2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)

Semantic segmentation of UAV aerial videos using convolutional neural networks

Girisha S
Department of Information and Communication Technology
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
girisha3893@gmail.com

Manohara Pai M M
Department of Information and Communication Technology
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
mmm.pai@manipal.edu

Ujjwal Verma
Department of Electronics and Communication Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
ujjwal.verma@manipal.edu

Radhika M Pai
Department of Information and Communication Technology
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
radhika.pai@manipal.edu

***Abstract*—Semantic segmentation of complex aerial videos enables a better understanding of scene and context. This enhances the performance of automated video processing techniques like anomaly detection, object detection, event detection and other applications. But, there is a limited study of semantic segmentation in aerial videos due to non-availability of the relevant dataset. To address this, an aerial video dataset is captured using DJI Phantom 3 professional drone and is annotated manually. In addition, the proposed research work investigates the performance of semantic segmentation algorithms for aerial videos implemented using Fully Convolution Networks (FCN) and U-net architectures. In this study, two classes (greenery, road) are considered for semantic segmentation. It is observed that both architectures perform competitively on the aerial videos of Unmanned Aerial Vehicle (UAV) with a pixel accuracy of 89.7% and 87.31% respectively.**

***Index Terms*—Semantic segmentation, aerial videos, convolutional neural network, U-net, FCN**

I. Introduction

Analyzing videos captured through UAV have wide applications like tracking vehicles, object detection, anomaly detection etc. For a majority of these applications, there is a need to infer spatial and contextual information from these images. For example, tracking of vehicles will be easier in the presence of knowledge about the roads. Semantic segmentation is one of the tools used to divide the image into different semantic regions and classify these regions into predefined classes. Semantic segmentation helps in understanding the layout of the scene and hence it is increasingly becoming a vital factor in anomaly detection, autonomous driving vehicles, object detection, etc. [1]. Semantic segmentation remains challenging because of variation within a class, loss of perspective, the context of the scene, the presence of noise, illumination changes etc. Semantic segmentation can be achieved by using traditional machine learning approach like Conditional Random Field (CRF) and deep learning approach based on Convolutional Neural Network (CNN).

CRF based algorithms are widely used because of their ability to capture context information. This framework generally consists of unary potential and pairwise potential energies. Unary potential energy captures local features which are dependent on pixel itself while pairwise potential energy captures spatial information. Different potential energies capturing various features like texture, colour location etc. needs to be manually encoded into the model. However, these hand-crafted features may fail to capture all the variations in the data.

Recently, a modification of multi-layer perceptron called as CNN gained great success in semantic segmentation, object detection and image classification [2], [3], [4], [5]. This is due to the fact that CNN captures context information which plays an important role in these tasks. The accuracy of segmentation depends on local features (colour, intensity etc.) and global features like (texture, context etc.). The ability of CNN to learn both these features in an end to end style has led to its success in semantic segmentation. Hence a deep learning approach is preferred for semantic segmentation because they are dependent on learned features.

The success of automated systems like anomaly detection, event detection etc. in aerial videos relies greatly on scene understanding for greater accuracy. However, there is a limited study on semantic segmentation of UAV videos due to the lack of available dataset. The aim of this study is to analyze the performance of CNN based semantic segmentation algorithms on aerial videos. Also, in the present paper, UAV aerial video dataset is presented for semantic segmentation. The proposed aerial video dataset consists of various scenarios taken from different regions along with annotations. Two widely used CNN based semantic segmentation algorithms,

978-1-7281-1488-0/19/$31.00 ©2019 IEEE
DOI 10.1109/AIKE.2019.00012

Fully Convolution Network (FCN) and U-Net, are used to evaluate the performance of semantic segmentation on aerial images.

The remaining part of the paper is organized into following sections. Section 2 deals with the recent developments in semantic segmentation of aerial videos. Section 3 describes the detailed methodology of the proposed system followed by results and discussions briefed in Section 4. Finally, conclusions are given in Section 5.

