IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Suspicious Activity Detection in Surveillance Video Using Fully Convolutional Networks Segmentation

S. Santhiya[1*], T. Ratha Jeyalakshmi[2]

[1, 2] Department of computer applications, Sri Sarada College for women,Tirunelveli -11.

DOI: https://doi.org/10.26438/ijcse/v7si8.139142 | Available online at: www.ijcseonline.org

Abstract— In Recent Years, suspicious activity detection is used to detect traffic in different surveillance videos with high accuracy and high speed in daytime. This surveillance video detection method includes Adaptive Background, Object Modeling, Object Tracking, Activity Recognition, and Segmentation. The semantic segmentation using suspicious activity detection techniques plays a major role in the segmentation of the surveillance video. U-Net is one of the popular Fully Convolutional Networks (FCN) which is applicable for image segmentation. This method could found the different anomalies activity from the videos.

Keywords— Segmentation, FCN, Object Modeling, Suspicious Activity Detection, Surveillance Video.

I. INTRODUCTION

Fully Convolutional Networks (FCN) unit is revolutionizing many fields of computer vision. This paper presents Suspicious Activity Detection in investigating video. Among Segmentation Techniques, FCN style is applied to tackle the matter of investigating video object segmentation, that is, the segmentation of all pixels of a video sequence into background and foreground, given the manual annotation of one (or more) of its frames. We have a tendency to stipulate that a completely convolutional network (FCN) is processed by end-to-end, pixels-to-pixels on video segmentation. Suspicious activity and anomaly activity detection play an important role in vehicle identification. Image data may well be a necessity to provide such identifications.
We used annotated vehicle knowledge set provided by Udacity. The 4.5 GB knowledge set was composed of frames collected from 2 of videos whereas driving the Udacity automobile around Mountain read space in serious traffic. The info set contained a label file with bounding boxes marking different cars, trucks and pedestrians. The complete knowledge set is comprised of about 22000 pictures. we have a tendency to combine cars and trucks into one category vehicle, and born all the bounding boxes for pedestrians.

Data preparation and augmentation:
We first divided information the info the knowledge the knowledge into work and testing data sets. as a result of the frames were obtained from a video feed, each frame was dependent upon the previous frames, we've an inclination to thus last 2000 footage for testing, and remaining footage for work. We have an inclination to then performed augmentation on work info set. We've an inclination to performed solely three augmentations during this paper.

Fig 1 Original and Segmentation image

Stretching:
Figure below shows however stretching augmentation works. We have a tendency to 1st outline four points close to corners of the first image (shown in purple). We have a tendency to then stretch these points thus these points become the new boundary points.

Translation:
We next apply translation transformation, to model the impact of automotive moving at completely different locations.

Target set preparation: In typical pixel-wise prediction, we tend to draw polygons round the object of interest to draw masks. During this case, we tend to failed to have that info, we tend to thus use the region among the bounding boxes as masks for outlining objects. We tend to then use these masks to get a mask of constant size that once applied to the

first pictures provides U.S.A. vehicles back. This so often conjointly illustrated within the figure below. The picture son the left panel square measure obtained victimization augmentation on associate degree naive image, the middle panel presents the vehicle mask we tend to shall predict and therefore the final panel shows the results of applying the mask back on the first image to substantiate that the mask actually identifies vehicles. The goal of our neural network model is to predict the mask within the center. We next apply translation transformation, to model the effect of car moving at different locations

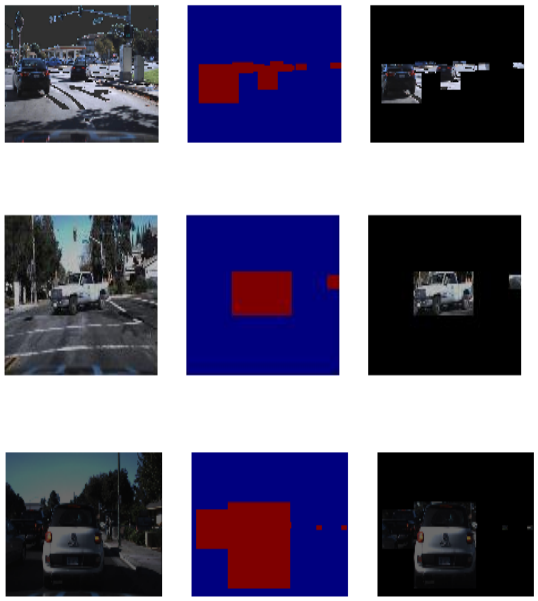

Fig 1.6 Augmented images, masks and applied masks generated from scaling, translation and brightness augmentation.

