experiment to earlier research. The quantitative outcomes of this comparison are shown in Table 1. The performance of the proposed method is compared to the RGB data set [30]. Table 2 demonstrates that the proposed technique outperforms the Hockey-Fight dataset. Overall, performance increased in all categories, but particularly in those with increased movement in terms of greater displacements. This makes it easier to acquire motion data using optical flow.

Table 1. Accuracy of action recognition

Actions	Existing [22] (%)	Proposed (%)
Gun	74	86.9
Chase	91.3	89.27
Clap	98.6	89.76
Selfie	74	94.82
Crouch	82	89.11
Drop	48.6	95.17
Fight	86.6	98.45
Handshake	76	90.26
Hop	90.67	94.35
Hug	77.33	96.67
Kick	64	96.57
Pass	60	90.21
Pickup	50	95.92
Punch	53.3	96.97
Push	60	94.76
Video	80.67	96.84
Run	83.33	91.14
Throw	89.33	93.17
Walk	76.67	87.92
Wave 1	93.33	97.87
Wave 2	81.33	98.21

Table 2. Comparison in different dataset

Dataset	Actions	Precision	Recall	F1-Score	Accuracy
Proposed (Thermal)	Violent Activity	0.97	0.98	0.98	98.5%
	Normal Activity	0.98	0.97	0.97	
Hockey-Fight Dataset (RGB) [30]	Violent Activity	0.96	0.97	0.97	96%
	Normal Activity	0.97	0.96	0.96	

4.5. TRAINING AND TESTING ON TSF DATASET

For the training phase, only standard ADL clips are utilized. The video clips are utilized to train the models were not labelled since they all represented as normal ADL. During the testing phase, fall videos with both normal and fall frames are employed. In these videos, the fall frames were labelled manually. The calculation of a test error may be used as an anomalous score to designate a fall frame as an abnormality. Quantitative findings are obtained by evaluating the proposed framework on 30% samples of the dataset.

4.5.1. Performance analysis of proposed method on TSF dataset

We provide an example of qualitative indoor environment outcomes. Figure 10 illustrates various sample findings of HAR in sequential video frames. In other instances, abnormal behaviors such as falling from standing, falling from a chair, and falling from a seated position are also recognized. Individuals, on the other hand, exhibit normal behaviors such as sitting, walking, laying down, bending, etc. Also, noted that partial activities are complex to recognize.

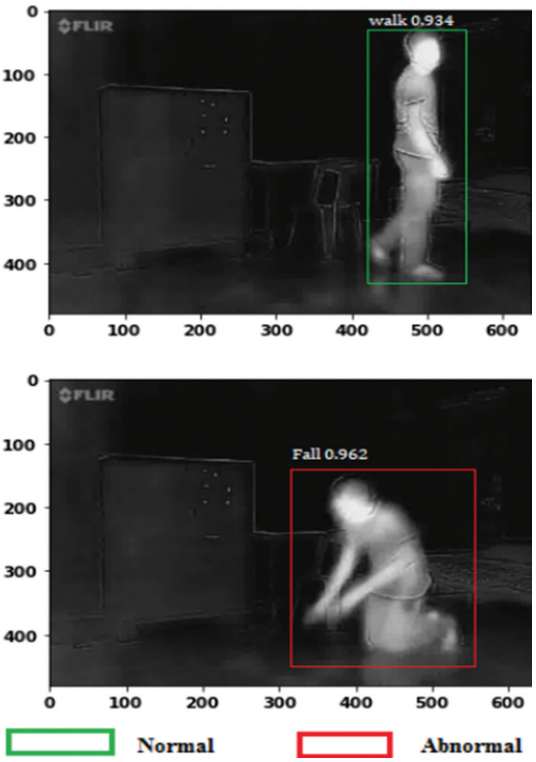


Fig. 10. Recognition results of normal and abnormal actions in an infrared video

Figures 11 and 12 illustrate the confusion metrics for indoor scenarios. It is evident from the data that the majority of complicated action examples are accurately identified. The lowest recognition accuracy (92%) is attained for routine indoor activities that capture actions. Maximum regular acts have the most accuracy (95%), whereas abnormal actions have the highest accuracy (97%). Only walk and stand classes confuse with each other since they have only minor changes. On the whole dataset, we reach 94.85% accuracy. Figure 13 demonstrates that the ROC curves for both the normal and abnormal categories have desirable qualities and that the classification impact and generalizability are rather robust.

