Real-time Aerial Suspicious Analysis (ASANA) System for the Identification and Re-Identification of Suspicious Individuals in Crowds using the Bayesian ScatterNet Hybrid (BSH) Network

Kranthi Kiran GV, Onkar Harsh, Rishav Kumar, Koushlendra Singh Rajput, Chandra S S Vamsi
Skylark Labs LLP.
Warangal, India
{kgkiran, oharsh, rkumar, ksrajput,cssvamsi}@skylarklabs.ai

Amarjot Singh
Skylark Labs LLC.
San Francisco, USA
amarjot@skylarklabs.ai

Abstract

Video monitoring and safety systems have been used to keep track of hostiles, conduct border control operations as well as to monitor the suspicious entities in public spaces. However, these systems are inadequate for the monitoring of large crowds due to the limited field of view of cameras. This paper introduces the Aerial Suspicious Analysis (ASANA) System for the Identification and Re-Identification of suspicious Individuals in large public areas using the Bayesian ScatterNet Hybrid (BSH) Network. The BSH network first estimates the human pose in each frame. Next, a batch of frames is used by the Bayesian 3D ResNext to identify individuals with suspicious postures. The system can also re-identify the identified suspicious individuals as they tend to move after committing the suspicious event. The proposed architecture is advantageous as it can learn meaningful representations quickly using the ScatterNet with fewer labelled examples. This is of great importance as real-life annotated training samples are hard to collect, especially for these applications. The pose estimation, suspicious individual identification, and re-identification performance of the proposed framework is compared with the state-of-the-art techniques. The proposed dataset is also made public which may encourage other researchers who are interested in using the deep learning technique for aerial visual crowd monitoring.

1. Introduction

Due to the significant rise of criminal activities in recent years, law enforcement agencies have been motivated to use automated video monitoring systems to identify suspicious activities. Several such automated systems have been developed in the past to monitor abandoned objects (bags) [11], theft [4], violent activities [7], etc.

Li et al. [11] developed a video monitoring system to detect abandoned objects using Gaussian mixture models and Support Vector Machine. Their system detected the objects of interest with an accuracy of 84.44%. Such systems can be vital for the identification of abandoned bags in public areas, which may contain bombs. Chuang et al. [4] developed a system to detect robberies using ratio histogram and a finite state machine. The system has been proven to be useful at automatic teller machines (ATMs) and has successfully identified 96% cases of theft. Goya et al. [7] presented a public safety system (PSS) for automatic detection of crimes such as purse snatching, child kidnapping and fighting using distance, velocity, and area to determine the human behavior. This system performed with an accuracy of around 85%.

The above systems have been very successful in detecting and reporting various criminal activities. Despite this, these systems are unable to monitor large areas due to the limited field of view of the cameras. Governments have recently deployed drones to invigilate vast territories to track hostiles in war zones, to spy on drug cartels [14], conduct border control operations [25] etc. One or more soldiers usually pilot these drones for long durations which makes these systems prone to mistakes due to human fatigue.

A handful of autonomous aerial systems have recently been introduced to overcome this challenge. Kumar et al. [10] proposed a real-time system to track moving objects in aerial videos using different image-processing techniques. Layne et al. [10] introduced an offline framework

Figure 1. Illustration presents the suspicious activities from the introduced ASIIR dataset namely (clockwise from top) (i) Strangling, (ii) Pushing, (iii) Sword Fighting.

to re-identify manually selected humans in videos recorded from a drone. Surya et al. [15] proposed an offline UAV framework to detect suspicious humans in public areas with an accuracy of 76%. Most recently, Singh et al [24]. introduced the aerial violent individual recognition system. This system can identify violent individuals with an average accuracy of 79%. The performance of these autonomous systems is inferior as compared to the performance of ground surveillance systems ($\sim$75% vs. $\sim$90%). An ideal system should be able to identify and re-identify persons of interest in real-time with reasonable accuracy.

This paper introduces the real-time Aerial Suspicious Analysis (ASANA) System for the Identification and Re-Identification of suspicious individuals in large public crowds using the proposed Bayesian ScatterNet Hybrid (BSH) Network. The 4k resolution video recorded by the drone is first decomposed into frames. The BSH network then extracts the humans and estimates their poses using the ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network constructed by replacing the first convolutional, relu and pooling layers of the Part Affinity Fields (PAFs) network [1] with the hand-crafted ScatterNet [20] as shown in Fig. 3. The poses are then used to identify suspicious (Fig. 1) individuals using the Bayesian 3D ResNext [5] with uncertainty estimates. The features obtained from the deeper layer of the BSH network are also used to perform one-to-one person re-identification.

