Segmenting Technical Drawing Figures in US Patents

Md Reshad Ul Hoque,[1] Xin Wei,[2] Muntabir Hasan Choudhury,[2] Kehinde Ajayi,[2] Martin Gryder,[2] Jian Wu,[2] Diane Oyen[3]

[1]Electrical and Computer Engineering, Old Dominion University
{mhoqu001}@odu.edu
[2]Computer Science, Old Dominion University
{xwei001,mchou001,kajay001,j1wu}@odu.edu
[2]Los Alamos National Laboratory
doyen@lanl.gov

Abstract

Image segmentation is the core computer vision problem for identifying objects within a scene. Segmentation is a challenging task because the prediction for each pixel label requires contextual information. Most recent research deals with the segmentation of natural images rather than drawings. However, there is very little research on sketched image segmentation. In this study, we introduce heuristic (point-shooting) and deep learning-based methods (U-Net, HR-Net, MedT, DETR) to segment technical drawings in US patent documents. Our proposed methods on the US Patent dataset achieved over 90% accuracy where transformer performs well with 97% segmentation accuracy, which is promising and computationally efficient. Our source codes and datasets are available at https://github.com/reshadshuvo123/Sketched-Image-Segmentation

Introduction

Combining information contained in text and images is an important aspect of understanding scientific documents. However, patents and scientific documents often contain compound figures containing subfigures, each having its own label, caption, and reference text. To associate individual subfigures with the appropriate caption and reference text, we must first segment the full figure into its individual subfigures. Although much research has been done on figure understanding and extraction for scientific documents, existing methods rely on either (1) manually-designed rules and human-crafted features which do not generalize well for new dataset (Taschwer and Marques 2018); or (2) machine learning approaches most of which were trained on natural images (Tsutsui and Crandall 2017). We demonstrate that we cannot simply apply approaches developed for other datasets to patent drawings, and develop a novel approach while addressing questions about how to extend existing methods to novel datasets.

Image segmentation has been extensively studied with rule-based methods such as watershed, and machine learning methods applied to natural images (Bai and Urtasun 2017; Ren and Zemel 2017; Minaee et al. 2021). In patent drawings, there are usually white space between individual drawings. A simple sweeping line method, which detect boundaries of subfigures by counting the maximum number of black-pixels along a horizontal (or vertical) pixel array, e.g., Rane et al. (2021), does not work well because (1) other components such as figure labels may be present, and (2) one figure may contain multiple disconnected parts. There are few existing papers on segmenting technical drawings in patent documents. Most existing tools were developed for extracting figures in research papers. For example, Clark and Divvala (2016) developed a framework that extracts figures from scientific papers in PDF format. Viziometrics (Lee, West, and Howe 2017), a figure-oriented literature mining system, was developed which works on certain pattern figures. These tools could not be used for segmenting compound figures. For compound figure separation, background color, layout patterns, spaces and lines between sub figures were used as important cues for rule-based methods (Taschwer and Marques 2018). Tsutsui el. at. developed a data driven deep learning model to segment compound figures (Tsutsui and Crandall 2017). They fine-tuned the pre-trained YOLO-2 model to segment compound images on ImageCLEF Medical dataset.

In this paper, we report our preliminary work on automatically segmenting scientific figures appearing in patent documents, focusing on technical drawings. We propose a heuristic model and compare it with the state-of-the-art convolutional neural network (CNN) based models, including U-Net, HR-Net, and transformer-based models, including MedT and DETR. The method we propose, called "point-shooting" correctly segments over 92.5% of the patent figures (compound and single). We perform a comparative study between the point-shooting method and the state-of-the-art deep learning methods on a benchmark dataset. The transformer-based model, called MedT, fine-tuned on a small set of training samples, works the best with high accuracy (97%) and efficiency. The model is also computationally efficient compared with other methods. We release the benchmark dataset that can be used for future work on the task of segmenting technical drawings.

Data

The data for this project is obtained from the United States Patent and Trademark Office (USPTO). The ground truth dataset is developed on a corpus of 500 randomly selected

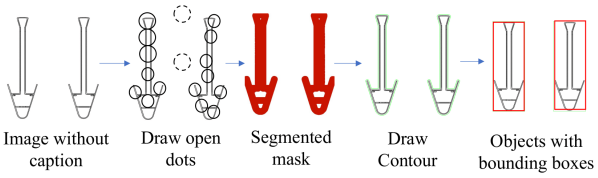

Figure 1: Point-shooting method.

figures from the *design* category of patent. The dataset consists of 20 figure files with single drawings and 480 figure files, which containing at least two sub-figures. We preprocess each figure to remove text labels. The number of subfigures in each figure file is inferred by the number of text labels detected identifying subfigures so segmentation is only necessary for figures containing multiple subfigures.