II. Related Work

Semantic segmentation divides the image into regions and labels each region with a predefined class label. Identifying the layout of the scene provides information about the spatial distribution of the object and their relationship with the environment. A brief overview of semantic segmentation is presented below. A detailed review of semantic segmentation of images can be found in [1].

The importance of texture, context and spatial information of pixel in semantic segmentation is explained in [6]. Here neighbouring pixels are used for labelling the pixel in consideration which will help in segmentation accuracy. Immense work has been carried out for semantic segmentation using CRF based approach since they are able to capture spatial information [6], [7], [8], [9]. CRF framework consists of unary potential described by local features and pairwise potential. This unary potential can be varied according to the necessity [10], [11]. Also, higher order potential energies are explored [12] in order to improve segmentation accuracy.

Recently, CNN's are widely explored for semantic segmentation due to their ability to learn complex features [13], [3], [14], [2]. Encoder and decoder-based architecture are more popular because they learn deep features which give more accurate results compared to handcrafted features [13], [3]. In a separate study, the convolutional layer is replaced by atrous convolutional operation [14]. This enables to capture more spatial information with a smaller number of parameters. There are several numbers of works wherein CRF and CNN frameworks are combined in order to improve the segmentation accuracy [2].

There are few works exploring semantic segmentation of Po1SAR aerial images using CNN [15]. These images are able to capture information from the vast area however, quality of image acquisition depends on the climate in Po1SAR images, therefore, limiting this application.

Aerial videos captured through UAV are generally used for detecting and tracking objects in a given scene [1]. Few aerial video datasets are developed for multiclass object tracking [16] and anomaly detection [17]. These datasets are provided along with annotations for object detection. To analyze the scenario more precisely, understanding the layout of the scene is essential which gives an insight into the context of the scenario. Hence a sophisticated aerial video dataset along with ground truth for semantic segmentation is need of the hour.

Due to the lack of aerial video dataset and annotation set for semantic segmentation a new aerial video dataset is proposed to be developed in this study which will be mainly used for semantic segmentation. U-Net and FCN algorithms are used for semantic segmentation of captured dataset. Finally, the system is evaluated by various performance metrics.

III. Dataset: Manipal UAV aerial video dataset

DJI Phantom 3 professional drone is used to capture aerial videos. The resolution of the captured videos is 1280 x 720 resolution at 29 frames per second. These videos are collected at an altitude of 25 meters approximately. The videos are collected inside Manipal Institute of Technology campus, Manipal, India. For the present study, two semantic classes are considered namely greenery and roads. Greenery class includes trees, gardens and foliage. Road class includes the parking lot, footpath and roads. The aerial videos are captured from 8 different regions covering various scenarios such as parking lot, library, etc. The maximum and minimum duration of the videos is about 12 minutes and 10 seconds respectively.

Processing all the frames is time-consuming and tedious. Hence keyframes are identified using the shot boundary detection algorithm. These keyframes are given as input to both the segmentation algorithms. Annotations for semantic segmentation are provided for keyframes using LabelMe tool [18]. Ground truth images are necessary for establishing the reliability of the system by stating the accuracy of prediction. But the creation of these ground truth images is a challenging task for aerial images since there exists ambiguity in labelling the pixels at the boundary of two classes of objects. For instance, the foliage of the trees appears to be indefinite from the top view (Figure 1). Few original frames and corresponding ground truth masks are shown in Figure 1. In the ground truth image, green colour represents the greenery class and grey colour represents the road class. Table I presents a brief description of the data set.

TABLE I
Description of Data set

Total number of videos	13
Minimum duration	10 sec
Maximum duration	12 min
Total number of frames	2494
Number of class considered for annotations	2 (road, greenery)
Image resolution	1280x720 pixels
Frames per second	29
Approximate altitude	20-30 mts

IV. Methodology

In the present study, the shot boundary detection algorithm is first used to identify the keyframes. Subsequently, semantic segmentation is performed on these keyframes by using FCN and U-Net.