The novelties of the proposed BSH network are detailed below:

- ***Rapid learning with ScatterNet and Structural Priors***: The proposed SH-PAF network is constructed by with the hand-crafted ScatterNet (front-end) that extracts translation, rotation, and scale invariant low-level edge features from the input images (similar to the replaced layer). These features can be used by the PAF (back-end) network to learn more complex features from the start of learning as the edges are already present, resulting in accelerated training. The ScatterNet invariant features are particularly useful for this application as the human can appear at different locations, orientations, and scales. The training of the PAF network is further accelerated by initializing the filter weights with structural priors learned (unsupervised) using the PCANet [3] framework (Fig. 3). The initialization with priors also ***reduces the need for sizeable labeled training datasets*** for effective training which is especially advantageous for this task or other applications [19, 8] as it can be expensive and time-consuming to generate keypoint annotations.
- ***Bayesian Uncertainty***: The proposed network uses dropout at test time to make several predictions. These predictions are then used to compute the mean and standard deviation which can aid the user in deciding if a certain prediction can be trusted.
- ***Suspicious Individuals Re-identification***: The features obtained from the deeper layer of the BSH network are also used to perform one-to-one suspicious person re-identification. This is important for a safety and security system as the suspicious individuals may disappear and then reappear (without necessarily performing the suspicious activity) in subsequent images.
- ***Real-time Processing***: The proposed system performs the computation and memory demanding suspicious individual identification and re-identification processes on the cloud while keeping short-term navigation onboard. This allows the system to identify suspicious individuals in real-time.
- ***Aerial Suspicious Individuals Identification and Re-identification (ASIIR) Dataset***: The paper proposes ASIIR dataset that contains images with humans engaged in different suspicious activities (Section 2) recorded at different variations of scale, position, illumination, blurriness, etc. The dataset is collected such that it is suitable for human re-identification task with suspicious individuals appearing, disappearing and reappearing (without necessarily performing the suspicious activity) in subsequent images. This dataset may encourage researchers interested in using deep learning for aerial safety and security applications.

The pose estimation, individuals identification, and re-identification performance of the system is compared with

the state-of-the-art techniques.

The paper is divided into the following sections. Section 2 presents the introduced ASIIR dataset while Section 3 introduces the proposed ASANA system. Section 4 details the experimental results and Section 5 concludes this research.

2. Aerial Suspicious Individuals Identification and Re-identification (ASIIR) Dataset

This paper proposes a labelled aerial suspicious individuals identification and re-identification (ASIIR) dataset which is used by the proposed system to identify suspicious individuals. The dataset comprises of 2000 images with each image containing more than 50 persons. The dataset in total consists of 108570 humans with 51240 (48%) engaged in one or more of the six suspicious activities such as (i) Strangling, (ii) Punching, (iii) Kicking, (iv) Shooting (v) Stabbing and (vi) Sword fighting. Each person in the aerial image is annotated with 18 key-points which are used by the network as labels to estimate pose as shown in Fig. 2. The activities are performed by 25 subjects between the ages of 18-25 years. The images are recorded from the Parrot AR. Drone 2.0 at four heights of 50m, 100m, 150m and 200m (m: meters). The dataset is suitable for re-identification problem since the images are captured such that a person or more currently in a frame would be out of the frame later because of the moving drone and then the person(s) re-appear in the subsequent frame due to (1) Re-spotting by moving drone, (2) Drone increases field-of-view by flying upward and hence the person(s) re-appearing in the frames.

The suspicious person identification and re-identification tasks from these aerial images is an extremely challenging task as these images can be affected by variation in illumination, shadows, poor resolution, and blurring. The persons can also appear at different locations, orientations, and scales. The proposed dataset includes images with the above-detailed variations since they can significantly alter the appearance of the humans and affect the performance of the surveillance systems. The BSH network, when trained on the dataset with these variations, can learn to model them and recognize suspicious humans in spite of such variations.

3. Aerial Suspicious Analysis (ASANA) System

This section introduces the Aerial Suspicious Analysis (ASANA) System which first uses the proposed Bayesian ScatterNet Hybrid (BSH) Network for first human pose estimation using the ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network followed by the classification of individuals as suspicious or non-suspicious using the Bayesian 3D ResNet. The ScatterNet features are also used to perform re-identification of the recognized suspicious individuals. The system uses cloud computation to achieve the identification and re-identification in real-time. Each part of the ASANA system is explained in the following sub-sections.

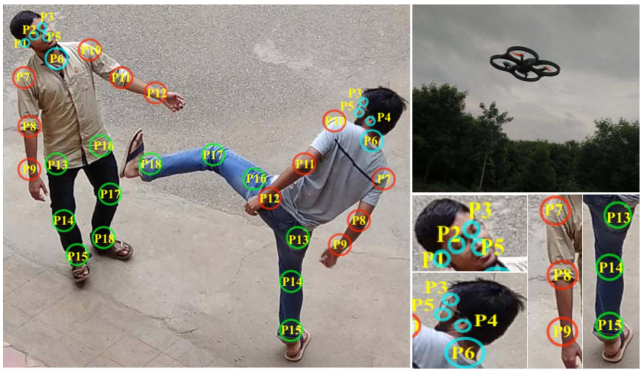

Figure 2. The figure (left) illustrates the 18 anatomical keypoints annotated in the dataset. The description of the keypoint is as: Head region (cyan): P1-right ear, P2-right eye, P3-left eye, P4-left ear, P5-nose, P6-Neck; Arm region (red): P7-right shoulder, P8-right elbow, P9-right wrist, P10-left shoulder, P11-left elbow, P12-left wrist; Leg region (green): P13-right hip, P14-right knee, P15-right ankle, P16-left hip, P17-left knee, P18-left ankle. The figure (right) shows the Parrot AR Drone 2.0 which is used to capture the images in the dataset and close-ups of a few keypoints