We use VGG Image Annotator (VIA) to annotate our dataset. VIA is a manual annotation open source software for annotating images, videos, and audio. We draw rectangles bounding boxes around subfigures. Each figure consists of 2–12 subfigures. We also performed an independent human verification to ensure the bounding boxes were drawn correctly. VIA allows exporting annotation results including filename, file size, region count (e.g., number of the bounding boxes for each figure in an image), region id, and coordinates of bounding boxes.

Segmentation Methods

Point Shooting Method

We propose a heuristic method for segmenting figures containing technical drawings. We call it point-shooting because it mimics the shooting of darts onto a dartboard. The goal is to draw bounding boxes around individual subfigures on a figure containing multiple technical drawings.

Figure 1 illustrates the procedures of this method. After removing the figure labels, we randomly pick a pixel in the figure , and draw an open dot of a radius r. For our experiment, we chose an empirical value $r = 2$. If a black pixel *in the original figure* was detected inside the open dot, the dot retrained. Otherwise, the dot was removed. We constrain the circle centers so they do not fall outside the figure boundary. We then fill all retained circles and draw contours[1]. Using the contour information, we draw rectangular bounding boxes to segment a figure.

Deep Learning Methods

The point-shooting method is easy to implement and successfully segments most figures containing multiple technical drawings. However, the method does not generalize well for certain figures in our dataset. One example is shown in Figure 2.

Therefore, we consider applying deep learning-based methods including U-Net (Ronneberger, Fischer, and Brox 2015), HR-Net (Wang et al. 2020) and transformer models (MedT (Dosovitskiy et al. 2020), DETR (Carion et al. 2020)). One challenge is that the ground truth only contains bounding boxes, while these models produce pixel-level masks. Therefore, the ground truth cannot be directly used for training these deep learning models. To overcome this challenge, we first use the point-shooting method to generate masks for training figures and use them as input to train U-net, HR-net, and fine-tune MedT and DETR. It is worth mentioning that the point-shooting method achieves an accuracy of 92.5%. Although the result output by the point-shooting method is not 100% accurate, we hope that the neural networks can still encode and capture the right features and achieve better generalization for reasonably good performance. Figure 3 shows the deep learning segmentation pipeline. All of our deep learning-based models except DETR (Carion et al. 2020) are semantic segmentation models where the models produce foreground-background masks on the input image. DETR is a transformer based end-to-end object detection model that directly predicts the bounding boxes on the input image.

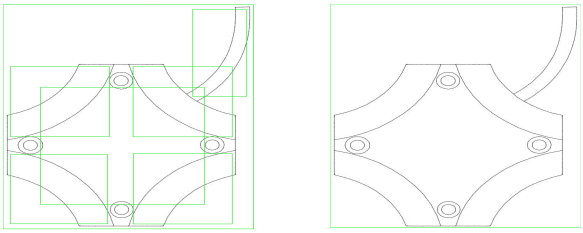

Figure 2: An example of the point-shooting methods failure. [Left] Point-shooting creates wrong bounding boxes. [Right] Deep learning model (U-Net) produce correct bounding box. All other deep learning models are successful for this figure.

To reduce unnecessary computational cost, we rescale the resolution of the input figure to 128×128 and use them to train the deep learning models. The model produces the segmentation mask with a dimension of $128 \times 128 \times 3$ which we use to draw contours and then bounding boxes around contours. After obtaining bounding boxes from the low-resolution image, we linearly scale up the predicted bounding boxes to fit the original figure.

U-Net: The architecture of U-Net consists of a contracting path and an expanding path. The contracting path is a typical convolutional network containing a series of convolutional layers, each followed by a rectified linear unit (ReLU) and a max pooling layer with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. In the expanding path, each step consists of an upsampling of the feature map followed by an "up-convolution", a concatenation with cropped feature map from a contracting path, and two convolutions, each followed by a ReLU.

