UNSUPERVISED PANOPTIC SEGMENTATION

by

Sajeel Aziz, Bachelor of Engineering, Ryerson University 2017

An MRP presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Engineering
in the program of
Electrical and Computer Engineering

Toronto, Ontario, Canada, 2020,

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this MRP. This is a true copy of the MRP, including any required final revisions. I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research. I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my MRP may be made electronically available to the public

ABSTRACT

Unsupervised Panoptic Segmentation
Sajeel Aziz
Master of Engineering
Electrical and Computer Engineering
Ryerson University, Toronto, Canada, 2020

The contributions of this paper are two-fold. We define unsupervised techniques for the panoptic segmentation of an image. We also define clusters which encapsulate the set of features that define objects of interest inside a scene. The motivation is to provide an approach that mimics natural formation of ideas inside the brain. Fundamentally, the eyes and visual cortex constitute the visual system, which is essential for humans to detect and recognize objects. This can be done even without specific knowledge of the objects. We strongly believe that a supervisory signal should not be required to identify objects in an image. We present an algorithm that replaces the eye and visual cortex with deep learning architectures and unsupervised clustering methods. The proposed methodology may also be used as a one-click panoptic segmentation approach which promises to significantly increase annotation efficiency. We have made the code available privately for review[1]

[1] `https://github.com/ShujaKhalid/project_cygnus`

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Naimul Khan *Ryerson University* for his guidance with this project. We would also like to acknowledge the contributions of Shuja Khalid *University of Toronto* for his technical and moral support.

Contents

List of Tables

List of Figures

INTRODUCTION

Scene understanding has generated significant interest in recent times and complete scene understanding aims to not just learn about the presence of various entities in a 2D or 3D scene, it aims to form relations between these un-defined entities. Recent work has focused on techniques such as semantic [15], instance [12] and panoptic segmentation [11] to detect and annotate the exact specifications of an object in a scene. However, the task of relating these objects requires intuition that existing models have yet to master. If we are able to identify all of the possible "things" and "stuff" classes, we can then work towards creating informed relations between them. This paper proposes an algorithm for the creation of unsupervised domain representation of scenes. To achieve this, pre-trained models are used for boundary detection (HED [24]) and feature extraction (ResNet [20]).

The features are mapped to closed contours within the original image and subsequently clustered in feature space to identify variations based on texture, colour, edge, shape or orientation. The feature space is modelled as a Gaussian Mixture Model (GMM) [18] where each Gaussian represents a unique class. The resulting class information is then mapped back to the original image to create a two-channel map where each pixel in each contour is then assigned a tuple (*stuff*, *thing*) as per the definition of panoptic segmentation.

Specifically, these methods are expected to work by using the properties of the trained Convolutional Neural Networks(CNNs). The HED model uses a trained CNN to output only the edges of an image by running the image through it's

filters. Next, the ResNet model outputs a set of feature maps on the original image resulting from many filtering operations. This finds the important spatial and textural properties of an image. The output is stopped before the classification layer so that the output of each filter is obtained in a stacked structure. Each output is assumed to have different information about the image. Therefore, information-rich pixels can be obtained by using the stacked outputs from the filters/feature maps.

Another important concept to mention for this approach is of manifold learning[8]. Manifold Learning is for situations where there are many connected data points and each data point has a neighbourhood of points associated with it. The requirement from Manifold Learning is to recognize the most important dimensions, or features, to find a less complex representation of the same data. This is generally done by finding the highest distances, or variances, between points in the original dimension space and only preserving those points. This distinguishes the points more clearly, helping subsequent models perform classification more efficiently.

We hypothesize that this algorithm will aid in the creation of a feature space where objects of the same class, from different images, will be clustered similarly. This mapping of similar objects in a similar manner has the added benefit of creating explainable representations of visual objects.

RELATED WORK

2.1 Semantic/Instance Class Boundary Detection

A variety of works have focused on the task of boundary detection and classification, where the goal is to not only detect boundaries of key objects but also to assign classes to the identified boundaries.