3.1. ScatterNet Hybrid Part Affinity Fields

This section details the proposed ScatterNet Hybrid Part Affinity Fields (SH-PAF) Network, inspired from Singh et al.'s work in [21, 22, 19, 23], composed by combining the hand-crafted (front-end) two-layer parametric log ScatterNet [20] with the pruned Parts Affinity Fields (PAFs) [2] network (back-end) as shown in Fig. 3. The ScatterNet accelerates the learning of the SH-PAF network by extracting invariant edge-based features which allow the network to learn complex features from the start of the learning [21]. The pruned PAF also uses structural priors to expedite the training as well as reduce the dependence on the annotated datasets. The ScatterNet (front-end) and pruned Parts Affinity Fields (PAF) are presented below.

ScatterNet (front-end): The parametric log based ScatterNet [20] is a two-layer hand-crafted network which extracts translation, rotation, and scale invariant feature representations from multi-resolution images obtained at 1.5 times and twice the size of the input image. Below we present the formulation of the parametric ScatterNet for a single input image which may then be applied to each of the multi-resolution images.

The invariant features are obtained at the first layer by filtering the input image or signal x with dual-tree complex wavelets (better than cosine transforms [9]) $\psi_{j,r}$ at different scales (j) and six pre-defined orientations (r) fixed to $15°, 45°, 75°, 105°, 135°$ and $165°$. To build a more translation invariant representation, a point-wise L_2 non-linearity (complex modulus) is applied to the real and imag-

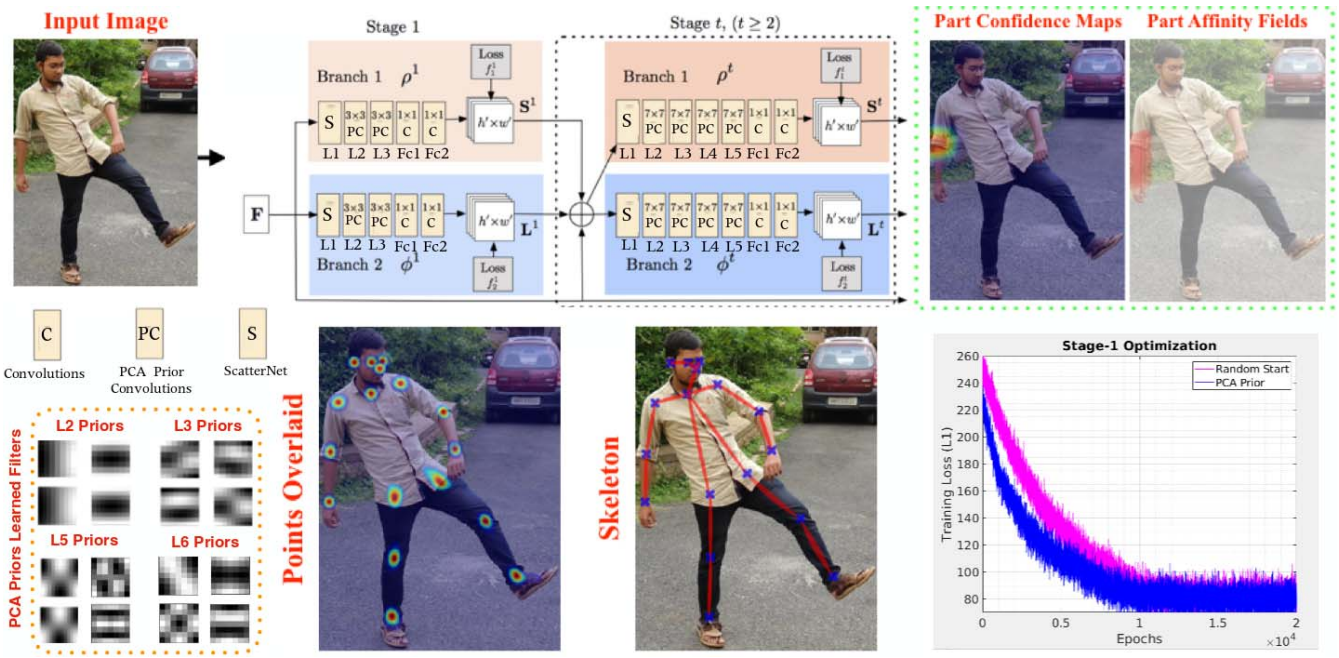

Figure 3. Key-point detection pipeline: The illustration shows the location of all P1-P18 keypoints generated as heatmaps by the part affinity fields deep regression convolutional network. The keypoints are overlaid and a skeleton connecting them is formed as shown in the right. The outputs of two branches of the network, Part Confidence Maps and Part Affinity Fields (PAF), are also shown.