A. Shot boundary detection

The captured UAV aerial video dataset has 29 fps. Hence the variation between each consecutive frame is minute. Therefore, keyframes are identified using the shot boundary

Fig. 1. Sample images from dataset and its annotation

detection algorithm to ease the analysis of frames. Figure 2 describes the process of keyframe identification where shot boundaries are identified from consecutive frames and the entire block is represented by the middle frame. This middle frame is the keyframe.

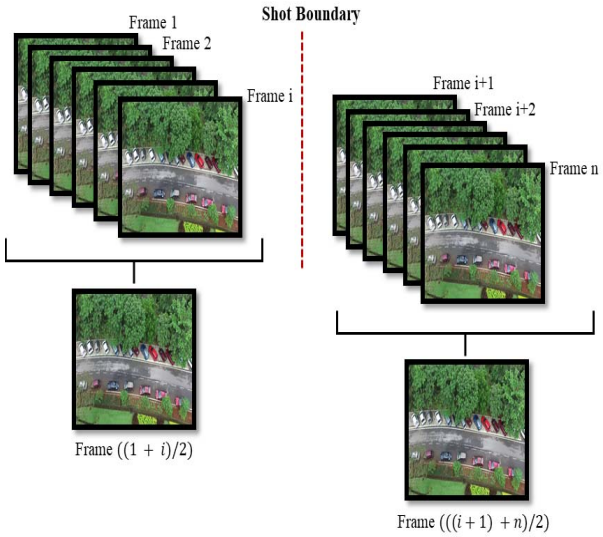

Fig. 2. Identification of Key frames

Shot boundaries are identified in each frame by dividing each frame into non-overlapping grids of size 16x16. Corresponding grid histogram difference is computed between two consecutive frames by adopting chi-square distance and is given as follows,

$$d(H_i, H_{i+1}) = \sum_i \frac{(H_i(I) - H_{i+1}(I))^2}{H_i(I)} \quad (1)$$

Where, H_i represents the histogram of i^{th} frame and H_{i+1} represents the histogram of $(i+1)^{th}$ frame. I represents the image patch at an identical location in both the frames. The average histogram difference between two consecutive frames is calculated as follows,

$$D = \frac{1}{N} \sum_{K=1}^{N} d_k(H_i, H_{i+1}) \quad (2)$$

Where, D represents the average histogram difference of two consecutive frames and d_k represents the chi-square difference between k^{th} image patches. N represents the total number of image patches in an image. Shot boundary is identified at frames where the histogram difference is greater than threshold T_{shot}.

$$ShotBoundary = \begin{cases} 1 & D_i - D_{i+1} > T_{shot} \\ 0 & otherwise \end{cases} \quad (3)$$

Where i and $i+1$ represents two consecutive frames. Key frames identified are further processed by using semantic segmentation algorithms to identify the various objects present in the scene (greenery, roads). U-Net and FCN models are considered in the present study to perform semantic segmentation because they don't require large datasets.

B. U-Net

U-Net model was originally proposed by [13] and has been used for various applications such as biomedical image segmentation [13], remote sensing [19], [20], [21], etc. This architecture mainly consists of two paths namely contracting path and symmetric expanding path. Region localization is done in expanding the path with the help of features extracted in the contracting path. Convolution operation followed by Rectified Linear Unit (ReLU) activation function is implemented in the contracting path for the purpose of extracting features. Among the extracted features the relevant ones are identified by applying the maxpool function. The model learns patterns by adopting data augmentation and gradient descent method. Softmax activation is used in the last layer of architecture to obtain the probability of the pixel belonging to each class.

