OBJECT SEGMENTATION IN THE DEEP NEURAL NETWORK FEATURE DOMAIN FROM HIGHLY CLUTTERED NATURAL SCENES

Hayder Yousif, Zhihai He

University of Missouri-Columbia
Dept of Electrical and Computer Engineering
hyypp5@mail.missouri.edu, hezhi@missouri.edu

Roland Kays

North Carolina State University
Dept of Forestry and Environmental Resources
rokays@gmail.com

ABSTRACT

Deep convolutional neural networks (DCNNs) offer an effective hierarchical representation of images for various vision analysis tasks, including classification and detection. In this paper, we propose to study background modeling and object segmentation from highly cluttered natural scenes in the DCNN feature domain instead of traditional pixel domain. Specifically, we first design and train a DCNN for animal-human-background object classification, which is used to analyze the input image to generate multi-layer feature maps, representing the responses of different image regions to the animal-human-background classifier. From these feature maps, we construct the so-called *deep objectness graph* for accurate animal-human object segmentation with graph cut. The segmented object regions from each image in the sequence are then verified and fused in the temporal domain using background modeling. Recognizing that the DCNN is very computation-intensive, we explore a fast and efficient design of the DCNN which finds a good trade-off between complexity and the classification-segmentation performance. Our experimental results demonstrate that our proposed method outperforms existing state-of-the-art methods on the camera-trap dataset with highly cluttered natural scenes.

Index Terms— Camera-trap, graph cut, Convolutional Neural Network, image segmentation, object detection.

1. INTRODUCTION

Camera-traps are stationary camera-sensor systems attached to trees in the field. Triggered by animal motion, they capture short image sequences of the animal appearance and activities along with other sensor data, such as light level, moisture, temperature, and GPS sensor data. Operating in a non-invasive manner, they record animal appearance without disturbance [1, 2]. During the past several years, a vast amount of camera-trap images have been collected, far exceeding the capability of manual image processing and annotation by human. There is an urgent need to develop automated animal detection, segmentation, tracking, and biometric feature extraction tools for automated processing of these massive camera-trap datasets. Images captured in natural environments represent a large class of challenging scenes that have not been sufficiently addressed in the literature. These types of scenes are often highly cluttered and dynamic with swaying trees, rippling water, moving shadows, sun spots, rain, etc. It is getting more complicated when natural animal camouflage adds extra complexity to the analysis of these scenes.

Detecting and segmenting moving objects from the background is an important and enabling step in intelligent video analysis. For camera-traps image sequences, the major challenge is to handle highly dynamic background scenes. [3] proposes a method for constructing a background model using the median pixel value in dynamic scenes. Methods based on robust principle component analysis (RPCA) [4] has been developed for foreground object detection from noisy and dynamic background scenes.

Selective Search [5] is based on computing multiple hierarchical segmentations based on superpixels and placing bounding boxes around them as object proposals. R-CNN [7] is agnostic to a particular region proposal method that uses selective search. [8] proposes to exploit a pre-trained large convolutional neural network (CNN) to generate deep features for conditional random fields CRF that used for image segmentations. Mask R-CNN [9] extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. [10] introduces novel features that combine together colour and texture information for semantic segmentation purpose.

In this paper, we propose to explore deep learning methods for human-animal segmentation and classification from highly cluttered camera-trap images. Using pixel domain image representation to segment a highly cluttered scene is significantly inefficient. To address this issue, we propose to couple effective region proposals system with deep learning classification to develop a fast and accurate scheme for human-animal detection.

The rest of the paper is organized as follows. Section 2 provides an overview of the proposed method. In Section 3, we study the complexity-accuracy trade-off of DCNNs and

propose a fast and accurate DCNN scheme to segment and classify human and animal. In Section 4, we introduce the proposed deep objectness graph for cross-frame region proposals extraction. Section 5 reviews our background modeling and subtraction. Section 6 presents our experimental results and comparison with existing methods.

2. SYSTEM OVERVIEW

Our proposed system consists of four major components. First, the foreground objects (human or animals) are detected and segmented from the background using our background modeling and subtraction method. Second, we introduce a deep objectness graph region proposals method. Third, we fuse the region proposals and verify them temporally. Fourth, our fast DCNN module is used to classify the foreground proposals into human, animal, and false positive background classes. The proposed system flow chart is shown in Fig. 1.

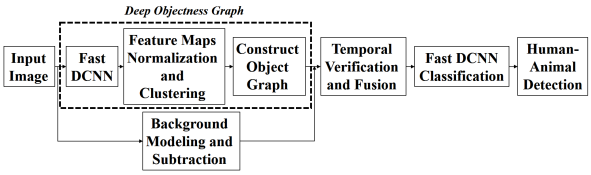

Fig. 1. Flow chart of the proposed system.