inary part of the filtered signal:

$$U[\lambda_{m=1}] = |x \star \psi_{\lambda_1}| = \sqrt{|x \star \psi^a_{\lambda_1}|^2 + |x \star \psi^b_{\lambda_1}|^2} \quad (1)$$

The parametric log transformation layer is then applied to all the oriented representations extracted at the first scale $j = 1$ with a parameter $k_{j=1}$, to reduce the effect of outliers by introducing relative symmetry of pdf [20], as shown below:

$$U1[j] = \log(U[j] + k_j), \quad U[j] = |x \star \psi_j|, \quad (2)$$

Next, a local average is computed on the envelope $|U1[\lambda_{m=1}]|$ that aggregates the coefficients to build the desired translation-invariant representation:

$$S_1[\lambda_{m=1}] = |U1[\lambda_{m=1}]| \star \phi_{2^J} \quad (3)$$

The high frequency components lost due to smoothing are retrieved by cascaded wavelet filtering performed at the second layer. Translation invarinace is introduced in these features by applying the L2 non-linearity with averaing as explained above for the first layer [20].

The scattering coefficients at L0, L1, and L2 are:

$$S = \left(x \star \phi_{2^J}, S_1[\lambda_{m=1}], S_2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J}\right) \quad (4)$$

The rotation and scale invariance are next obtained by filtering jointly across the position (u), rotation (θ) and scale(j) variables as detailed in [18].

The features extracted from each resolution at L0, L1, and L2 are concatenated and given as input to the pruned Part Affinity Fields (PAFs) network, to learn high-level features for human pose estimation. The ScatterNet features help the proposed SH-PAF network to converge faster as the convolutional layers of the PAF network can learn more complex patterns from the start of learning as it is not necessary to wait for the first layer to learn invariant edges as the ScatterNet already extracts them.

Pose Estimation with Structural Priors (back-end): The invariant ScatterNet features are used by the pruned Parts Affinity Fields (PAFs) network [2] (initial layers replaced with ScatterNet) to learn pose estimation using the introduced ASIIR Dataset. The ASIIR dataset contains aerial images with 18 annotated key-points with 36 coordinates (section 2) on the human body which are used by the network to learn the human poses.

The pruned PAF network is composed of a feedforward network which is divided into two branches which simultaneously predicts a set of confidence maps S of body part locations and a set of vector fields L of part affinities for each limb which preserves both the position and orientation information of the limb. The predictions of each of the branch are iteratively refined over successive stages [27]. Finally, the confidence maps and part affinity fields are parsed by greedy inference to output the body keypoints for all people in the image.

Structural Priors: In order to accelerate the training,

Figure 4. Human Detection at different Scales: Pose estimation is performed for humans which appear at different scales by dividing the 4K resolution frames recorded by the UAV into 72 regions as shown. Selected regions are shown. The reasoning for this processing is detailed in section 3.1.

each convolutional layer of the joints identification network of the SHDL network is initialized with structural priors. The structural priors are obtained for each layer using the PCANet [3] framework. The structural priors for the first layer of the pruned PAF network are learned on the ScatterNet features, the structural priors for the following layers are learned on the previous layers outputs and so on. This is applied to both of the branches present in the network. The structural priors for the pruned PAF networks layers learn filters that respond to a hierarchy of features which is similar to the features learned by CNNs. These learned structural priors are used to initialize each of the convolutional layer resulting in accelerated training. Since the determination of structural prioris is fast, the training process is much faster than that of CNNs with random weight initializations. However, the PCA framework may learn undesired checkerboard filters. In order to detect the checkerboard filters from the learned filter sets, we use the method defined in [6] and are then avoided as filter priors.

Human Pose Estimation at different Scales: Pose estimation is performed for humans which appear at different scales by dividing the 4K resolution frames recorded by the UAV into 72 regions as shown in Fig. 4. The resultant subregions correspond to the scale and resolution of the patches which were used to train the pose estimation model. This allows the model to estimate the pose of humans at various scales.

3.2. Suspicious Individual Classification using Bayesian 3D-ResNext

A 3D ResNext [5] is trained on 16 subsequent frames to recognize the individuals with suspicious as shown in Fig. 5. In the proposed system, we use Monte Carlo dropout at prediction time to measure the 3D ResNext model's uncertainty. We make 50 predictions at test time with dropout enabled. The variance of these predictions can be used to measure how certain the model is about the prediction.