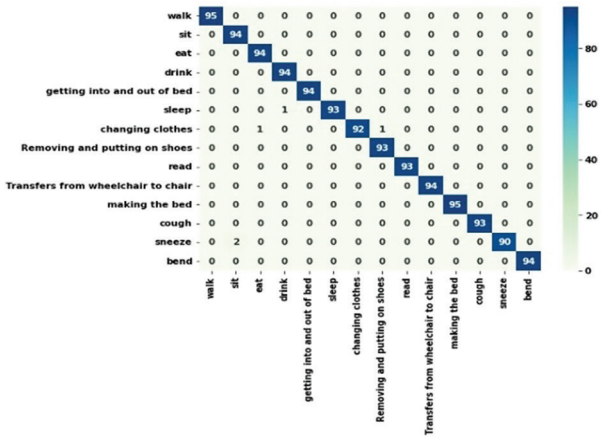


Fig.11. Confusion matrix for Normal activities

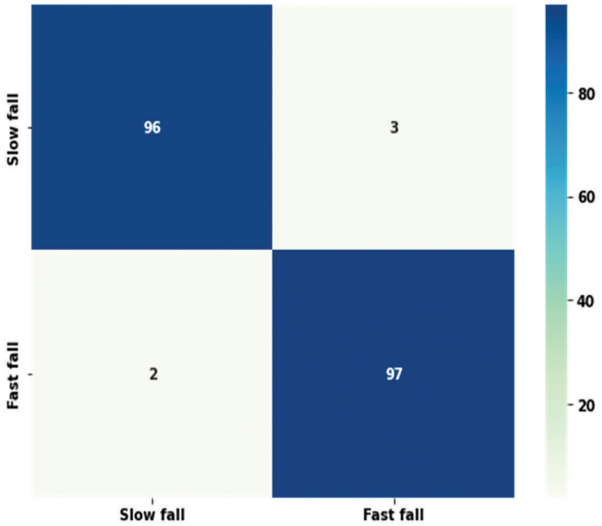


Fig. 12. Confusion matrix for abnormal activities

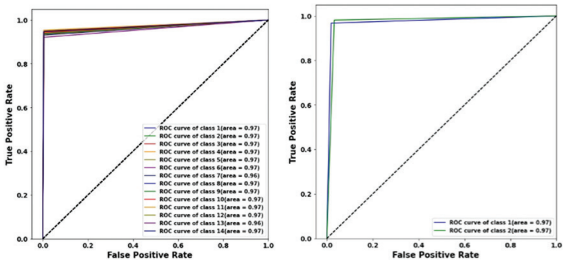


Fig. 13. ROC for normal activities and abnormal activities

Figure 14 depicts accurate predictions with their highest confidence ratings for a single action video based on our proposed system. A set of frames for each action is also provided for readers' convenience. It is noted from the experiments that the complex actions such as "kneel down" and "fall" which has a minor difference in the same action have more than 90% accuracy. The confidence score for normal actions like walking, sitting, kneeling, lying, sleeping, making the bed, bend show 86.88%, 92.17%, 93.3%, 95.48%, 95%, 97.22% and 90.08% respectively. The abnormal action like falls shows about 95.96% which reveals an effective result.

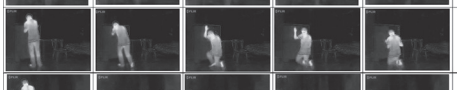
Sequence of frames representing an action	Predictions	Confidence score (%)
	walk	86.88
	sit	92.17
	Kneel down	93.3
	lying on the table	95.48
	sleep	95
	making the bed	97.22
	bend	90.08
	fall	95.96

Fig.14. Some prediction results on the TSF dataset with a maximum confidence score

4.5.2. Comparison with existing methods

The proposed model is compared with Deep Spatio-Temporal Convolutional Autoencoders (DSTCAE) [31] since they too are implemented on a similar dataset. The results for comparison of existing with the proposed method are shown in Table 3. It is noted that the proposed model performs better than the DSTCAE model. This is because the proposed approach extracts both spatial and temporal characteristics from videos that are essential for fall detection.