Compared to existing methods developed in the literature, our method has the following three major contributions. **First**, the proposed background modeling and object segmentation is performed in the DCNN feature domain, instead of traditional pixel domain. **Second**, instead of generating thousands of region proposals per image for object detection, our method generates less than fifty region proposals from a camera-trap image. This results in significantly reduced detection time. **Third**, our system couples semantic understanding of image scenes using DCNN classification with spatiotemporal background modeling to achieve significantly improved performance in object segmentation from highly cluttered natural scenes.

3. FAST DCNN FOR IMAGE ANALYSIS

Due to highly cluttered scene of camera-trap images, RGB or Lab color domain do not provide sufficient graph information that can separate the scene's objects from the background. Instead, we trained a human/animal CNN model that is able to transfer the image into a deep representative domain. The designed model is used to obtain the deep objectness graph and to classify the region proposals.

We study the complexity-accuracy trade-off in the DCNN. As we know, the DCNN is very computation-intensive. Note that the total number of region proposals to be classified by the DCNN is large. Therefore, there is an urgent need to speed up the DCNN while largely maintaining its classification accuracy. We have identified three major parameters: input size, number of layers, and number of filters, which can be used to effectively control the computational complexity of DCNN. Table 1 shows the 14 layers of our designed DCNN architecture. Each convolutional layer is followed by Rectified Linear Unit (ReLU) and normalization layers.

Table 1. Our CNN networks architecture. Conv is the convolutional layer, Pool is the pooling layer, and FC is the fully connected layer.

Layer	Type	K size	Filter #	Stride	Pad
1	Conv1	6	128	2	0
4	Pool1	3	-	2	0
5	Conv2	5	128	1	2
8	Pool2	5	-	2	0
9	Conv3	5	128	1	1
11	FC4	7	2048	1	0
14	FC5	7	2048	1	0

The final foreground region proposals are verified through DCNN classification. We extract a 4048-dimensional feature vector from each region proposal using FC4 and FC5. For each class, we score the extracted feature vector using the linear SVM trained for that class. We refine the overlapped foreground proposals by suppressing the low confidence classified regions. Table 2 demonstrates the resource requirements and the associated testing speed using different DCNN architectures. We can see that designing smaller input image size (96×96) with less layers reduces the complexity by 18 times at the small loss of classification accuracy.

Table 2. DCNN training memory allocation and patch testing time comparison.

DCNN	Memory(MB)	(patch/ms)	Acc(%)
AlexNet [11]	217	336.1	91.98
VGG-F [12]	316	395.6	92.43
VGG-S [12]	393	304.7	**93.99**
ours	**50**	**17.9**	91.39
ours+SVM	-	20	92.05

4. DEEP OBJECTNESS GRAPH

Each feature map represents a response of the input to a specific distinctive image characteristics. Human, animal, and background have different responses for the same filter bank. Feature maps do not provide the scene description and objects boundaries. To transfer the feature maps into semantics representation of images for object segmentation using graph

cut, we propose to cluster the deep feature maps with K-mean clustering before building the deep objectness graph using the graph-cut. The overview of our deep objectness graph cut is depicted in Fig. 2. Specifically, an image $I \in \Re^{H\times W\times D}$ convolve with a filter bank $f \in \Re^{R\times C\times D\times N}$ and biases s yields $I' \in \Re^{H'\times W'\times D'}$

$$I'_{i'j'n} = s_n + \sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{n=1}^{N} f_{rcn} \times I_{i'+r-1,j'+c-1,d,n} \quad (1)$$

All negative outputs are removed through ReLU:

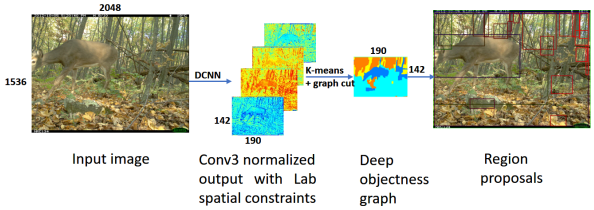

Fig. 2. Our deep objectness graph for cross-frame region proposals.

$$I'_{ijn} = \max\{0, I_{ijn}\} \quad (2)$$

The feature maps are concatenated with a re-scaled Lab color input image. The $X \times Y \times Z$ size concatenated matrix is then normalized to obtain the final image representations:

$$I_{xyz} = \sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z} \frac{I_{xyz} - \min_z I_{xy}}{\max_z I_{xyz} - \min_z I_{xyz}}. \quad (3)$$

