MACHINE-ASSISTED ANNOTATION OF FORENSIC IMAGERY

Sara Mousavi[⋆] *Ramin Nabati*[⋆] *Megan Kleeschulte*[†] *Dawnie Steadman*[†] *Audris Mockus*[⋆]

[⋆] Department of Electrical Engineering and Computer Science
[†]Department of Anthropology
The University of Tennessee Knoxville, USA

ABSTRACT

Image collections, if critical aspects of image content are exposed, can spur research and practical applications in many domains. Supervised machine learning may be the only feasible way to annotate very large collections. However, leading approaches rely on large samples of completely and accurately annotated images. In the case of a large forensic collection that we are aiming to annotate, neither the complete annotation nor the large training samples can be feasibly produced. We, therefore, investigate ways to assist manual annotation efforts done by forensic experts. We present a method that can propose both images and areas within an image likely to contain desired classes. Evaluation of the method with human annotators showed highly accurate classification and reasonable segmentation accuracy that was strongly affected by transfer learning. We hope this effort can be helpful in other domains that require weak segmentation and have limited availability of qualified annotators.

Index Terms— Semantic segmentation, Proposed annotations, Pattern recognition, Forensic Imagery

1. INTRODUCTION

Certain image collections, such as images of human decomposition, represent high potential value to forensic research and law enforcement, yet are scarce, have restricted access, and are very difficult to utilize. To utilize such image collections, they need to be annotated with relevant forensic classes so that a user can find images with the desired content. This work is motivated by our attempt to annotate over one million photos taken over seven years in a facility focused on studying human decomposition.

Annotating images is a difficult task in general, with a single image taking from 19 minutes [1] to 1.5 hours [2] on average. Human decomposition images present additional difficulties. First, forensic data cannot be crowd-sourced due to its graphic nature and need for anonymity. Second, annotating forensic classes requires experts in human decomposition that are hard to come by. Therefore, it is natural to consider approaches to support such manual effort with machine learning (ML) techniques [3]. Unique challenges specific to forensic images prevent direct application of state-of-the art techniques described in, for example, [3]. This is mainly due to the primary focus of the annotation in being used by researchers, not algorithms. Particularly, when it comes to creating relevant training samples for ML approaches, we encountered the following challenges and discuss them afterwards:

- It is not feasible to annotate images completely. In other words, the user may choose to only annotate some instances of a class in an image, or only a subset of classes.
- The locations of forensically-relevant objects is not precisely outlined but, instead, roughly indicated via rectangular areas.
- It is not feasible to annotate a very large number of examples of a forensic class.

The first challenge results from the numerous instances of certain classes (for example, there may be tens of thousands of maggots in a single image, spread in multiple groups). Annotators may only tag classes relevant to their investigation or classes that they have sufficient expertise to identify accurately. The second challenge is caused by the primary objective of the annotator to provide indicators to other researchers and the need to maximize the number of manually annotated images irrespective of the ability of machine learning to generalize from them (i.e. using simple rectangles instead of more time-consuming masks). The last challenge is imposed by the limited availability of forensic experts. Furthermore, since it is not possible to annotate the entire set of images, the expert needs to choose which images to annotate. Choosing images randomly, as it turns out, is highly inefficient since such images rarely contain relevant forensic classes.

LabelMe [4] and similar polygon-drawing interfaces have been used to annotate image collections [2, 5, 6, 7]. The annotators need to manually select the areas of interest and label them with the correct label. Given the amount of time needed to annotate a single image, such approaches are not suitable for annotating one million forensic images.

Fluid Annotation [3] assists annotators in fully annotating images by providing initial annotations, that can be edited as needed. Fluid annotation uses Mask-RCNN [8] as the primary deep learning model. For Mask-RCNN and other deep-learning based techniques such as Deeplabv3+ and YOLO [9, 10] to work, large, complete, and clean training datasets such

Fig. 1: A sample image from ITS-HD. The image highlights the texture-like nature of the data. The image resolutions vary from 2400×1600 up to 4900×3200.

as Open Images, Image Net and COCO [11, 12, 1] are required. Such approaches without additional training do not work for a dataset with a complete different set of object classes. Our attempts to train Mask-RCNN on the photos of human decomposition had extremely poor performance (even with transfer learning) due to incomplete set of annotations, approximate bounding boxes, and relatively few labeled instances per class.