The work in [1] uses weakly supervised learning to predict semantic segmentation labels from image level class labels in an end-to-end manner. Similarly, [2] also proposes an end-to-end trainable framework for assigning instance-level labels using image level class labels. The training required for such a process is significantly reduced as segmentation annotations are time-intensive to procure. We aim to provide panoptic segmentation labels without the use of image level class labels.

The study in [22] uses generative probabilistic modelling to model appearance level cues, such as colour, edge, shape and pose of objects within an image. This approach has motivated our work as we create probabilistic models for the representation of different classes within an image. However, we do so implicitly by using rich feature vectors extracted from a deep neural network. Our algorithm is also more powerful as it generalizes to datasets such as COCO [14] and Pascal VOC [21] which have challenging images consisting of multiple classes and segments.

The research in [17] proposes an end-to-end supervised learning architecture that utilizes semantically meaningful boundaries for the task of semantic segmen-

tation. This is similar to other supervised techniques that yield state-of-the-art results in semantic/instance segmentation such as Mask-RCNN [10], U-Net [19], probabilistic U-Net [23], amongst others. Each of these techniques make use of expensive annotations for their training process and fail to generalize to cases where an image contains a class that that the model has not been trained on.

The work of [4] uses boundary neural fields in conjuction with FPNs [13] as a global energy model that incorporates boundary cues for the enhancement of semantic segmentation predictions. The use of boundary cues for the purpose of semantic and instance segmentation has thus been explored. We attempt to improve on this by utilizing boundary cues and by designing an algorithm that does so in an unsupervised manner.

There have thus been a number of attempts to improve the performance of models for the tasks of semantic, instance, and panoptic segmentation, using supervised and semi-supervised techniques. We attempt to improve on the state-of-the art by using pre-trained models for boundary detection and feature extraction. The proposed approach is thus unsupervised and does not require pixel or image level annotations for the purpose of panoptic segmentation.

2.2 Knowledge Representation/Feature representation

From [26] we can see the demonstration of a method to perform feature selection and extraction simultaneously. The technique in the work uses manifold learning on spectral and spatial features from remote-sensing image data for classification, where spectral and spatial features are concatenated and manipulated in a way to preserve only important individual features while also utilizing complimentary attributes of the spectral and spatial information.

The experiments in [25] show a similar concept to [26] where the spectral and spatial features are picked by using successive dimensionality reduction techniques which help preserve the relevant information required for classification of image data.

We can also see the work in [7] which uses a similar technique as [26] and [25] called SC-MK, where a superpixel is obtained from the image data via running through multiple kernels. These pixels are then used as input to a classifier to perform the final classification.

The study in [3] uses CNN kernels for feature extraction at different levels, then performs SVM or Random Forest classification directly on the feature maps obtained. Using this method, the authors got a competitive result to regular CNN learning techniques even if the CNN is only partially trained.

METHOD

3.1 Datasets

We use the MS COCO dataset[1] to perform the evaluation as this dataset is very frequently used to further state-of-the-art techniques in panoptic segmentation. This dataset consists of 80 "thing" categories, 32 "stuff" categories, and 17 "stuff" categories which were merged or ignored. To assign the thing/stuff classes to each image, the image has to be transformed by setting the first channel to instance id, second channel to category id, and the third to zero.

[1] COCO, `http://cocodataset.org/`

3.2 Metrics

The metrics used to evaluate the method proposed in this paper is based the Panoptic Quality (PQ) metric. The PQ metric is accompanied by 2 other metrics called Segmentation Quality (SQ) and Recognition Quality (RQ). These 3 metrics are divided into 3 more categories, namely, "All", "stuff", and "thing".