Fig. 3 shows samples from the normalized feature maps. Then, I_{xyz} is converted to a K-D image by incorporating spatial and deep image information to K-means objective function with K clusters as follows:

$$C_{xyk} = \arg\min_{K} \sum_{k=1}^{K}\sum_{z=1}^{Z} \parallel I_{xyz} - \mu_k \parallel^2, \quad (4)$$

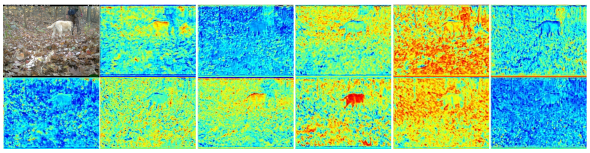

Fig. 3. Samples from the normalized feature maps.

where μ_k is the mean of pixels in the feature map I_{xyz}. Our deep objectness graph $G =< V, E >$ consists of a set of nodes V and a set of directed edges E that connect them. A graph normally contains some additional special nodes that are called classes [13]. There are two types of edges in the graph: N-links and T-links. N-links connect pairs of neighboring pixels or voxels. Thus, they represent a neighborhood system in the deep image representations. The cost of N-links corresponds to a penalty for discontinuity between the deep image representation pixels. T-links connect pixels with class [14]. The Euclidean norm is used to measure the distance of each pixel value to the class prototype:

$$D_{xyk} = \parallel I_{xyk} - C_{xyk} \parallel^2 \quad (5)$$

We set the cost as a square matrix M with K rows and columns:

$$M_{xy} = \begin{cases} 0, & \text{if } x = y, \\ A, & \text{otherwise}, \end{cases} \quad (6)$$

where A is constant, we set A to 3 to strengthen the neighborhood constraints more and makes the segments larger. Graph cut is then applied to the distance D with the cost M to obtain the final segmentation result.

It should be pointed out that the deep image representation CNN network is a part of classification network. Specifically, there is no further calculations after Conv3. All region proposals extraction is done in low scale resolution and then the bounding boxes coordinates are scaled back to fit with original image size.

5. TEMPORAL VERIFICATION AND FUSION

In traditional background modeling, the foreground-background decision is performed at the pixel domain in a small local neighborhood. In this paper, we have constructed a semantic region representation of the image using the DCNN classifier. We expect that background modeling in the temporal domain based on this semantic region representation will be much more efficient than traditional background modeling in the pixel domain. From the above two sections, we have successfully generated object regions for each frame in the sequence. The next step is to verify and fuse these region segmentation results across frames using background modeling to produce the final moving object (animals and human) segmentation results.

The key idea of our background modeling is to find the block in the sequence which has the minimum feature distance to all other co-located blocks. Specifically, frame I_n, $1 \le n \le N$, is divided into overlapping $B \times B$ blocks b. The block in I_n at block location (i, j), $1 \le i, j \le B$, is denoted by $b_n(i, j)$.

$$b_n(i,j) = \arg\min_{n} \sum_{k=1}^{N} \Phi[b_n(i,j), b_k(i,j)], \quad (7)$$

$\Phi[b_n(i,j), b_k(i,j)]$ represents the feature distance between blocks $b_n(i,j)$ and $b_k(i,j)$. The HOG histogram of the block represents the feature vector. We use the χ^2 distance between two histograms. Rather than using hard thresholding, the foreground regions for each frame are obtained by thresholding the feature difference between the frame and the constructed background.

We build a spatial verification model to remove noisy false positive regions. Let n_i be the number of occurrences of intensity level i, then the Shrinked Histogrm Length (SHL) for an image of size $w \times h$ is defined as:

$$SHL = \sum_{i=0}^{L-1} p_i, p_i = \begin{cases} 1, & \text{if } \frac{n_i}{wh} > th \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where L is the total number of intensity levels in the image. In our experiments, we set the threshold th to zero and SHL to 70. If a patch has SHL less than 70, it will be classified as background otherwise further verification is needed. Our experimental results demonstrate that this verification module is able to reduce the background false alarms while maintaining foreground regions.

6. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed human-animal detection method on real-world camera-trap images. Our camera-trap test dataset consists of 100 samples. Each sample is a sequence of 10 images, being triggered by motion or heat. We manually labeled all animals and persons with bounding boxes in these 1000 images, which serves as the ground truth for performance evaluations. The foreground segmentation output patches extracted from a large number of sequences of Reconyx camera type are used to build our training dataset that contains 17,960 image patches consisting of 4,357 human, 6,160 animal, and 7,443 background.