II. METHODOLOGY

The model we tend to selected could be a scaled down version of a deep learning design known as U-net. U-net could be a encoder-decoder kindspec for image segmentation. The name of the design comes from its distinctive form, wherever the feature maps from convolution half in down sampling step ar fed to the up-convolution half in up-sampling step. U-net has been used extensively for medicine applications to find cancer , urinary organ pathologies and pursuit cells etc. U-net has tested to be terribly powerful segmentation tool in canaries with restricted knowledge (less than fifty coaching samples in some cases). Another advantage of employing a U-net is that it doesn't have any absolutely connected layers, so has no restriction on the dimensions of the input image. This feature permits North American nation to extract options from pictures of various sizes, that is a horny attribute for applying deep learning to sound reproduction medicine imaging knowledge. the flexibility of U-net to figure with little knowledge and no specific demand on input image size build it a powerful candidate for image segmentation tasks. Another reason to settle on the U-net design is that the letter U. because the knowledge set was provided by Udacity and as am presently listed in Udacity's self-driving automotive, selection of U-net was a fitting tribute to Udacity.

The input to U-net could be a resized 960X640 3-channel RGB image and output is 960X640 1-channel mask of predictions. we tend to wished the predictions to replicate likelihood of a picture element being a vehicle or not, therefore we tend to used associate degree activation perform of sigmoid on the last layer.

Training:

This 1image was randomly samples and augmented from all training images. As we chose a batch size of 1, we chose adam optimizer with a learning rate of 0.0001. Setting up the training itself was straight forward, but training the segmentation model made my Titan X gpu cringe. To perform 10000 iterations, my titan X machine took about 20 minutes.

Figure 4. Two-stream FCN architecture: The main foreground branch (1) is complemented by a contour branch (2) which improves the localization of the boundaries (3).

Fully convolutional networks layer of data in a convolutional network is a three-dimensional array of size h×w×d. The h and w are spatial dimensions, and d is the feature or channel dimension. The first layer is the image, with pixel size h×w, and d color channels. Locations in higher layers correspond to the locations in the image they are path-connected to, which are called their receptive fields. Convolutional networks are built on translation invariance. Their basic components (convolution, pooling, and activation functions) operate on local input regions, and depend only on relative spatial coordinates.

Writing xij for the data vector at location (I , j) in a particular layer, and yij for the following layer, these functions compute outputs
yij by yij = fks ({xsi+δi,sj+δj}0≤δi,δj≤k)
where k is called the kernel size, s is the stride or sub sampling factor, and fk determines the layer type: a matrix multiplication for convolution or average pooling, a spatial max for max pooling, or an element wise nonlinearity for an activation function, and so on for other types of layers. This functional form is maintained under composition, with kernel size and stride obeying the transformation rule fk While a general deep net computes a general nonlinear function, a net with only layers of this form computes a nonlinear filter, which we call a deep filter or fully convolutional network. An FCN naturally operates on an input of any size, and produces an output of corresponding (possibly re sampled) spatial dimensions. A real-valued loss function composed with an FCN defines a task. If the loss function is a sum over the spatial dimensions of the final layer,

$$`(x;\theta) = \sum_{ij} `0(x_{ij};\theta),$$

its gradient will be a sum over the gradients of each of its spatial components. Thus stochastic gradient descent on `computed on whole images will be the same as stochastic gradient descent on `0, taking all of the final layer receptive fields as a minibatch. Whole image training in our other experiments.

Class Balancing:
Fully convolutional training can balance classes by weighting or sampling the loss. Although our labels are mildly unbalanced (about 3/4 are background), we find class balancing unnecessary.

Dense Prediction:
The scores are up sampled to the input dimensions by deconvolution layers within the net. Final layer deconvolutional filters are fixed to bilinear interpolation, while intermediate up sampling layers are initialized to bilinear up sampling, and then learned. Augmentation We tried augmenting the training data by randomly mirroring and "jittering" the images by translating them up to 32 pixels (the coarsest scale of prediction) in each direction. This yielded no noticeable improvement.

