Deep Learning in Gait Recognition for Drone Surveillance System

Jonathan Phang Then Sien*, King Hann Lim, Pek-Ing Au
CDT 250, 98009 Miri, Sarawak, Malaysia

*Email: incrediblej8@gmail.com

Abstract. As implementation of video surveillance system becomes mainstream, tremendous amount of data is produced. A robust and efficient recognition measure is required for enforcement of public safety in smart city projects. One of a demanded bio-metric recognition measures is gait recognition. With gait recognition, automatic indexing of individual’s action in surveillance area allow identification of abnormal activities, hence keeping the area in order. Besides, with advancement in online computing, gait recognition is enabled in remote computing that promotes hybrid of static-mobile surveillance system. Essentially in this research project, a deep learning pipeline to detect suspicious human actions is proposed to recognize individual human’s gait via untrimmed continuous streaming video. It consists of Single Shot Multi-box Detector (SSD) for human detection, Inception-V3 based transferred learning convolutional neural networks (CNNs) to extract spatial features. Subsequently, the spatial features are integrated with temporal information by sequentially inserted into Long Short-Term Memory (LSTM) deep architecture for action recognition. The pipeline is trained with KTH human action dataset consists of six action classes, i.e. walking, running, hand clapping, hand waving and boxing. The proposed pipeline achieved 99.51% detection rate to spot the suspicious gait action for surveillance system.

1. Introduction
Video surveillance systems are deployed in the modern city by authorities to monitor daily activities for security purposes [1]. The current design of surveillance system is always aimed to be affordable, reliable and portable in the market. Hence, there are two major types of video surveillance systems, i.e. static and mobile systems. Static surveillance system has installed bulk number of static cameras on the wall to cover the limited sight of viewing and therefore the cost of installation and maintenance is significantly increased. For the mobile systems, robotic machines and drones with on-board cameras are often deployed to increase the coverage of security with limited battery life for power consumption. As a proposed solution, the hybrid of static and mobile approach in the surveillance system is novel in the market to improve the viewing coverage, battery life and installation costs through cloud computing service with the advancement of online machine learning platform.

Online machine learning platform is made possible to be scaled up exponentially, to solve more complex problems in the video surveillance system. It could adapt and learn the feed-in videos to detect and recognize suspicious human activity from gait analysis. Gait is a manner or pattern referred to human’s locomotion achieved through movement of human limbs. In digital context, gait can be described as sequence of human’s locomotion that contain spatial-temporal features [2][3]. With visual-based gait recognition using deep learning, the identification and classification can be done in distances without any cooperation and minimal effort is required [4]. Deep learning neural networks (DNNs), such as convolutional neural networks (CNNs [5][6]) can classify and recognizing objects and often beyond human’s ability. Inspired from biological processes, CNNs require minimal of pre-processing to extract human features and subsequently, those human features can be trained from the samples of sequence. These human features can be stacked up to be a series of time variant features for periodic and non-periodic actions classification. For the classification, recurrent neural networks (RNNs) acquire
long short-term memory (LSTM [7]) to learn about time variant features and correlation between features from a sequence of action. Subsequently, the actions are recognized by the machine.

In this project, a deep learning approach for gait recognition is proposed to enable enhanced video surveillance system such as hybrid static-mobile drone surveillance system. The concept of the proposed pipeline and system is first implemented in local and offline network for model training and testing (see Figure 1(a)). In advance of the project, Google Cloud Platform (GCP) is introduced for cloud computing, storage and remote access as shown in Figure 1(b). Combination of modules in GCP is designed to work similarly as the offline pipeline model to detect and recognize abnormal human action. When there are incoming alerting signals from the GCP, the drone is activated to track the abnormalities in a safe range. Close-up monitoring on a suspicious human enhances the effectiveness of surveillance system.

2. Related work
Sequential approach is one of category in single-layered approaches which are mainly implemented for action recognition. Action recognition by sequential approach is carried out through analyzing of the sequential features that is relating the feature contained in each frame in a video. Therefore, a feature vector of sequence of image is extracted prior to action recognition. Then, the feature vector is analyzed by learning the correlation between the features and the correlation is used for action recognition. In general, there are two types of sequential approach in human activity recognition, i.e. exemplar-based approaches and state-based approaches.