The U-Net architecture proposed in [13] was used for grayscale images of size 572x572. In this study, U-Net architecture is modified accordingly to process aerial images. The network is modified to handle colour images (RGB) of size 256x256 instead of only grey scale images. This is achieved by using 3D convolutional operation at each layer. Along with maxpool operation, padding is also considered in each layer to retain the most relevant features for further processing. Figure 3 represents the modified U-net architecture for segmenting of aerial videos.

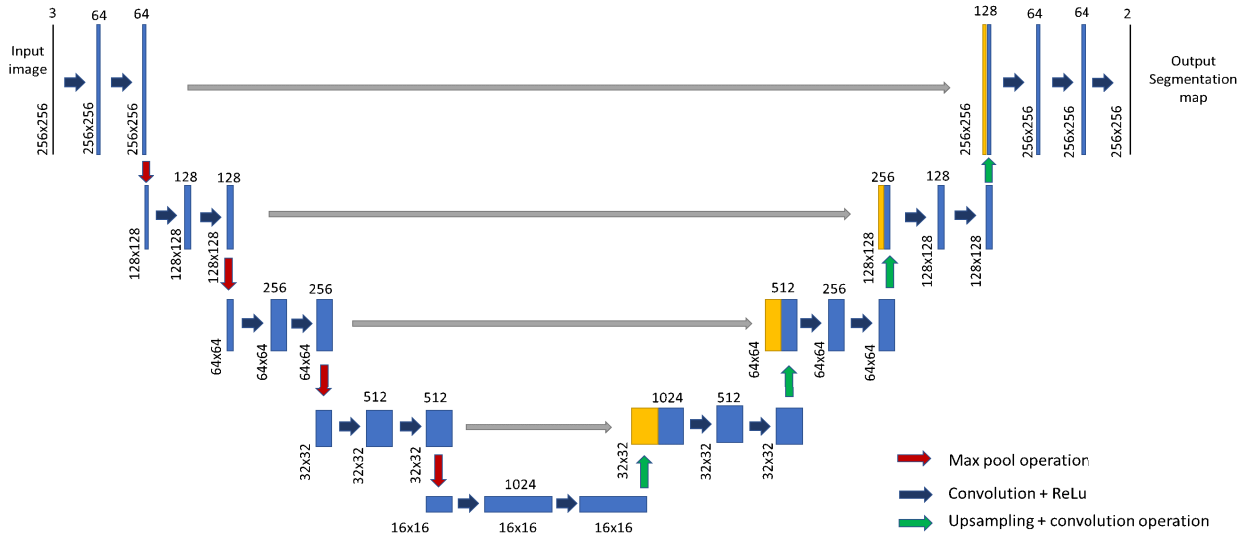

Fig. 3. Modified U-Net architecture for UAV aerial video semantic segmentation

C. FCN with VGG16 Backbone

Fully convolution networks proposed by authors in [3], are widely used for semantic segmentation because it can be developed by using other CNN layers like VGG16 [4], ResNet [14] etc. FCN uses CNN blocks of other architecture like [4], as backbone architecture to extract features. In the present study, FCN32 with VGG16 is incorporated as a backbone architecture. VGG16 architecture is used as a backbone because of its simple structure. The last dense layer of VGG16 is replaced by fully convolutional layers to obtain output segmentation map. In every layer, the convolution operation is followed by ReLU activation function. Maxpool operation is performed in order to preserve the most relevant features for later stages. The model is trained from scratch on the UAV aerial video dataset. The last layer is modified for binary class classification. Softmax operation is used in the last layer in order to obtain the probability of pixel belonging to each class. The architecture of FCN is shown in Figure 4.