The calculation of the PQ metric involves first segment matching. The IoU threshold for a predicted segment instance matching the ground-truth instance is 50 percent. Next, the predicted categories are compared to the ground truth to get the True Positives and False Positives. Overall, the formula for the measure for the PQ metric is below:

$$\text{PQ} = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{3.1}$$

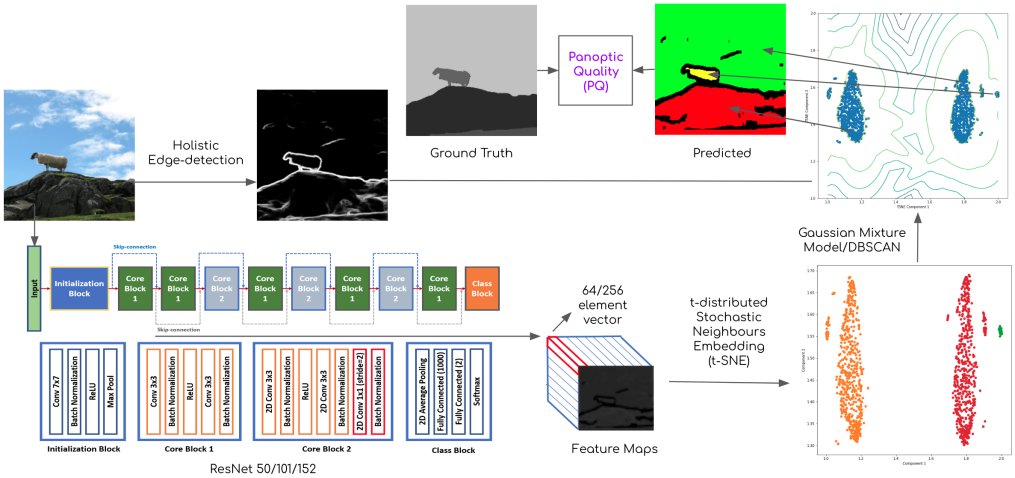
The decomposition of the PQ metric can also be done in terms of the product of SQ and RQ as follows:

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \tag{3.2}$$

Where parameters p and g denote the predicted, and ground truth label respectively.

3.3 Architecture

This architecture combines feature extraction techniques, dimensionality reduction, and probabilistic clustering techniques to create pixel-level labels of input images based on texture/shape information. This method is illustrated in Figure 3.1.



(a) Process Description

Figure 3.1: Illustration of the main architecture

3.4 Implementation

The algorithm defined in this paper doesn't explicitly optimize a cost function using traditional supervised learning approaches. We use a pre-trained model to define an image in terms of its edges and we complete a set of pixel level assignments as per the requirements of Panoptic Segmentation. The algorithm is defined in 1.

First, the image is passed through a pretrained HED model, which identifies all of the contours in the image. Next, each contour is assigned a label. This is achieved by processing the image to set every non-edge area to a zero-value, and is followed by strengthening the edges to make sure there are no empty spaces inside the contours. A generic flood-fill algorithm is then applied, which assigns a unique value to each closed contour.

A pre-trained Resnet model is then used to output a set of feature maps for the image. These feature maps are stored for each segment by only selecting the areas where the segment is equal to the unique value that was previously set by the flood-fill algorithm. Therefore, the feature maps stored in total will be Nx256 for Resnet152, where N refers to the no of segments stored and 256 refers to the no. of channels per segment.

Each identified segment is looped over and an average-pooling operation is performed to reduce the computational burden on the tSNE[16] module. 256 element vectors are stored for each pixel in these average pooled feature maps. Next, we incorporate spatial/shape information, by affixing a binary map of spatial features to each 256 element vector. The resulting N outputs are flattened, looped over for each segment, and the flattened output, consisting of a combination of spatial and textural features is assigned to all pixels. The final dimension input to tSNE will consist of MxN pixels with (256+MxN) features, where M is number of average-pooled pixels chosen per segment and N is the number of segments. Note, that the number of feature maps taken can be decreased in case of computational constraints.