Exemplar approaches work by comparing a sequence template of human activity to the input video frames for deviation. Afterwards, the activity can be recognized when the input frame has small deviation to the sequence template. This approach also give advantage for recognition of identical activity with variation of pose and rates. A dynamic time warping (DTW) algorithm is developed by Darrell et al. [8] in conjunction to exemplar-based approach, which can address optimal matching of two nonlinear sequence with variations using polynomial computations. In the works of Efros et al. [4], a visual-base action recognition is proposed. During the time of the methodology proposal, the resolution of input image is not ideal resulting the human height was on mere 30 pixels tall. To cope with this, two dimensional optical flows of the human is obtained by calculation of space-time volume of human using a temporal difference in the video frames. Subsequently, conversion of the two-dimensional optical flows to a spatial-temporal motion descriptor using blurry motion channels. Lastly, k-nearest method is applied for classification of sequence using the spatial-temporal motion descriptor.

State-based model approaches mathematically represent human action as set of state system. The state system is designed to exploit spatial-temporal feature between spatial feature of a video sequence. Therefore, the state system is trained statistically to output spatial-temporal feature based on the input.
spatial feature. In more specific, a statistical trained state model is designed to generate a probabilistic sequence (spatial-temporal feature). The probabilistic sequence is generated by calculating the deviation of spatial feature of video frames between action model. State model approaches such as Hidden Markov model (HMM) and Dynamic Bayesian Network (DBN) have been widely implemented for human activity recognition. Similarly, HMM and DBN are interpreting human activity as a set of hidden states. It is proposed that human activity in each frame of a video represent a state, hence the probability between transition between states are produced. These probabilities are corresponding to the human activity, therefore by threshold the output probability of the state model, human activity can be recognized such as work by Yamato et al.’s work [9]. In the most recent work, state model in sequential approaches are applied in more efficient deep learning model. Works presented by Montes et al. [10] demonstrated application of using CNN for spatial feature extraction. Similar in DMM, the output probability from DMM, also known as spatial-temporal feature is generated by RNN. The methodology demonstrated is able to be applied in today’s high-resolution input videos. Furthermore, Li et al. [11] demonstrated the integration of modern optical flows in Montes et al.’s work to further improve the recognition rate.

3. Human action detection and recognition

Our goal in implementation of action recognition for video surveillance-based system is capability of performing localization and local action recognition. The proposed pipeline is inspired by recent approach by Montes et al. [10], with a tweak by integrating human detection prior to action recognition. This method enables local action recognition and more importantly human localization which in contrast previous approaches only offer global action recognition. Figure 2 shows the proposed pipeline of human detection and recognition using deep learning approach. The pipeline comprises of human detection using Single Shot Multi-Box Detector (SSD) approach and gait analysis using Long-Short Term Memory (LSTM) deep learning model. In order to recognize gait, both spatial and temporal feature needs to be extracted. Hence, the novel idea of the proposed pipeline is local action recognition based on spatial-temporal features extracted from local detected region by SSD in the video using combination of CNNs and LSTM network as shown in Figure 3.

![Figure 2. Suspicious human action detection and recognition.](image)

3.1 Human detection using Single Shot Multi-box Detector

CNNs have evolved over the years, with higher detection rate, precision and performance. One of the CNNs that has achieved near state-of-the-art performance, while keeping cost of processing low is Single Shot Multi-Box Detector (SSD) [12]. Due to optimization and minimization to the hidden layer from state-of-the-art Faster Regional-CNN (FR-CNN) [13], SSD is capable of doing more detection per instant for several times. Hence, the cost of computational is made lower while having higher frames

| Method     | mAP (%) | FPS | Boxes |
|------------|---------|-----|-------|
| FR-CNN     | 73.2    | 7   | ~6000 |
| SSD        | 74.3    | 46  | 8732  |

Table 1. Comparison of FR-CNN and SSD on Pascal VOC2007 dataset [12]
per second (FPS) in object detection and high mean average precision (mAP) as tabulated in Table 1. With high detection rate and speed from SSD, it is ideal in the network for object detection. Pre-trained model from Common Contextual (COCO) dataset is used in the human detection. Detected humans are cropped and extracted for further processing.