Other approaches to reduce the annotation effort involve using weakly annotated data, image-level or object- level annotations, for object detection [13, 14, 15] and semantic segmentation [16, 17, 18, 19]. Although these approaches have been successful to some extent, there is still a large performance gap between the models trained on full segmentation masks and those trained on image-level or object-level labels.

The main goal of this work is to simplify and speed-up the annotation process of forensic imagery by developing a machine-assisted semantic segmentation system called Proposed Annotations (PA) that recommends potential annotations to annotators to simply accept or decline.

Semantic segmentation needs a large training set. Our technique relies on the fact that human decomposition images are dominated by texture-like patterns (Figure 1) repeated throughout a class. Our method, therefore, can work with a simple classifier and small training data. It utilizes the classifier in combination with a region selection technique to produce potential annotations and presents them to expert annotators.

In addition, our approach can be used to estimate probabilities of a specific forensic class being present in un-annotated images. While this is possible with other semantic segmentation methods, it is of particular use in forensic data, where a major problem faces the annotator: how to design sampling strategies to select images for manual annotation from the collection of one million images.

Therefore, our contribution in this work is twofold. First, we present a novel semantic segmentation technique using a classifier and a region selector for forensic data, leveraging their pattern-like nature. Second, we use this method to propose not only new regions of interest for annotation, but also new images that are likely to contain classes of interest.

The rest of the paper is as follows. Section 2 details our method and implementation. Section 3 discusses the results of our work and the paper is concluded in Section 4.

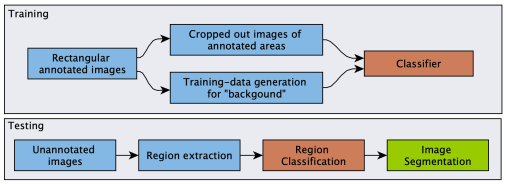

Fig. 2: An overview of the structure of PA. Blue, orange and green boxes represent data preparation, classification, and segmentation stages respectively.

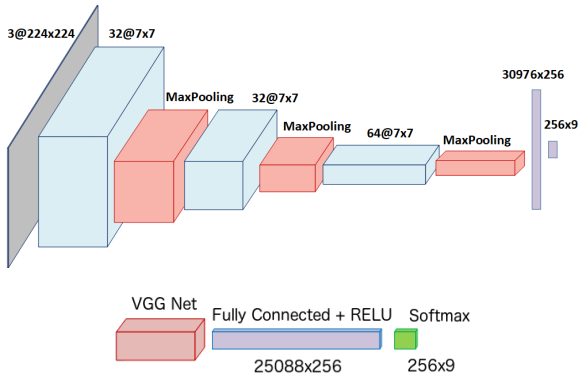

Fig. 3: The architecture of Model1 (top), and the VGG-based Model2 (bottom) is shown.

2. PROPOSED ANNOTATIONS (PA)

PA is comprised of a classifier and a region selection method. The classifier is trained on images that contain a single class. It is then used along with a region selection method to detect regions of new images. The classified regions are then merged into larger segments resulting in semantic segmentation. An overview of this process is shown in Figure 2.

This process has three main steps: data preparation (Section 2.2), classification (Section 2.3) and semantic segmentation (Section 2.4).

2.1. Human Decomposition Dataset

Our image collection includes photos that depict decomposing corpses donated to the Forensic Anthropology Center at the University of Tennessee. The photos are taken periodically from various angles to show the different stages of body decomposition. The collection spans from 2011 to 2016, and has over one million images. We call this dataset ITS-HD: Images Tracking Stages of Human Decomposition.

The annotation for a small subset of this dataset has been done manually by four forensic experts resulting in 2865 annotated images. However, as previously mentioned, these images are not fully annotated.

A sample image from ITS-HD is shown in Figure 1. The cadaver is mostly camouflaged in the background patterns.

The forensic classes used in this work along with the number of annotated instances for each, are shown in Table 1.

Table 1: An overview of the forensic classes of ITS-HD used in this work. The number of annotated instances is shown.