The resulting set of pixel level features is passed to the tSNE algorithm which

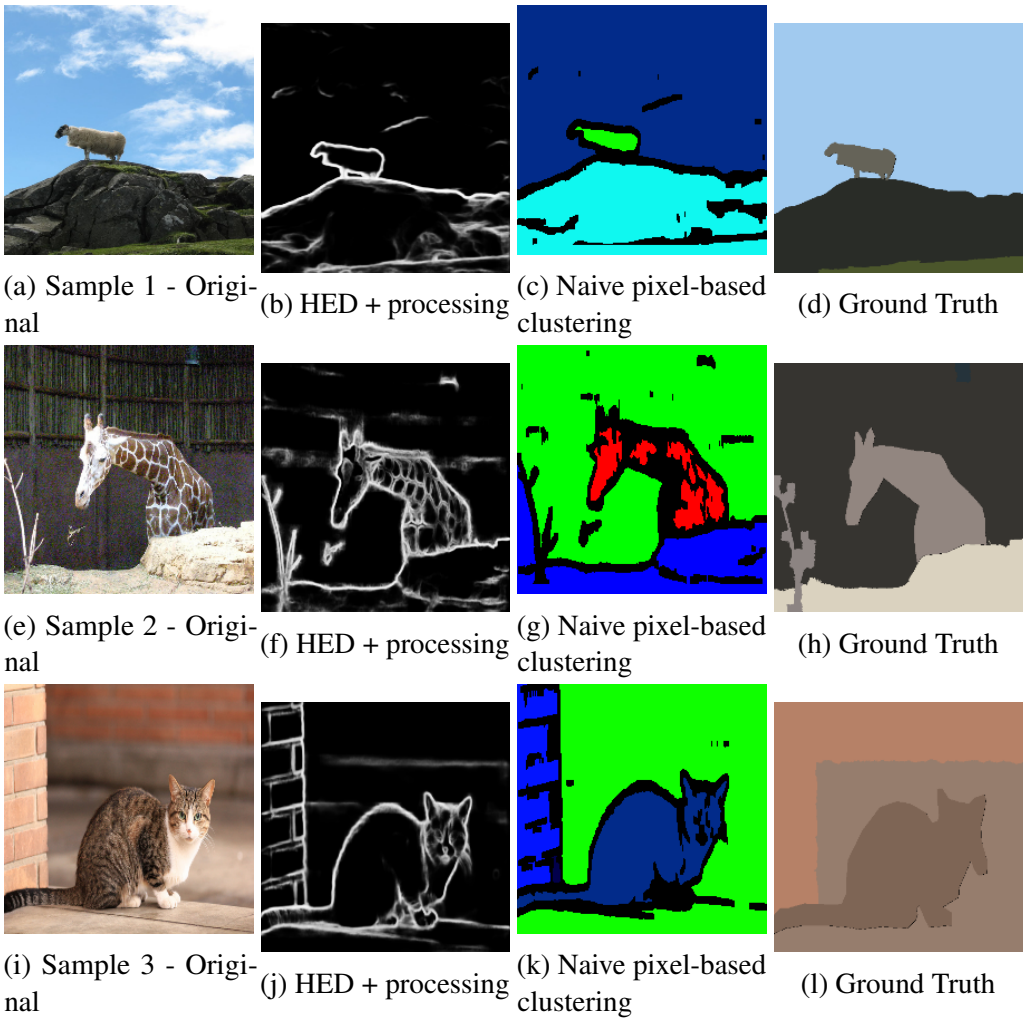
is used to decrease the number of per-pixel features from 256+(MxN) to a more manageable value. This allows for a lower dimension visualization of the high-dimensional clusters. The resulting clusters are visualized in 3.3. We input the pixel-level coordinates into a Gaussian Mixture Model(GMM) which calculates the log-likelihood of the points belonging to a cluster, while we set the component parameter simply to the number of segments found previously and multiply by 20 for a more compact cluster representation. The density of these clustered regions is then increased, by adding points on every cluster the where the log-likelihood is non-zero. Even if the amount of components is more than the classes, the closer points tend to be in a similar area of higher log-likelihood, and tend to cluster together in the GMM distribution. Next, the DBSCAN [5] algorithm is used to obtain the final labels for the clusters. Finally, using the indices of the points input to the DBSCAN algorithm, we can assign the classes to each segment and by extension, each pixel in the image.