Table 3. Comparison with existing methods

Methods	Precision	Recall	F1-Score	Accuracy
Proposed	94.12	94.14	94.14	94.85%
DSTCAE [31]	-	-	-	93%

In conclusion, the findings demonstrated that the approach provided by this research is effective in recognizing activities in complicated indoor and outdoor scenarios.

5. CONCLUSION AND FUTURE SCOPE

Human activity detection is a complex issue with a wide range of applications in entertainment, autonomous driving, human-computer interaction and visual surveillance. This research proposes a deep learning strategy for recognizing human behavior using Spatio-temporal data. The framework includes frame-level appearance attributes and spatial and temporal information with temporal-context features. The proposed work identifies multiple regions containing human activities, unlike single-action approaches. The experiment is conducted on IITR-IAR and TSF datasets. The system can distinguish 38 sequential behaviors divided

into normal, abnormal and violent activities. Qualitative and quantitative findings illustrate our framework's monitoring effectiveness. This research reveals new issues in detecting human behavior for optimum conditions. Future work of this research may include control appliances, yoga analysis, sports actions through human body poses etc. In addition, the proposed method may be evaluated in more realistic conditions, such as zooming in and out etc.

6. REFERENCES

[1] V. John, S. Mita, A. Lakshmanan, A. Boyali, S. Thompson, "Deep Visible and Thermal Camera-Based Optimal Semantic Segmentation Using Semantic Forecasting", Journal of Autonomous Vehicles and Systems, Vol. 1, 2021, pp. 106-119.

[2] F. S. Leira, H. H. Helgesen, T. A. Johansen, T. I. Fossen, "Object detection, recognition, and tracking from UAVs using a thermal camera", Journal of Field Robotics", Vol. 38, 2021, pp. 242-267.

[3] M. Kristo, M. Ivasic-Kos, M. Pobar, "Thermal object detection in difficult weather conditions using YOLO", IEEE Access, Vol. 8, 2021, pp. 125459-125476.

[4] J. Kim, J. Cho, "Low-cost embedded system using convolutional neural networks-based spatiotemporal feature map for real-time human action recognition", Applied Sciences, Vol. 11, 2021, p. 4940.

[5] V. Parameswari, S. Pushpalatha, "Human Activity Recognition using SVM and Deep Learning", European Journal of Molecular & Clinical Medicine, Vol. 7, 2020, pp. 134-144

[6] A. B. Sargano, X. Gu, P. Angelov, Z. Habib, "Human action recognition using deep rule-based classifier", Multimedia Tools and Applications, Vol. 79, 2020, pp. 30653-30667.

[7] M. E. N. Gomes, D. Macêdo, C. Zanchettin, P. S. G.de-Mattos-Neto, A. Oliveira, "Multi-human fall detection and localization in videos", Computer Vision and Image Understanding, Vol. 1, 2022, pp. 103442.

[8] F. S. Leira, H. H. Helgesen, T. A. Johansen, T. I. Fossen, "Object detection, recognition, and tracking from UAVs using a thermal camera", Journal of Field Robotics, Vol. 38, 2021, pp. 242-267.

[9] C. Dhiman, D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics", IEEE Transactions on Image Processing, Vol. 29, 2020, pp. 3835-3844.

[10] H. Liu, Y. Chen, W. Zhao, S. Zhang, Z. Zhang, "Human pose recognition via adaptive distribution encoding for action perception in the self-regulated learning process", Infrared Physics & Technology, Vol. 114, 2021, p. 103660.

[11] R. Kapoor, R. Goel, A. Sharma, "Deep learning based object and railway track recognition using train mounted thermal imaging system", Journal of Computational and Theoretical Nanoscience, Vol. 17, No. 11, 2020, pp. 5062-5071.

[12] M. A. Khan, Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman, S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network", Multimedia Tools and Applications, Vol. 80, 2021, pp. 35827-35849.