3.2 Spatial-temporal feature generation using LSTM

In order to construct spatial feature from the cropped video frames, a pre-trained model is acquired from ImageNet dataset, based on Inception-V3 CNN for number of reasons. In Figure 4(a) and Figure 4(b), Inception-V3 network is reasonably accurate in doing detection and relatively low operations required for a single forward pass in the network.

Therefore, Inception-V3 network is ideal for features extracting feature. As referring to Inception-V3 network architecture in Table 2, input frame is up sampled to match the input of Inception-V3 network and 2048-feature data points of input frame can be obtained by tapping the last pooling layer of network. Softmax layer is removed in this project for features extraction. By appending the 2048-feature data points by number of sequence (n) for \( n > 1 \), spatial feature can be constructed via the moving object in the frame. Spatial feature of a moving human is not sufficient for gait classifying, as spatial feature represents only the location and space of moving human [15]. Chronological relation or temporal features of spatial features are needed to fully understand the gait activity, hence spatial-

![Figure 3](image3.png)

**Figure 3.** Pipeline model of abnormal human action detection and recognition.

![Figure 4](image4.png)

(a) (b)

**Figure 4.** Performance of various network: (a) Top1 vs. network. Single-crop top-1 validation accuracy for top scoring single-model architectures. (b) Top1 vs. operations, size \( \propto \) parameters. Top-1 one-crop accuracy versus amount of operations required for a single forward pass [18].

**Table 2.** Outline of Inception-V3 network architecture [14]
temporal features are important in order to recognize human action [16][17]. Therefore, LSTM recurrent neural network is designed to handle a sequence of spatial features and return an output of spatial-temporal features. It is trained with dropout probability $p = 0.5$.

### 3.3 Post processing

Two weighted fully connected layers with ReLu and softmax activation are used to classify the sequence of spatial-temporal features as shown in Figure 5. The pooling layer are used as a filter to give higher level of spatial-temporal features. These higher levels of spatial-temporal features can ease the computation of probability of each class in the softmax layer. The number of nodes and input size of each layer of proposed network is shown in Table 3. The probability of each class of gait is predicted for each successive stacked in $n$ frames, where $n$ is the number of sequence of frames. In this project, the pipeline is trained with 6 classes of action, where the 5 classes are periodic action, i.e. walking, running, hand clapping, hand waving and boxing, while the 6th class is assigned as non-periodic action.

**Figure 5. LSTM network for temporal feature detection and classification.**

**Table 3. Outline of proposed LSTM network**

| Type       | Number of nodes | Input size |
|------------|-----------------|------------|
| Conv       |                 | 3×3/2 299×299×3 |
| Conv       |                 | 3×3/1 149×149×32 |
| Conv padded|                 | 3×3/1 147×147×32 |
| Pool       |                 | 3×3/1 147×147×64 |
| Conv       |                 | 3×3/2 71×71×80 |
| Conv       |                 | 3×3/1 35×35×288 |
| 3×Inception|                 | 35×35×288 |
| 5×Inception|                 | 17×17×768 |
| 2×Inception|                 | 8×8×1280 |
| Pool       |                 | 8×8×2048 |
| Linear     | Logits          | 1×1×2048 |
| Softmax    | Classifier      | 1×1×1000 |
4. Implementation of Google Cloud Platform (GCP)

The pipeline of human action detection and recognition is developed on the local computer in offline network. This pipeline is converted to native GCP-compatible model and implemented in GCP. With GCP, the feasibility and performance of implementation can be enhanced for remote, scalability and real-time application. In this project, the proposed algorithms are migrated from the backend computer into GCP. In order to fully migrate the proposed pipeline and system, Cloud Storage, Cloud function, Cloud Vision, Compute Engine and App Engine as shown in Figure 6 are used in the project. The input source of video frames is a camera-mounted quadcopter connected to a stable internet via wireless connection. The camera with a microcontroller onboard is programmed to upload video frames to incoming-frames’ bucket in the Cloud Storage. By utilizing Cloud Function, it is programmed to be triggered as video frames are buffered in bucket, Cloud Function will invoke Cloud Vision API to start human detection from the buffered frames which is inspired by SSD human detector in the pipeline. After detections are made, annotation JSONs from the detection is stored and logged in detection’s bucket in Cloud Storage, while also being send to Compute Engine. In Compute Engine, KERAS open source deep learning library is installed with Google’s TensorFlow as backend.