3.3. Drone Image Acquisition and Cloud Processing

The images in ASIIR dataset are collected using a Parrot AR Drone 2.0. It has two inbuilt cameras, an Inertial Measurement Unit (IMU), an ultrasound, a pressure-based altitude sensor and 1 GHz ARM Cortex-A8 CPU.Its front-facing camera has the resolution of 3840x1920 and capture at the rate of 30fps with a diagonal field of view of 92° and is used for taking images. Its down facing camera has lower resolution (320x240) and capture rate of 60 fps speed with a diagonal field of view of 64 ° and is used to estimate roll, pitch, yaw, and altitude for the drone. All the sensor measurements are updated at the 200Hz rate. To achieve real-time identification and re-identification of violent individuals, the images captured from drones are transferred to Amazon cloud, and the memory intensive and slow computations of SHDL network are processed on it.

3.4. Violent Individual Re-identification

The suspicious person is re-identified for subsequent frames using Deep ScatterNet features [20] as shown in Fig. 6. The feature vector is extracted from 5x5 patches around the 18 key points. The suspicious person is considered to be re-identified as one of the people in the frame if the cosine the similarity is maximum, or distance is minimum between the feature vector for the suspicious person in the initial and subsequent frames.

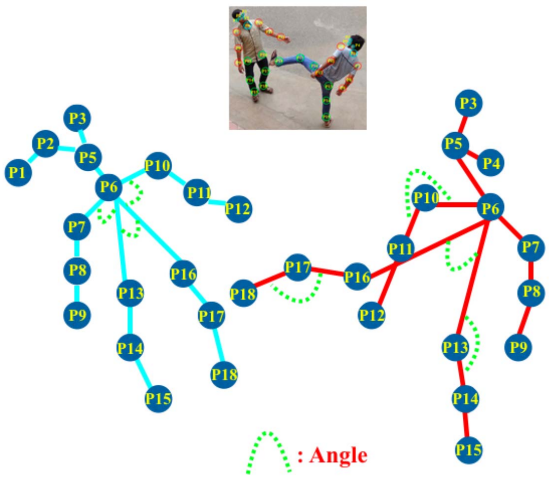

Figure 5. Suspicious Pose Identification: The figure shows the fashion in which the body keypoints are connected to form a skeleton corresponding to the suspicious action being performed in the small image above. Bayesian 3D-ResNext uses 16 frames to learn representations on the angles of this structure to identify suspicious poses. The model also produces uncertainty along with the predictions which is critical for this application.

4. Experimental Results

This section presents the results of the experiments performed on the introduced dataset, ASIIR dataset using the ASANA System for the identification and re-identification of suspicious individuals. The ASANA systems comprised of BSH network for the pose estimation of the humans. The detected joints are used to form human skeletons which are used by the 3D ResNext over a number of frames to identify suspicious individuals. The features of the BSH networks are also used to perform person re-identification. Further coming sections detail the performance of each subpart of the proposed framework. The performance of each part of the proposed system is presented and compared with the state-of-the-art.

4.1. SH-PAF Parameters and Training

The SH-PAF network is composed of the scatternet followed by the pruned PAF network.

ScatterNet: The scatternet extracts invariant low-level features using DTCWT filters at 2 scales, and 6 fixed orientations at layers L0, L1, and L2.

PAF Network with Structural Priors: The back end of BSH network is a PAF which is trained on the features obtained from the front end of SHDL network (scatternet). The scatternet features are extracted from the 108570 humans. Out of 108570 humans, 60% of extracted concatenated features used as training set and 20% for testing and 20% for the PAF network. The splitting of the dataset of extracted features is completely random. The network parameters are as follows: The base learning rate is 10^{-5}, which we decrease to 10^{-6} after 15 iterations, the dropout is 0.5, the batch size is 32, and the total number of iterations (epochs) is 140. To accelerate the training, the convolutional layers are initialized with structural priors.

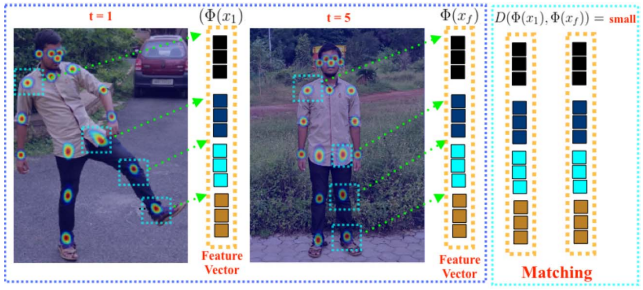

Figure 6. Re-identification: The figure shows a person at different time steps. Initially, the person is in a suspicious pose while in the subsequent frame the person is in normal pose. The system re-identifies him using the features extracted using the deep ScatterNet around the keypoints and Cosine Similarity, when a person is detected again.