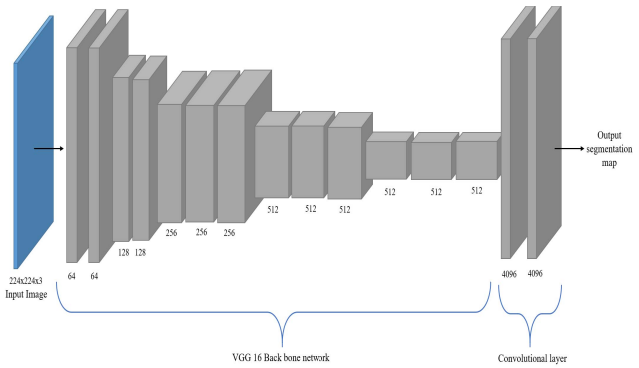

Fig. 4. FCN architecture [3]

V. Results and Discussion

In the present study, U-Net and FCN architectures are used to perform semantic segmentation on created UAV aerial video dataset. Results obtained by applying both the algorithms are presented henceforth. 80% (80 images) of the data captured is used for training the model, 10% (10 images) is used for the purpose of validation and remaining 10% (10 images) is used for testing. To evaluate the performance of both the algorithms, Mean Intersection over Union (MIoU), Pixel Accuracy (PA), precision, recall and F1-score methods are used. MIOU is calculated as follows.

$$MIoU = \frac{\sum_i x_{ii}}{C(\sum_i \sum_j x_{ij} + \sum_j x_{ij} - n_{ii})} \tag{4}$$

Where C is the number of classes which is two in this study. x_{ij} represents the number of pixels belonging to class i and predicted as class j. Pixel accuracy is calculated as follows.

$$PA = \frac{\sum_i x_{ii}}{\sum_i \sum_j x_{ij}} \tag{5}$$

A. Shot boundary detection

In the present study, the shot is identified by using shot boundary detection algorithm. Histogram difference between two consecutive frames are identified and is compared with the threshold T_{shot}. The value of T_{shot} is experimentally found to be 0.2.

B. U-Net

The U-Net model is trained from scratch on UAV aerial video dataset. Data augmentation and transfer learning are not necessary because sufficient training data is available in the

learning phase. Due to the memory constraints, the batch size is set to 10. Categorical cross entropy is used for obtaining loss function. The weights are initialized following the approach of [5]. The model loss and accuracy curve for training phase is obtained after it is trained for 100 epochs which are shown in Figure 5 and 6. The loss decreases nonlinearly as the number of epochs increases.

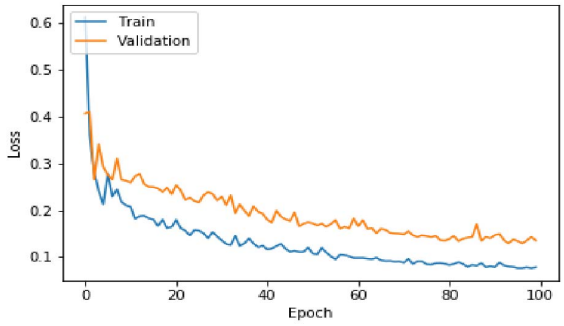

Fig. 5. Training and validation loss for U-Net architecture

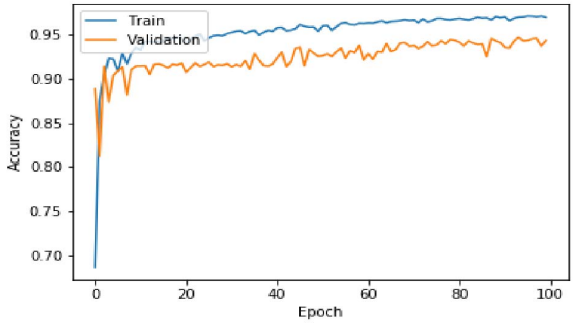

Fig. 6. Training and validation accuraccy for U-Net architecture

The segmentation result obtained from the U-net model on the aerial video dataset is shown in Figure 7. Segmentation accuracy measure is computed to establish reliability. The computed overall pixel accuracy, precision, recall, F1-score and MIoU of the system is given in Table II.