[13] K. Thapa, A. Al, Z. Md, B. Lamichhane, S. H. Yang, "A deep machine learning method for concurrent and interleaved human activity recognition", Sensors, Vol. 20, 2020, p. 5770.

[14] D. Zhou, S. Qiu, Y. Song, K. Xia, "A pedestrian extraction algorithm based on single infrared image", Infrared Physics & Technology, Vol. 105, 2020, p. 103236.

[15] Q. Kang, H. Zhao, D. Yang, H. S. Ahmed, J. Ma, "Lightweight convolutional neural network for vehicle recognition in thermal infrared images", Infrared Physics & Technology, Vol. 104, 2020, p. 103120.

[16] P. Kovács, B. Lehner, G. Thummerer, G. Mayr, P. Burgholzer, M. Huemer, "Deep learning approaches for thermographic imaging", Journal of Applied Physics, Vol. 128, 2020, p. 155103.

[17] M. Awais, X. Long, B. Yin, S. F. Abbasi, S. Akbarzadeh, C. Lu, W. Chen, "A hybrid DCNN-SVM model for classifying neonatal sleep and wake states based on facial expressions in video", IEEE Journal of Biomedical and Health Informatics, Vol. 25, No. 5, 2021, pp. 1441-1449.

[18] C. Dhiman, D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics", IEEE Transactions on Image Processing, Vol. 29, 2020, pp. 3835-3844.

[19] D. K.Vishwakarma, "A two-fold transformation model for human action recognition using decisive pose", Cognitive Systems Research, Vol. 61, 2020, pp. 1-13.

[20] H. Liu, X. Wang, W. Zhang, Z. Zhang, Y. F. Li, "Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition", Neurocomputing, Vol. 411, 2020, pp. 510-520.

[21] S. A. Manssor, S. Sun, M. Abdalmajed, S. Ali, "Real-time human detection in thermal infrared imaging at night using enhanced Tiny-yolov3 network", Journal of Real-Time Image Processing, Vol. 19, 2022, pp. 261-274.

[22] J. Imran, B. Raman, "Deep residual infrared action recognition by integrating local and global spatio-temporal cues", Infrared Physics & Technology, Vol. 102, 2019, pp. 103014.

[23] M. Krišto, M. Ivasic-Kos, M. Pobar, "Thermal object detection in difficult weather conditions using YOLO", IEEE Access, Vol. 8, 2020, pp. 125459-125476.

[24] G. Batchuluun, J. K. Kang, D. T. Nguyen, T. D. Pham, M. Arsalan, K. R. Park, "Action recognition from thermal videos using joint and skeleton information", IEEE Access, Vol. 9, 2021, pp. 11716-11733.

[25] M. Ding, Y. Y. Ding, X. Z. Wu, X. H. Wang, Y. B. Xu, "Action recognition of individuals on an airport apron based on tracking bounding boxes of the thermal infrared target", Infrared Physics & Technology, Vol. 117, 2021, p. 103859.

[26] H. Hei, X. Jian, E. Xiao, "Sample weights determination based on cosine similarity method as an extension to infrared action recognition", Journal of Intelligent & Fuzzy Systems, Vol. 40, 2021, pp. 3919-3930.

[27] A. M. De Boissiere, R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition", IEEE Access, Vol. 8, 2020, pp. 168297-168308.

[28] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, 2017, pp. 1137–1149.

[29] L. Yu, Y. Hu, X. Xie, Y. Lin, W. Hong, "Complex-valued full convolutional neural network for SAR target classification", IEEE Geoscience and Remote Sensing Letters, Vol. 17, 2019, pp. 1752-1756.

[30] S. Habib, A. Hussain, W. Albattah, M. Islam, M. S. Khan, R. U. Khan, Khan, K. "Abnormal Activity Recognition from Surveillance Videos Using Convolutional Neural Network", Sensors, Vol. 21, 2021, p. 8291.

[31] J. Nogas, S. S. Khan, A. Mihailidis, "Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders", Journal of Healthcare Informatics Research, Vol. 4, 2020, pp. 50-70.