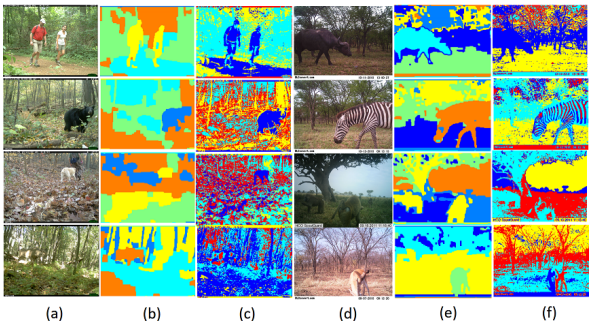

Fig. 4. Deep objectness graph segmentation comparison: (a) samples from our camera-trap dataset, (b) Our deep objectness graph segmentation regions, (c) original graph cut regions, (d) samples from Serengity dataset [15], (e) Our deep objectness graph segmentation regions, (f) original graph cut regions.

Fig. 4 illustrates the difference between the original Lab color based graph-cut and our proposed deep objectness graph with the same input image size and number of clusters used for segmentation. Note that the animal patches used for training the region proposal network do not contain any of Serengity dataset [15] animals, such as zebra and monkey. The trained DCNN filter banks are able to discriminate between these animals, that share some shape and texture features with camera-trap patches, and the background. Fig. 5

Fig. 5. Human-animal detection results. The ground truth, detected human, and detected animal are in red, green, and blue bounding boxes, respectively.

shows how our proposed region proposals and detection system is efficient for camera-trap sequences. We compare the result of our background segmentation with several state of the art background segmentation methods. Table 3 shows the performance of our fused background modeling with deep objectness graph over other background subtraction methods. We used our fast DCNN classifier to classify the foreground regions proposed from each method. The true positive regions are regions that have the same class as the ground truth and have Intersection over the Union $(IoU) \geq 0.5$. The misclassified ground truth regions or foreground classified regions that do not have $IoU \geq 0.5$ are assumed to be true negative regions. We can see that our proposed method outperforms existing state-of-the-art methods by large margin, up to 20%.

Table 3. Performance comparison on human-animal detection in the camera-trap dataset between our proposed work and other methods.

Method	Recall
RPCA-PCP [4, 16]	47.77%
FastLADMAP [17]	44.85%
Deep-Semi-NM [18]	46.05%
OR1MP [19]	43.30%
This work	**63.06%**

7. CONCLUSION

We proposed a new human-animal detection and recognition framework. We made an efficient region proposals extraction by fusing our background subtraction and the proposed deep objectness graph. We optimized the region proposals system through segmenting the image in a very low resolution. Our designed DCNN was able to reduce the classification time by 18 times and maintain high accuracy.

8. REFERENCES

[1] Tim CD Lucas, Elizabeth A Moorcroft, Robin Freeman, J Marcus Rowcliffe, and Kate E Jones, "A generalised random encounter model for estimating animal density with remote sensor data," *Methods in Ecology and Evolution*, vol. 6, no. 5, pp. 500–509, 2015.

[2] Jakub W Bubnicki, Marcin Churski, and Dries PJ Kuijper, "Trapper: an open source web-based application to manage camera trapping projects," *Methods in Ecology and Evolution*, 2016.

[3] Agnieszka Miguel, Sara Beery, Erica Flores, Loren Klemesrud, and Rana Bayrakcismith, "Finding areas of motion in camera trap images," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1334–1338.

[4] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[5] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[6] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1529–1536.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] Fayao Liu, Guosheng Lin, and Chunhua Shen, "Crf learning with cnn features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," *arXiv preprint arXiv:1703.06870*, 2017.

[10] Daniele Ravì, M Bober, Giovanni Maria Farinella, Mirko Guarnera, and Sebastiano Battiato, "Semantic segmentation of images exploiting dct based features and random forest," *Pattern Recognition*, vol. 52, pp. 260–273, 2016.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.

[13] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] Vladimir Kolmogorov and Ramin Zabin, "What energy functions can be minimized via graph cuts?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 2, pp. 147–159, 2004.

[15] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer, "Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna," *Scientific data*, vol. 2, pp. 150026, 2015.

[16] Paul Rodriguez and Brendt Wohlberg, "Fast principal component pursuit via alternating minimization," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 69–73.

[17] Zhouchen Lin, Risheng Liu, and Zhixun Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in neural information processing systems*, 2011, pp. 612–620.

[18] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W Schuller, "A deep semi-nmf model for learning hidden representations.," in *ICML*, 2014, pp. 1692–1700.

[19] Zheng Wang, Ming-Jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye, "Orthogonal rank-one matrix pursuit for low rank matrix completion," *SIAM Journal on Scientific Computing*, vol. 37, no. 1, pp. A488–A514, 2015.