As seen from Figure 8 and 9, few false positives are present for greenery class which occur for the pixels covering the parking area. Presence of shadow is the root cause for this ambiguity since the model struggles to handle illumination changes. The sensitivity of the model to these illumination changes can be addressed by incorporating colour and texture features.

As observed there is the presence of false negatives for greenery class in some of the cases which are again due to the presence of the shadow. This effect is clearly observed in Figure 10 where greenery class objects are classified as road pixel. The same is marked by the yellow circle in the figure.

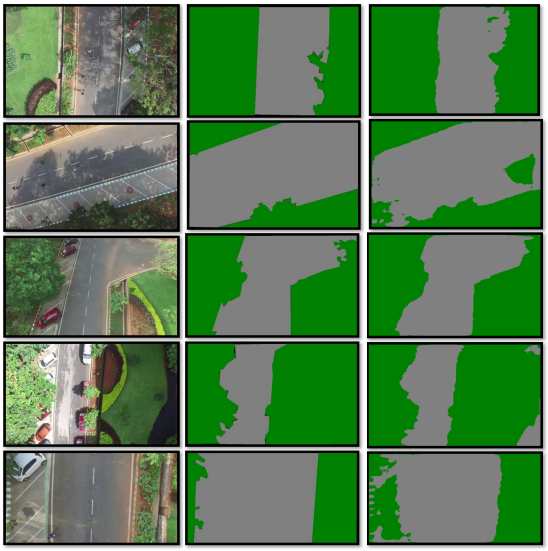

Fig. 7. Semantic segmentation results of U-Net. (a) Original images (b) Ground truth images (c) U-Net segmentation results

Fig. 8. Semantic segmentation results of U-net. (a) Original image (b) Ground truth image (c) Image showing false positives for U-Net model

Fig. 9. Semantic segmentation results of U-Net. (a) Original image (b) Ground truth image (c) Image showing false positives for U-Net model

Fig. 10. Semantic segmentation results of U-Net. (a) Original image (b) Ground truth image (c) Image showing false negatives for U-Net model

C. FCN

In the present study, VGG16 is used as a backbone architecture of FCN. The network weights are learnt by using the created UAV aerial video dataset. Transfer learning is not preferred in this case because of the availability of sufficient training data. The model adopted is trained for 100 epochs with a batch size of 10. The loss function used is categorical cross entropy and weights are initialized following the approach of [5]. The model training loss and accuracy curve obtained are shown in Figure 11 and 12.

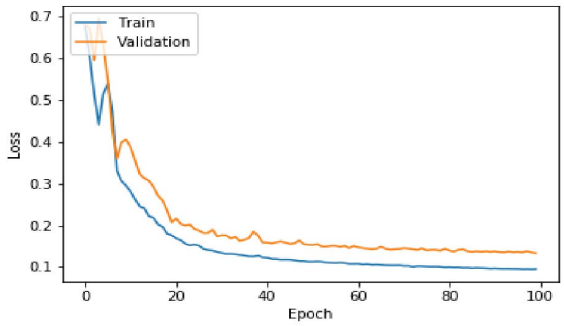

Fig. 11. Training and validation loss for FCN architecture

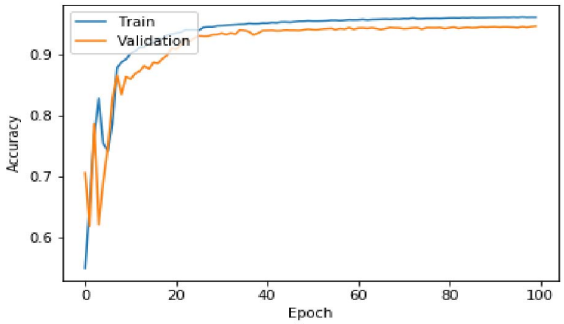

Fig. 12. Training and validation accuracy for FCN architecture

TABLE II
PERFORMANCE METRICS OF U-NET AND FCN MODEL

	Precision	Recall	F1-score	PA	MIoU
U-net	0.95	0.95	0.95	87.31%	0.911
FCN	0.96	0.96	0.96	89.7%	0.918