Table 3.1: Comparison of Panoptic Quality (PQ) metric across existing state-of-the-art techniques.

	PQ	PQ (thing)	PQ (stuff)
Baseline	0.372	0.454	0.249
AUNet	0.465	0.559	0.325
UPSNet	0.466	0.532	0.367
JSISNet	0.272	0.296	0.234
Ours	0.022	0.013	0.036

Algorithm 1 CEReBRO algorithm

Inputs: $I_{orig}, \mathcal{N}$
Outputs: I_{mask}
$I_{edge} = HED(I_{orig})$
$I_{FM} = ResNet(I_{orig})$
$I_{CC} = FF(I_{orig})$
$K \leftarrow$ regions in I_{cc}
for $region = 0$ **to** K **do**
    $\{(x_1, y_1)...(x_N, y_N)\} \leftarrow (x, y) \sim$ region; Sampling N points from each region
    $\{F_1...F_N\} \leftarrow I_{FM}(\{(x_1, y_1)...(x_N, y_N)\})$; Extracting feature vectors from the sampled points
end for
$(U, V) \leftarrow tSNE(\{F_1...F_N\})$; Represent the feature vectors as a set of 2D coordinates
$X_{segs} \leftarrow GMM(\{F_1...F_N\})$; Use Gaussian mixture models to represent the clusters in space
$X_{class}, X_{segs} \leftarrow DBSCAN(X_{segs})$; Determine the no. of unique classes in an image
$I_{mask} \leftarrow$X$_{segs}$; Assign per-pixel (stuff, thing) tuples to the original image