CLASS NAME	#SAMPLES	CLASS NAME	#SAMPLES
MAGGOTS	1375	EGGS	533
SCALE	716	MOLD	339
PURGE	709	MARBLING	241
MUMMIFICATION	557	PLASTIC	107

2.1.1. Manual annotation

To enable manual annotation of the small subset, we built an online platform that allows browsing, querying, and annotating ITS-HD. The annotator starts by first selecting a rectangular bounding box around the region of interest and then enters the appropriate class name in an input dialog. The bounding boxes' coordinates along with the class names are stored in a database.

2.2. Data Preparation

Preparing training data is a crucial step for making a highly accurate classifier. Due to the similarity of some forensic classes to the background, both in terms of color and texture, we added an additional class to the actual forensic classes for "background". We then cropped areas designated as the forensic classes from the annotated images and used the class name to label each cropped section. Therefore, each annotation became a new training image by itself. For the images cropped for "background", in order to create a diverse range of training data, we used a sliding window to extract smaller images from each training image. We re-sized all images to 224*224 and, as is commonly done, we also generated additional training data from the existing annotations using data augmentation.

2.3. Classification

We used a CNN with a multinomial logistic regression classifier to train a model for classifying regions of the un-annotated images. The preponderance of texture-rich classes did not call for very deep neural networks. We started with Model1 that uses a simple neural network shown in Figure 3:top. The CNN network in this model has three convolutional and two fully connected layers. We used normalization after each layer and also a drop-out of 0.5 before the last layer. In addition to Model1, we also experimented with Model2 , a standard VGG [20] with two fully connected layers added on top (Figure 3:bottom). Images generated from section 2.2 were used to train and validate these two models. We trained Model1 from scratch. However for Model2, we tested both pre-trained weights obtained from ImageNet as well as random weights.

2.4. Semantic Segmentation

Locating the forensic objects within images is done using the classifier described in section 2.3. Algorithm 1 shows how semantic segmentation is done in PA. Regions of un-annotated images are fed into the classifier model to be classified. The regions are generated by sliding a window of size 224*224 with a stride of 200. Since the training data is not fully annotated, many regions within an image may contain classes that the classifier has not been trained on. To reduce the number of such false positives, we use a threshold of 0.85 to accept a classification done on a region, otherwise it will be ignored.

The contiguous classified regions of the images need to be organized so that neighbor regions belonging to the same class are proposed as a single composite segment. To do so, we group the classified regions by first finding overlaps. Then, we create an adjacency matrix of size $n \times n$ where n is the number of regions for the class. A cell (i, j) (for two regions i and j) is set to 1 if the two regions overlap. We then create a graph from the adjacency matrix and find the connected components of the graph for each class using the iGraph library [21]. Next, we find the convex hull for each connected component. The resulting hulls are presented to the annotator as proposed annotations. The confidence of a recommended annotation is calculated based on the average confidence of the individual regions within that component.

Algorithm 1 Semantic segmentation in our method

```
procedure SEGMENT(image)
    for every region in image do
        CLASSIFY(region)
        Store region's coordinate, class_id and confidence
    end for
    for every c in classes do
        Find all regions classified in class c
        Create an adjacency matrix of regions
        Create connected-components to group neighboring regions
        Draw the convex hull of each group
        Calculate score for each colored area
    end for
    Present the segmentation as proposals to the annotator
end procedure
```

3. RESULTS AND DISCUSSION

To evaluate PA, we measure the accuracy in comparison to the manual annotation done by a forensic expert. The results include the performance of Model1 and Model2. We also tested the effects of including the background as a separate class in both models and also the effect of transfer learning on Model2. Section 3.1 describes tuning parameters for both models and evaluation setup. Section 3.2 discusses our findings.

3.1. Evaluation Setup

PA is implemented using Keras, TensorFlow and Python. We used MongoDB as our database. For both CNN networks we used the SGD optimizer with a learning rate of 0.001.

Over two hundred distinct classes of samples were present in the dataset. To select a more manageable number of classes for the experiments, we first excluded classes with fewer than 100 ground truth instances and asked forensic experts to select the most important classes for the forensic community. We used one third of images per class for validation and the remaining images for training.

a: Original image b: Ground-truth c: Model1's result d: Model2's result

Fig. 4: Detected forensic classes using Model1-bg and Model2-bg-tl are shown in (c) and (d) respectively. Sub-figures (a), and (b) show the original image and the ground truth respectively. Concave annotations are a result of overlaps between two convex hulls where one overlays part of the other.