Model of the proposed pipeline which is trained on the offline computer is loaded to Compute Engine to enable the spatial-temporal feature process. Next, Compute Engine will pull buffered frames from bucket, and based on annotation JSONs acquired from Cloud Vision, gait recognition can be carried out and annotation is stored in recognition bucket. Finally, App Engine is used to build a mobile app, to enable access of data from buckets in Cloud Storage, such as getting real-time monitoring status and logged activities detection remotely.

5. Result and discussion

In the experiment, the proposed pipeline was built on KERAS library and it was implemented on Google’s Tensorflow as backend. Training dataset in the network was retrieved from KTH activity [19]. KTH dataset consists of six type of human actions, i.e. walking, jogging, running, boxing, hand-waving and hand clapping. There were 25 subjects performing actions several times in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors conditions. There were 2391 sequences in the dataset and all sequences were taken over homogenous background with a
static camera with 25 frame-per-second. In this research, static action. The video sample are in untrimmed 160×120 resolution video format. The dataset was split into 70% training and 30% for validation. In the proposed system, spatial feature network was implemented with pre-trained model and subsequently model was evaluated using categorical accuracy metric, where the index of maximal predicted value from softmax layer is compared to index of true prediction.

Several data augmentation was made to simulate various possible scenario during online gait recognition. In the first scenario where the number of sequences of frame was fixed at \( n = 5 \), videos from dataset was frequency down sampled with factor of \( f = 2, 5, 10 \) to identify the best resolution for efficient pipeline detection and classification. In the second scenario, the number of sequences was determined with \( n = 12, 24 \) and stride is set to 1 for non-down sampled videos. This experiment was to identify the minimum sequence of number used in defining an action of human. Video that was down sampled by factor, \( f = 2 \) is ideal for idle processing as it required less computational power while having reasonably high accuracy. The higher accuracy and lower loss were caused by the down sampling effect due to the frames were skipped in interval of 2. Hence, every sequence will contain more significant spatial feature than non-down sampled video. However, further down sampling the video caused information lost and resulting lower accuracy and higher loss as tabulated in Table 4. In the lossless processing, where video was not down sampled, while having longer number of sequence and each sequence is apart from another sequence by stride of 1.

In Figure 7(a), it can be observed that a human object required averagely 24 frames to complete one full cycle of walking gait. Due to walking gait is the dominant in most of the recognition scenario, the number of sequences is set \( n = 12, 24 \), where \( n = 12 \) was the half cycle of walking gait and \( n = 24 \) was the full cycle of walking gait for all the recognition. As illustrated in Figure 7(b), half cycle of human gait is defined when opposite heel strike is achieved, and same sided heel strike is achieved in full cycle. Simulated results were expected to detect human actions when lossless information is given,

![Figure 7](image.png)

**Figure 7.** Human gait cycle. (a) Illustration of walking gait cycle. (b) Number of frames in conjunction to illustrated walking gait cycle.
the proposed network can recognise up to 99.51% with the defined number of sequences based on the walking gait as shown in Table 5. Despite the results of the gait recognition on test KTH dataset, the proposed pipeline was implemented on actual real-time streaming video as shown in Figure 8. Actions such as walking, boxing and hand-waving was detected in the video with 10 fps. It is notably to mention as the pipeline is currently solely trained with KTH dataset, which the properties of the dataset are grey-scaled video and low resolution, the recognition on actual real-time video yielded remarkable results. In the project, the definition of suspicious human action is categorized as non-periodic non-repetitive action (6th class). The detected suspicious action is used to enable the mobile surveillance system into a drone to monitor the person in a closer manner.