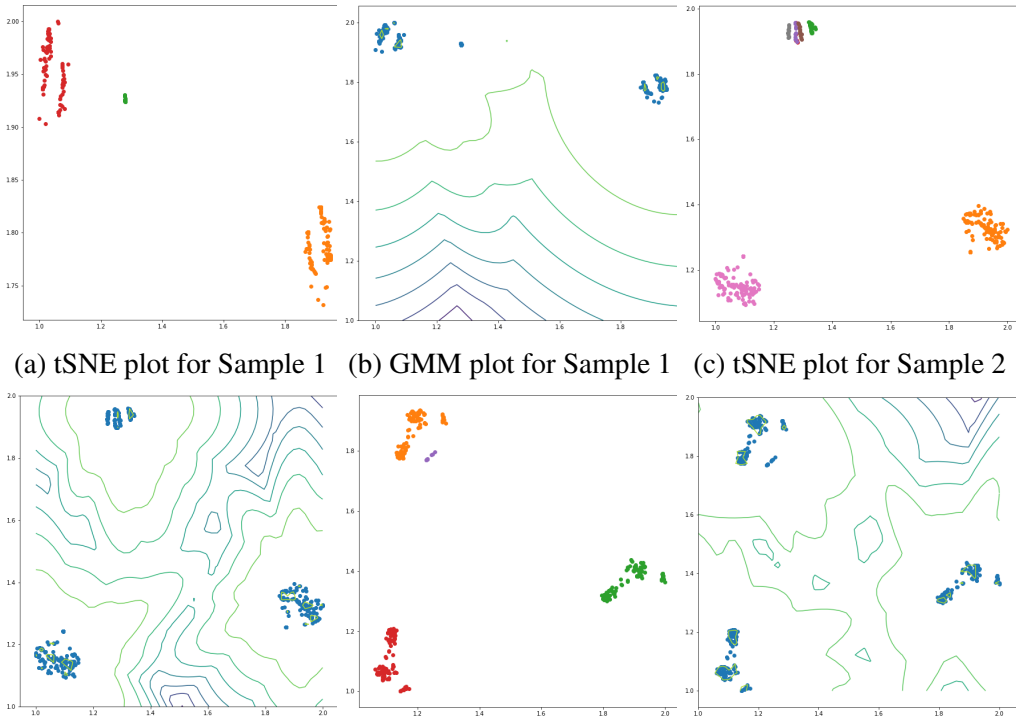


(a) Sample 1 - Original (b) HED + processing (c) Naive pixel-based clustering (d) Ground Truth

(e) Sample 2 - Original (f) HED + processing (g) Naive pixel-based clustering (h) Ground Truth

(i) Sample 3 - Original (j) HED + processing (k) Naive pixel-based clustering (l) Ground Truth

Figure 3.2: (a),(e),(i) show original input images, (b),(f),(j) show the output of the HED model which provides closed contours within the image for segment initial guess (c),(g),(k) After inputting each closed contour to tSNE,GMM, and DBSCAN, the segments are classified by clustering similar superpixel/spatial information , (d),(h),(i) are the corresponding ground truth panoptic segments



(a) tSNE plot for Sample 1 (b) GMM plot for Sample 1 (c) tSNE plot for Sample 2

(d) GMM plot for Sample 2 (e) tSNE plot for Sample 3 (f) GMM plot for Sample 3

Figure 3.3: tSNE and GMM Clustering used to obtain final classes for each initial segment found from HED. Each colour indicates a segment found from HED. The axes correspond to tSNE and GMM component coordinates

Table 3.2: Comparison of Segmentation Quality (SQ) metric across existing state-of-the-art techniques.

	SQ	SQ (thing)	SQ (stuff)
Baseline	0.771	0.815	0.706
AUNet	0.810	0.837	0.770
UPSNet	0.805	0.815	0.789
JSISNet	0.719	0.716	0.723
Ours	0.459	0.426	0.508

Table 3.3: Comparison of Recognition Quality (RQ) metric across existing state-of-the-art techniques.

	RQ	RQ (thing)	RQ (stuff)
Baseline	0.457	0.544	0.325
AUNet	0.561	0.663	0.407
UPSNet	0.569	0.646	0.453
JSISNet	0.359	0.396	0.306
Ours	0.030	0.019	0.047

RESULTS AND DISCUSSION

As this method is unsupervised, the presented results are inferior to that of supervised state-of-the-art techniques. The quantitative results presented in tables 3.1, 3.2, an 3.3. During our pixel designation process, we assign pixels with low confidence, the "N/A" class. This approach has certainly contributed to the low PQ scores in the paper. Since we envision this approach being used as a first-pass to aid annotators with the panoptic annotation task, we believe that there is significant utility in the approach even with low quantitative statistics.

The PQ score consists of two sub-scores, the *Segmentation Quality* SQ and the *Recognition Quality* RQ, eq. (2). The results for SQ, in table 3.2, indicate that the quality of segmentation for the unsupervised approach is not quite as good as that of the supervised techniques. However, the removal of artifacts in the image, and thinning of the contour lines might yield improved results in future works. In contrast the RQ metric is extremely low and requires significant improvement. This indicates that the algorithm is not able to correctly differentiate between objects of the same class. Future work will focus on improving the clustering of the model such that items of the same class are clustered sufficiently far from each other in feature space.

Further improvements may be made by strategically choosing hyperparameters for each module in Figure 3.1. Sensitive hyperparameters such as perplexity for tSNE, number of components for GMMs, and cluster size for DBSCAN are crucial for effective segmentation. More research will have to be done to find optimal

values for these hyperparameters. Another factor which needs to be investigated is an optimal method to incorporate both texture and spatial/shape information. In the current setup, the spatial/shape information is appended to the feature maps, so, the input dimensions being reduced from the tSNE module is dominated by the spatial/shape information as the feature maps are only 256 dimensions deep while the spatial/shape information is the flattened feature map itself.