4.2. Key-Point Detection Performance

In this section, we analyze the performance of key point detection by the proposed BSH network. The coordinates of the detected 18 key-points and the corresponding ground truth values in the annotated dataset are compared to analyze the performance of the module. The performance of the proposed framework is presented in the form of graphs that plot accuracy vs. the distance from the ground truth pixels (with a key-point considered to be correctly located if it is within a set distance of d pixels from the marked key-point center in the ground truth).

The key-point detection performance for the dataset is plotted for each key-point as shown in Fig 6. We observe that the accuracy increases as the distance from the ground truth pixel d, within which we consider that the key-point is marked correctly, increases.

4.3. Region-wise Key-Point Performance Analysis

This section further presents the analysis of detection performance of the various key-points present in each of the regions of the body.

1) Facial Key-Points Detection Performance: The following are two points are constituted by arms region: P1- on the head, P2- on the neck as shown in Fig. 1. We observe that the accuracy of keypoint of the neck (P2) is substantially higher than that of the head. This can be seen in figure 6, the third graph constitutes the head region, where it can be seen that the neck keypoint detection has significantly higher accuracy.For a pixel distance of $d = 5$, the neck point has an accuracy of 97% as compare of the head point with an accuracy of 84%.

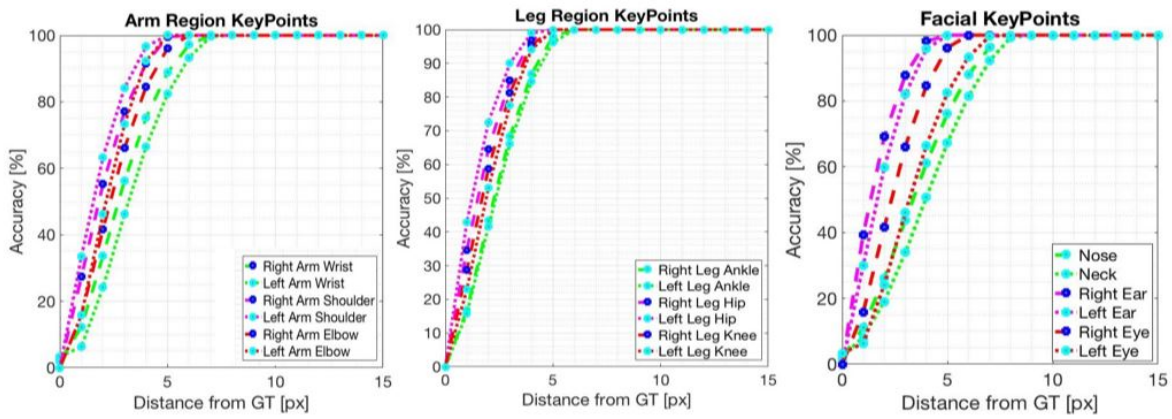

Figure 7. The figure shows the keypoint detection performance graphs for each of the 18 keypoints considered. The first graph represents the arm region, which constitutes of the wrist, shoulder and elbow keypoints; second graph represents the leg region which constitutes of the ankle, hip and knee keypoints and the third graph represents the head region which constitutes of the ear, eyes, nose and neck keypoints.

2) Arms Region Key-Point Detection Performance: The following are six points are constituted by arms region: P6-right shoulder, P7-right elbow, P8-right wrist, P3-left shoulder, P4-left elbow and P5-left wrist as shown in Fig. 2. Fig. 7(a) indicates that the SHDL network can detect the wrist region key-points with an accuracy of around 60%, for a pixel distance of $d = 5$. The accuracy of detection of elbow and shoulder is comparatively higher, roughly 85% and 95% respectively, for the same pixel distance ($d = 5$).

3) Legs Region Key-Points Detection Performance: The following are six points are constituted by legs region: P12-right hip, P14- right knee, P13-right ankle, P9-left hip, P17-left knee, P10- left ankle. In the legs region, the accuracy of keypoint detection of the hip region is higher nearly 100% for a pixel distance of $d = 5$. It can be inferred from the second graph in Fig 6. While detection accuracy for knee region lies in 85% to 90% and for ankle region, it falls to around 85%.

4.4. SH-PAF Performance and comparison

The human pose estimation performance of the BSH network on the ASIIR dataset is presented in Table 1. We have considered 3 architectures namely CoordinateNet (CN), CoordinateNet extended(CNE) and SpatialNet. As observed from the Table for the comparison. The BSH network estimates the human pose based on the 18 key-points at $d = 5$ pixel distance from the ground-truth, with 91.6% accuracy.

The key-point detection accuracy results for the dataset are 79.6%, 80.1%, 83% and 87.66% for Coordinate Network [16], Coordinate Network Extended [16], SpatialNet [17] and proposed SH-PAF network, respectively. The SH-PAF network outperforms the other networks by a significant margin and performs as well as SpatialNet.