To evaluate the performance of our PA, we randomly selected 46 images and asked a forensic expert to provide us with the ground truth annotation masks only for the forensic classes used in this work. These images were annotated carefully and completely with polygonal selections, taking about 3 hours to complete. We evaluated the performance of our proposed annotations against these round truths.

3.2. Discussion

Table 2 shows the performance of PA. We calculated mean average precision (mAP) for the classification done by both Model1 and Model2 over all classes. We also calculated mean average recall and precision over all classes (mAR, mAP) for our semantic segmentation against the ground truth. These values are used as mAP and mAR in Table 2.

The mean average precision is calculated as the ratio of correct predicted pixels over the total predicted pixels for each class. This value is then averaged for each class over all 46 images. We used a similar method for mAR, however we used the the ratio of correct predicted pixels to the total ground truth pixels for each class.

Table 2 shows that transfer learning improves the performance of Model2. Comparing Model2 with Model2-tl, we can see that transfer learning has improved both mAP of the classifier model and mAR of the semantic segmentation.

Comparing Model2 with Model1 in Table 2, we believe that we might get even better results using Model1 if we first train it on another dataset such as ImageNet, considering the fact that Model1 is a very simple model and its training takes less time compared to Model2.

A trade off between using a model with high recall or high precision can also be observed from the table. For the purpose of suggesting classes to a human annotator, it is more important to detect a forensic class if it exists, as opposed to exactly pinpointing the location of the class within the image. Thus, we want to have a model with higher recall and a reasonable mAP. Our results also show that including the background as a class improves mAP for segmentation.

Figure 4 shows a segmentation using Model1 without transfer learning and Model2 with transfer learning, and compares it to the ground truth. Both models were trained on 8 forensic classes plus the background class. We can see that a better segmentation is obtained when transfer learning is employed.

Table 2: Performance of classifier models and semantic segmentation in PA. bg and tl stand for background and transfer learning.

Method	Semantic Segmentation		Classification
	mAP	mAR	mAP
Model2-bg-tl	0.26	0.45	0.95
Model2-tl	0.15	0.59	0.92
Model2	0.30	0.28	0.79
Model1	0.16	0.32	0.84
Model1-bg	0.17	0.23	0.88

4. CONCLUSION

In this work, we discuss an annotation-assistance system that proposes annotations within an image as well as images likely to contain a desired class to forensic experts. At the core of our system we introduce a semantic segmentation method composed of a classifier in conjunction with a sliding-window-based region selection method. We also evaluate its applicability in the context of imagery documenting human decomposition where classes are primarily determined by patterns. We demonstrate the feasibility of semantic segmentation in this domain using a relatively small set of training samples. As is expected with small training samples, transfer learning has been effective. Inclusion of the background as a class also brought improvements, possibly because background is at times difficult to distinguish from focal classes.

In the future, we would like to evaluate if our method would work with other types of texture-like data. In addition, we plan to utilizing body pose detection methods to improve the ability to exclude background and increase the accuracy of our system for forensic class segmentation.

Acknowledgements

This work was supported by National Institute of Justice Awards 2016-DN-BX-0179 and 2018-DU-BX-0181.

5. REFERENCES

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari, "Coco-stuff: Thing and stuff classes in context," *CoRR,*

abs/1612.03716, vol. 5, pp. 8, 2016.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari, "Fluid annotation: a human-machine collaboration interface for full image annotation," *arXiv preprint arXiv:1806.07527*, 2018.

[4] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William 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.

[5] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891–898.

[6] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva, "Sun database: Exploring a large collection of scene categories," *International Journal of Computer Vision*, vol. 119, no. 1, pp. 3–22, 2016.

[7] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, vol. 1, p. 4.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan 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.

[10] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al., "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv preprint arXiv:1811.00982*, 2018.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[13] Hakan Bilen and Andrea Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[14] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Multi-fold mil training for weakly supervised object localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2409–2416.

[15] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari, "Localizing objects while learning their appearance," in *European conference on computer vision*. Springer, 2010, pp. 452–466.

[16] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.

[17] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1796–1804.

[18] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele, "Simple does it: Weakly supervised instance and semantic segmentation.," in *CVPR*, 2017, vol. 1, p. 3.

[19] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 616–625.

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

[21] Gabor Csardi and Tamas Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, pp. 1695, 2006.