4.0.1 Evaluation

We encoded our pixel-specific results in a format that allowed us to run the evaluation scripts provided by COCO. This was done to ensure consistency in the reported metrics. All images were processed sequentially and thus did not require a GPU. Our architecture is thus highly adaptable and does not require the use of a GPU as is the case for collaborative assistant tools that might utilize deep neural nets.

4.1 Limitations

This approach suffers from an number of limitations the most important one of which is its ability to generalize to more complex datasets.

4.1.1 Application to larger datasets/videos

We have attempted to use this algorithm on complex panoptic segmentation datasets, specifically Cityscapes [6]. The results are not promising as the noise in the data results in very tight clusters during the dimensionality reduction step of the proposed algorithm. The following GMM and DBSCAN steps are highly susceptible to creating erroneous clusters which in-turn results in incorrect classifications of the pixels. It is our intention to introduce a hyper-parameter for controlling the sensitivity of these clusters. This will allow us to experiment with larger and more complex datasets.

4.1.2 Processing time

Since our algorithm uses computational approaches such as tSNE, GMM and DBSCAN for each image, it is time-intensive. We would like to improve the operation times of the algorithm by streamling and parallelizing it as much as possible so that it may be used for its intended purpose, a collaborative panoptic segmentation assistant.

4.2 Extension to model explainability

The presented algorithm is inspired by the learning methodologies of infants. Infants do not understand the names of classes but are able to make out the shapes of individuals and items that are presented to them. By being able to differentiate objects, they can then attempt to classify them without even knowing the name of the object, using simple visual and textural cues. They are subsequently taught the names of the objects and through trial and error, and a timely signal from a teacher, their learning is complete. Through the repeated application of this structured approach to learning they are able to use visual sensory information and couple that with their visual cortex [9], which processes this information for a particular purpose. We replicate this physiological process by replacing the visual sensors with HED and a ResNet model (pre-trained on COCO), and the visual cortex with clustering of pixel spectra in a low-dimensional pixel space. This model is thus intuitive as it contains parallels from physiology, it is also more explainability than deep learning models which consists of a finite but large number of layers that learn complex representations internally and are essentially black-boxes.

CONCLUSION

We propose a panoptic segmentation approach in this paper that is completely unsupervised. The proposed architecture is significantly more interpretable than existing deep learning approaches as it derives inspiration from physiology, specifically the visual cortex system. The results presented in the paper indicate that this method has the potential to serve as a first pass for decreasing the workload for image annotators by making pixel-based predictions about segments of interest, within images, that it is confident about. However, more research is required to improve the algorithm to make it competitive with current state of the art, supervised panoptic segmentation techniques.

Bibliography

[1] J. Ahn, S. Cho, and S. Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.

[2] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

[3] B. Athiwaratkun and K. Kang. Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313*, 2015.

[4] G. Bertasius, J. Shi, and L. Torresani. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3602–3610, 2016.

[5] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[6] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 2, 2015.

[7] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson. Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels. *IEEE transactions on geoscience and remote sensing*, 53(12):6663–6674, 2015.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[9] K. Grill-Spector and R. Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27:649–677, 2004.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[12] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2395, 2017.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] 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*, pages 3431–3440, 2015.

[16] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[17] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018.

[18] D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009.

[19] 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*, pages 234–241. Springer, 2015.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[21] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 41–48, 2014.

[22] J. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 756–763. IEEE, 2005.

[23] X. Xia and B. Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[24] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[25] L. Zhang, L. Zhang, D. Tao, and X. Huang. A modified stochastic neighbor embedding for multi-feature dimension reduction of remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 83:30–39, 2013.

[26] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao. Simultaneous spectral-spatial feature selection and extraction for hyperspectral images. *IEEE Transactions on Cybernetics*, 48(1):16–28, 2016.