Table 1. Comparison of KeyPoint detection accuracies(of various architectures namely Coordinate Net (CN) [16], Coordinate extended (CNE) [16], Spatial net [17] and SH-PAF network on ASIIR dataset

Dataset	other architecture			
	SHDL	CN	CNE	Spatial Net
ASIIR	87.6	79.6	80.1	83

4.5. Suspicious Individuals Identification

The estimated body jointed are connected together to form a human skeleton structure as shown in Fig. 3. A set of 16 frames containing these skeleton structures are given as input to the 3D ResNext which performs the binary (suspicious vs non-suspicious) classification with an accuracy of 88.7% as shown in Table. 2. The proposed method is also able to outperform the state of the art methods by a decent margin.

Table 2. Table shows the suspicious activity classification accuracy (%) compared against the two state-of-the-art method.

	Comparison		
	ASANA	Surya et al. [15]	Singh et al. [24]
Acc.	84.8	69.2	76.8

The classification accuracy for suspicious activities recorded at different heights is also shown in Table 3.

As the number of humans increases in the image frame, the accuracy of the proposed system decreases as often the humans which are large distances are not detected.

4.6. Suspicious Individual Re-identification

This section presents the violent person Re-identification performance for rank 1 and 5 on the Aerial Suspicious person dataset using the Deep ScatterNet with cosine similarity

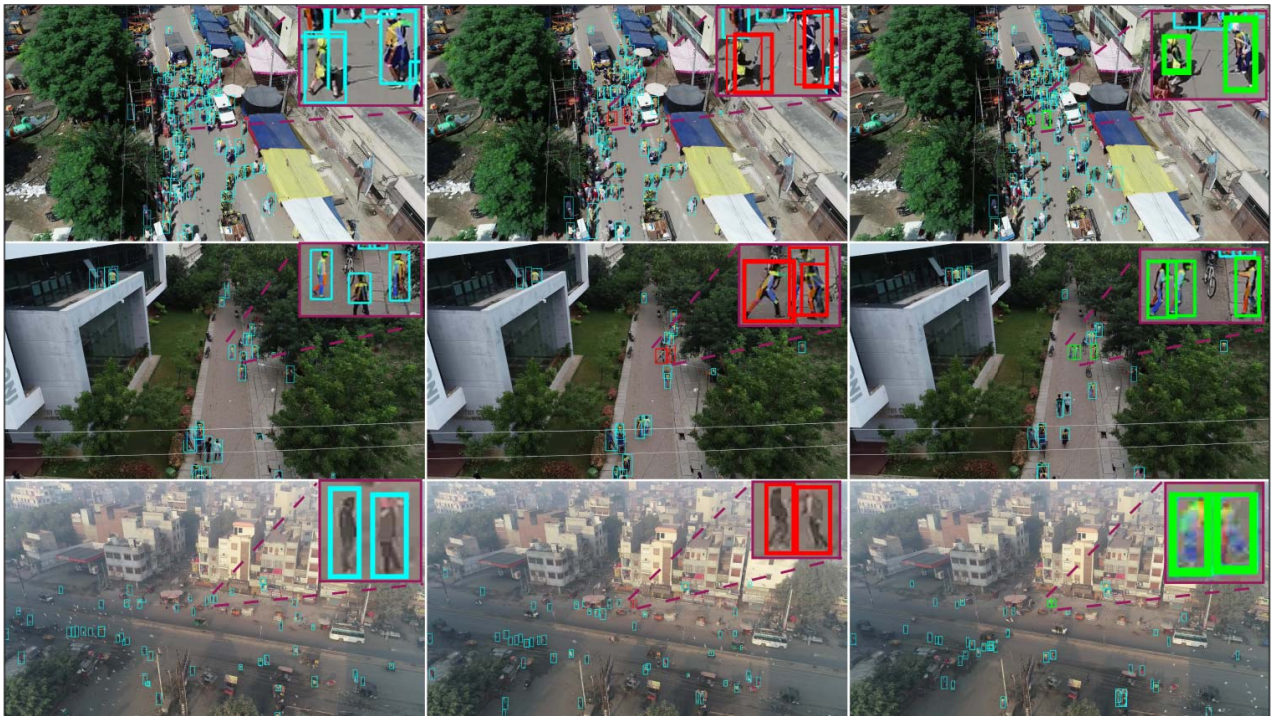

Figure 8. The figure represents three different actions namely- sword fighting, kicking and punching. Images in the first column show the situation before the actual event (in green), the second column presents the suspicious individuals (engaged in suspicious activities) detected by the software (in red), the third column presents the suspicious individuals re-identified again by the software.

Table 3. The table presents the classification accuracies(%) with the increase in height (m) for individuals engaged in suspicious activities in the ASIIR dataset.

	Height (m)			
	50	100	150	200
ASANA	94.1	90.6	87.3	84.8

and its comparison with other methods, shown in Table. 4.