**Table 4.** Down sampling factor per training epoch, accuracy and loss for n = 5.

| Down sampling factor, f | Number of epochs | Accuracy  | Loss   |
|-------------------------|------------------|-----------|--------|
| 0                       | 104              | 93.95%    | 0.1541 |
| 2                       | 116              | 96.96%    | 0.0818 |
| 5                       | 116              | 95.68%    | 0.1581 |
| 10                      | 52               | 82.40%    | 0.4577 |

**Table 5.** Number of sequences per training epoch, accuracy and loss for original videos.

| Number of sequences, n | Number of epochs | Accuracy  | Loss   |
|------------------------|------------------|-----------|--------|
| 12                     | 83               | 99.46%    | 0.01576|
| 24                     | 56               | 99.51%    | 0.0143 |
6. Conclusion
A human action detection and recognition pipeline is proposed in this project for hybrid static-mobile video surveillance system. First, spatial feature is extracted from cropped human by consecutively stacking successive detected human by SSD and feature extractor CNN-based network. Recurrent neural network with LSTM receives the spatial features to generate spatial-temporal feature that will be used for human action classification. The experimental results showed the proposed pipeline can achieve 99.51% of accuracy in gait recognition on KTH dataset. Degradation of performance is observed with various data augmentation, such as down sampling and different number of sequences. Down-sampled video by factor of 2 can be used for idle processing as the computational cost is twice lower from non-down sampled video and has high recognition rate. Despite the experimental results, the proposed pipeline is able to achieve remarkable result on actual streaming data. However, the recognition is extremely sensitive to camera angle, where if the angle of human in the video is completely novel prior to trainings, the recognition will fail. It is also worth noting that any false positive during detection phase will also bring recognition failure throughout the whole pipeline. Lastly, the proposed pipeline is built on top of KERAS library, which run on Google’s Tensorflow backend. Hence, it can be converted into GCP compatible model and allow migration to GCP and it can tremendously reduce the size of the hardware on the drone while enabling the remote functionality and increase the feasibility of the system.

References
[1] PRO-VIGIL, INC., “Understanding the Benefits of Security and Surveillance,” 4 May 2016. Available: https://pro-vigil.com/2016/05/04/understanding-the-benefits-of-security-and-surveillance/
[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks, 2014.
[3] Z. Shou, D. Wang, and S. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In CVPR, 2016
[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in IEEE International Conference on Computer Vision, IEEE, 2003, p. 726.
[5] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations, 2015.
[6] C. Szegedy, W. Liu, and Y. Jia. Going deeper with convolutions. 1409.4842, 2014.
[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, pp. 1735-1780, 1997.
[8] T. Darrell and A. Pentland, “Space-time gestures,” in Computer Vision and Pattern Recognition, 1993. Proceedings CVPR’93., 1993 IEEE Computer Society Conference on, IEEE, 1993, pp. 335–340.
[9] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time sequential images using hidden markov model,” in Computer Vision and Pattern Recognition, 1992. Proceedings CVPR’92., 1992 IEEE Computer Society Conference on, IEEE, 1992, pp. 379–385.
[10] A. Montes, A. Salvador, S. Pascual, and X. Giro-i-Nieto, “Temporal activity detection in untrimmed videos with recurrent neural networks,” arXiv preprint arXiv:1608.08128, 2016.
[11] Z. Li, K. Gavrilivuk, E. Gavves, M. Jain, and C. G. Snoek, “Video LSTM convolves, attends and flows for action recognition,” Computer Vision and Image Understanding, vol. 166, pp. 41–50, 2018.
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, “SSD: Single shot multibox detector,” in European conference on computer vision, 2016, pp. 21-37.
[13] R. Girshick, “Fast R-cnn,” arXiv preprint arXiv:1504.08083, 2015.
[14] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens. Rethinking the Inception Architecture for Computer Vision, 2015.
[15] I. Laptev, “Local spatio-temporal image features for motion interpretation,” PhD thesis, Numerisk analys och datalogi, 2004.
[16] T. Darrell and A. Pentland, “Space-time gestures,” in Computer Vision and Pattern Recognition, 1993. Proceedings CVPR’93., 1993 IEEE Computer Society Conference on, IEEE, 1993, pp. 335–340.

[17] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, “Modeling spatial-temporal clues in a hybrid deep learning framework for video classification,” in Proceedings of the 23rd ACM international conference on Multimedia, ACM, 2015, pp. 461–470.

[18] A. Canziani, A. Paszke, and E. Culurciello, “An analysis of deep neural network models for practical applications,” arXiv preprint arXiv:1605.07678, 2016.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, Aug 2004.