It is visible from the curves that the training loss decreases as the number of epochs increases. The pixel accuracy, precision, recall, F1-score and MIoU calculated for FCN is shown in Table II. The segmentation results of FCN on UAV aerial videos are shown in Figure 13. The performance of FCN architecture on segmentation is similar to the performance of U-Net architecture. However, it is seen that there is less occurrence of false positives for greenery class compared to that of U-Net architecture which indicates that, FCN model is able to handle illumination changes. It is also observed that MIoU of both the algorithms are similar. Pixel accuracy of FCN is 2.39% greater than U-Net architecture. In spite of ambiguities in class boundaries in the ground truth image, the U-Net and FCN models are able to segment the regions with high accuracy.

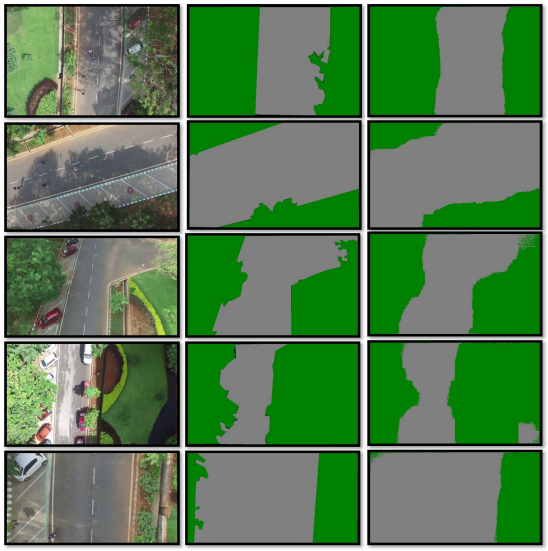

Fig. 13. Semantic segmentation results of FCN. (a) Original images (b) Ground truth images (c) FCN segmentation results

VI. CONCLUSION

To address the scarcity of annotated aerial videos for semantic segmentation a new UAV aerial video dataset is created for the same. These videos are collected from various regions which include different scenarios. The dataset is annotated manually for semantic segmentation into two major classes (road, greenery). U-Net and FCN are employed on developed aerial video dataset to achieve semantic segmentation. A comparative analysis is carried out between these two algorithms and the obtained results are presented. It is observed that CNN based algorithms like U-Net and FCN do not always require large dataset for learning the patterns. Accuracy difference of FCN and U-Net model is 2.7 per cent which indicates that both the algorithms perform competitively well for small dataset. In future, we would extend this dataset to cover more locations covering a wider area for multi-class classification. Further, the proposed dataset may be used for several applications like anomaly detection, event detection, object tracking etc.

REFERENCES

[1] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[6] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[7] N. Dinesh Reddy, P. Singhal, and K. Madhava Krishna, "Semantic motion segmentation using dense crf formulation," *arXiv preprint arXiv:1504.06587*, 2015.

[8] X. He and S. Gould, "An exemplar-based crf for multi-instance object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 296–303.

[9] B.-s. Kim, P. Kohli, and S. Savarese, "3d scene understanding by voxel-crf," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1425–1432.

[10] C. Russell, P. Kohli, P. H. Torr *et al.*, "Associative hierarchical crfs for object class image segmentation," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 739–746.

[11] S. Zheng, M.-M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, and P. H. Torr, "Dense semantic image segmentation with objects and attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3214–3221.

[12] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1698–1705.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[15] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, "Polsar image semantic segmentation based on deep transfer learning–realizing smooth classification with small training sets," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[16] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 4. IEEE, 2015, pp. 1–6.

[17] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[19] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[20] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 2115–2118.

[21] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, 2009.