Table 4. Comparison of ScatterNet and low-level features using recognition rate (%) at different ranks on the ASIIR dataset

Rank	Other techniques			
	Scat	VCues [13]	DVR [26]	HoG [12]
rank=1	56	32	38	28
rank=5	81	51	62	48
Average rank	83	57	57	n

4.7. Runtime Performance

The ASANA framework consisted of two parts:: (i) human pose estimation using the SH-PAF network, and (iii) classification of the estimated pose as suspicious or non-suspicious using the Bayesian 3D ResNext. Its runtime performance is computed on the cloud. The framework was trained and evaluated using the cuDNN framework and NVIDIA Tesla GPUs. For an aerial image frame, the system detected suspicious individuals at 5 fps per second to 16 fps for a maximum of ten and a minimum of two people respectively. The processing time varies in accordance with the number of individuals in the image frame.

5. Conclusion

The paper proposed a real-time Aerial Suspicious Analysis (ASANA) framework that can identify one or more persons engaged in violent activity from aerial images and re-identify them later when they are in the drones field of view again in subsequent frames. The framework first uses the proposed SHDL network consisting of Joints Identification network to detect humans and estimate their pose. The estimated poses are used by the 3d ResNext to detect any suspicious poses. The proposed SHDL network uses ScatterNet features with structural priors initialization to achieve accelerated training using relatively fewer labelled examples. The use of fewer labelled examples is beneficial for this application since it is expensive to collect annotated examples. The paper also introduced the ASIIR dataset which can benefit other researchers aiming to use deep learning methods for aerial surveillance applications. The proposed framework outperforms the state-of-art techniques on the dataset. We believe the framework would be instrumental in detecting violent individuals in public areas or large gatherings.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *IEEE International Computer Vision and Pattern Recognition*, 2017. 2

[2] Z. Cao, T. SImon, S.-E. W. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017. 3, 4

[3] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 2015. 2, 5

[4] C.-H. Chuang, J.-W. Hsieh, L.-W. Tsai, S.-Y. Chen, and K.-C. Fan. Carried object detection using ratio histogram and its application to suspicious event analysis. *IEEE transactions on circuits and systems for video technology*, 2009. 1

[5] D. et al. Spatio-temporal channel correlation networks for action classification. In *Proceedings of the European Conference on Computer Vision*, 2018. 2, 5

[6] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster. Automatic camera and range sensor calibration using a single shot. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3936–3943, 2012. 5

[7] K. Goya, X. Zhang, K. Kitayama, and I. Nagayama. A method for automatic detection of crimes for public security by using motion analysis. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009. 1

[8] S. Jain, S. Gupta, and A. Singh. A novel method to improve model fitting for stock market prediction. *International Journal of Research in Business and Technology*, 3(1):78–83. 2

[9] V. Jeengar, S. Omkar, A. Singh, M. K. Yadav, and S. Keshri. A review comparison of wavelet and cosine image transforms. *International Journal of Image, Graphics and Signal Processing*, 4(11):16, 2012. 3

[10] R. Layne, T. Hospedales, and S. Gong. Investigating open-world person re-identification using a drone. In *Proceedings of the European Conference on Computer Vision*, pages 225–240, 2014. 1

[11] X. Li, C. Zhang, and D. Zhang. Abandoned objects detection using double illumination invariant foreground masks. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 436–439, 2010. 1

[12] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999. 8

[13] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. In *arXiv:1611.08512*, 2016. 8

[14] T. Padgett. Drones join the war against drugs. *Time Magazine, June*, 2009. 1

[15] S. Penmetsa, F. Minhuj, A. Singh, and S. Omkar. Autonomous uav for suspicious action detection using pictorial human pose estimation and classification. *ELCVIA: electronic letters on computer vision and image analysis*, 13(1):18–32, 2014. 2, 7

[16] T. Pfister. Advancing human pose and gesture recognition. In *University of Oxford*, 2015. 7

[17] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision*, 2015. 7

[18] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240, 2013. 4

[19] A. Singh, D. Hazarika, and A. Bhattacharya. Texture and structure incorporated scatternet hybrid deep learning network (ts-shdl) for brain matter segmentation. *International Conference on Computer Vision Workshop*, 2017. 2, 3

[20] A. Singh and N. Kingsbury. Dual-tree wavelet scattering network with parametric log transformation for object classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 2, 3, 4, 5

[21] A. Singh and N. Kingsbury. Efficient convolutional network learning using parametric log based dual-tree wavelet scatternet. *IEEE International Conference on Computer Vision Workshop*, 2017. 3

[22] A. Singh and N. Kingsbury. Scatternet hybrid deep learning (shdl) network for object classification. *International Workshop on Machine Learning for Signal Processing*, 2017. 3

[23] A. Singh and N. Kingsbury. Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018. 3

[24] A. Singh, D. Patil, and S. Omkar. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1629–1637, 2018. 2, 7

[25] W. Walters and J. Weber. Ucav surveillance, high-tech masculinities and oriental others. *presentation to A Global Surveillance Society*, 2010. 1

[26] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Proceedings of the IEEE European Conference on Computer Vision*, 2014. 8

[27] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 4