III. RESULTS

This is awfully attention-grabbing research work for several reasons. This is first time segmentation model on a comparatively wild knowledge set is implemented. It absolutely was the primary time I saw my Titan X laptop struggle to run through convolutional networks. Overall, i used to be very proud of the results, and shocked by however well the U-net design learned to observe cars. In some cases, it performed higher than humans marking the first knowledge set. We used to be particularly shocked once it properly known automotive within the opposite late that we had incomprehensible till I saw the red blob over railings the rule was astonishingly quick. It took 200ms to form ten predictions (average of 20ms per image), this enclosed reading file off of disk, and drawing the blobs. Finally to check however well the model generalizes to unseen knowledge, we tend to run the U-net rule on one amongst the pictures from route driving. Figure below shows that the model properly known the cars, each in its lane and within the opposite lane. What's even additional shocking is that the model known cars that were occluded by the railings on the facet. I didn't notice the automotive till I saw red marks from U-net segmentation. The rule did establish some further region as attainable automotive location.

IV. CONCLUSION

Fully convolutional networks are a rich class of models, of which modern classification convnets is a special case. Recognizing this, extending these classification nets to segmentation, and improving the architecture with multi-resolution layer combinations dramatically improves the state-of-the-art, while simultaneously simplifying and speeding up learning and inference.

REFERENCES

[1]K.He,X.Zhang,S.Ren,andJ.Sun.patialpyramidpooling in deep convolutional networks for visual recognition. In ECCV, 2014. 1, 2
[2]Y.Jia,E.Shelhamer,J.Donahue,S.Karayev,J.Long,R.Girshick,S. Guadarrama, and T. Darrell. Caffe: Convolutionalarchitectureforfastfeatureembedding. arXivpreprint arXiv:1408.5093, 2014. 7
[3] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. Biological cybernetics, 55(6):367–375, 1987.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 1, 2, 3, 5
[5]Y.LeCun,B.Boser,J.Denker,D.Henderson,R.E.Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to hand-written zip code recognition. In Neural Computation, 1989. 2, 3
[6] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. M¨uller. Efficient backprop. In Neural networks: Tricks of the trade, pages 9–48. Springer, 1998. 7
[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In ICLR, 2015. 6
[8] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In ECCV, 2016. 2, 3
[9] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016. 2
[10] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In BMVC, 2014. 5
[11] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen. Jumpcut: Non-successive mask transfer and interpolation for video cutout. ACM Trans. Graph., 34(6), 2015. 2, 3, 5, 7
[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. TPAMI, 35(8):1915– 1929, 2013. 2, 3

[13] K. Fragkiadaki, G. Zhang, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In CVPR, 2012. 5
[14] R. Girshick. Fast R-CNN. In ICCV, 2015. 1, 3
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 1
[16] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. CVIU, 117(10):1245–1256, 2013. 5
[17] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In CVPR, 2010. 2, 3, 5, 7
[18] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015. 2, 4
[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 1, 3, 4
[20] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In ECCV, 2014. 2, 5, 8
[21] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In CVPR, 2017. 3
[22] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In CVPR, 2017. 3
[23] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. In ICLR, 2016. 1, 4
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 1, 3, 4
[25] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In ICCV, 2011. 5
[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016. 1
[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015. 2, 3, 4
[28] M. Kristan et al. The visual object tracking VOT2015 challenge results. In Visual Object Tracking Workshop 2015 at ICCV 2015, Dec 2015. 8
[29] K. Maninis, J. Pont-Tuset, P. Arbel´aez, and L. Van Gool. Convolutional oriented boundaries. In ECCV, 2016. 1, 3, 4, 5
[30] K. Maninis, J. Pont-Tuset, P. Arbel´aez, and L. Van Gool. Deep retinal image understanding. In MICCAI, 2016. 2, 3, 4
[31] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In CVPR, 2014. 5
[32] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In CVPR, 2016. 3, 8
[33] N. Nicolas M¨arki, F. Perazzi, O. Wang, and A. SorkineHornung. Bilateral space video segmentation. In CVPR, 2016. 2, 5, 7
[34] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In ICCV, 2015. 2, 3
[35] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. TPAMI, 36(6):1187– 1200, 2014. 5
[36] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In ICCV, 2013.
[37] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In CVPR, 2016. 2, 5, 6
[38] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In ICCV, 2015. 2, 5, 7
[39] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Doll´ar. Learning to refine object segments. In ECCV, 2016. 3
[40] J. Pont-Tuset, P. Arbel´aez, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. TPAMI, 2017. 3, 5

AUTHORS PROFILE

SANTHIYA.S is an M.Phil Scholar studying in Department of Computer Applications of Sri Sarada College for Women, Tirunelveli. She received her Bachelor‘s degree in Information Technology and Master of Computer Application from Sri Sarada College For Women,Tirunelveli.