HR-Net: In the contracting path of U-Net, feature maps are downsampled to the lower resolution using polling and later up-sampled in the decoder part. In this process, high-resolution information is lost. Although skip connections are used to copy the high-resolution information to the expansive path. They can not fully recover high-resolution infor-

[1]https://learnopencv.com/contour-detection-using-opencv-python-c/

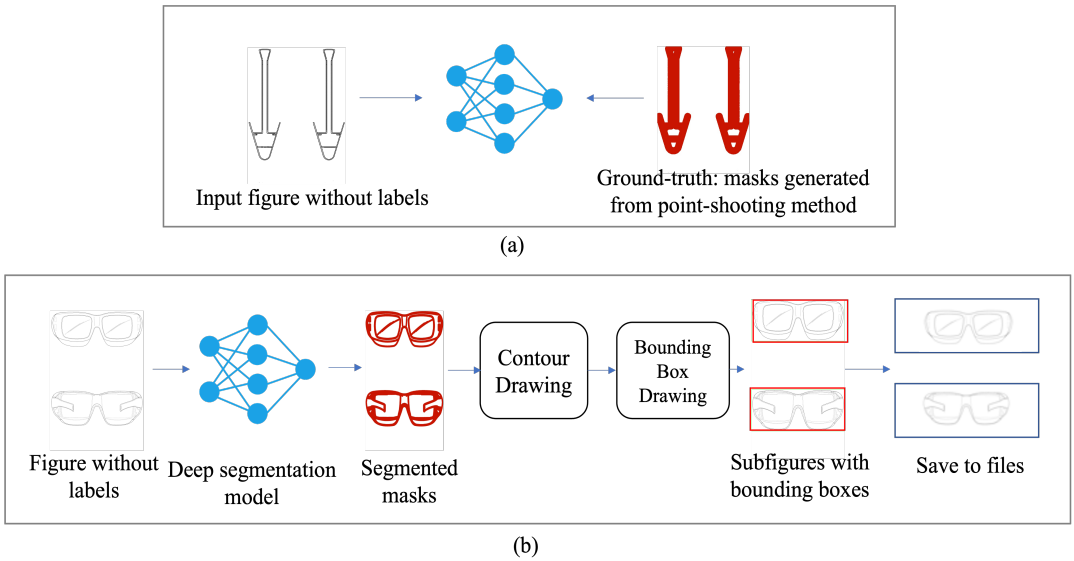

Figure 3: The training (a) and testing (b) of deep learning-based segmentation models.

mation. To overcome this drawback, we apply the HR-Net model which retains both high and low resolution information throughout the training process. The preserved information may be useful to reconstruct the segmentation mask. We simplified the original HR-Net, which contains three resolution channels, each capturing high, mid, and low resolution information, respectively. The three channels contain five, three, and two convolutional blocks, respectively. Each convolutional block contains two convolutional layers followed by a batch normalization layer, and a ReLU activation layer. The resolution gap between two channels is 2.

Transformer Although the CNN-based models have shown impressive performance on the segmentation tasks (Ronneberger, Fischer, and Brox 2015), they can not capture the long-range dependencies between pixels due to inherent inductive biases (Dosovitskiy et al. 2020). Transformers have significantly improved many fundamental natural language processing tasks. The novel idea behind the success is "Self Attention" (Vaswani et al. 2017). This mechanism automatically weights more on more important features and can capture the long-range dependencies. The computer vision domain has borrowed this idea to improve vision-related tasks. We consider two transformer-based models.

MedT: The core component of MedT is a gated position-sensitive axial attention mechanism designed for small size datasets (Valanarasu et al. 2021) . Gated control axial attention which introduces an additional control mechanism in the self-attention module is used to train a transformer on a small dataset. These mechanisms control the influence of the relative positional encoding on non-local context. This architecture contains two branches, including a global branch that captures the dependencies between pixels and the entire image and a local branch that captures finer dependencies among neighbouring pixels.

The training figures are passed through a convolution block before passing through the global branch. The same figure is broken down into patches and sent through a similar convolution block before passing through the local branch sequentially. A re-sampler aggregates the outputs from the local branch based on the position of the patch and generates output feature maps. Outputs from both branches are add together followed by a 1×1 convolutional layer to pool these output feature maps into a segmentation mask.

DETR: DETR is an end-to-end object detection transformer model (Carion et al. 2020). The architecture is simple and does not require specialized layer or a custom function (such as the non-maximum suppression function) for predicting the bounding boxes. The original DETR model predicts 80 classes of bounding boxes. We fine-tuned this model and directly predicts the bounding boxes of subfigures given a compound figure.

Results and Discussion

The segmentation task can be seen as a classification problem, in which individual subfigures are foreground objects, and the blank area between subfigures is the background. Although we use a training corpus with noisy labels, the deep learning models successfully capture latent representations and correctly segmented individual drawings. The evaluation results are shown in Table 1. Visual comparisons of segmentation results of our models are also available on this link[2].

The performance of each model is measured using the accuracy, which is calculated as the fraction of subfigures that are correctly segmented. We set aside 200 figures for evaluation. We use Intersection over Union (IOU), which compares overlaps between the predicted bounding boxes with the ground truth bounding boxes. The segmentation is deter-

[2]https://bit.ly/3oF21TF

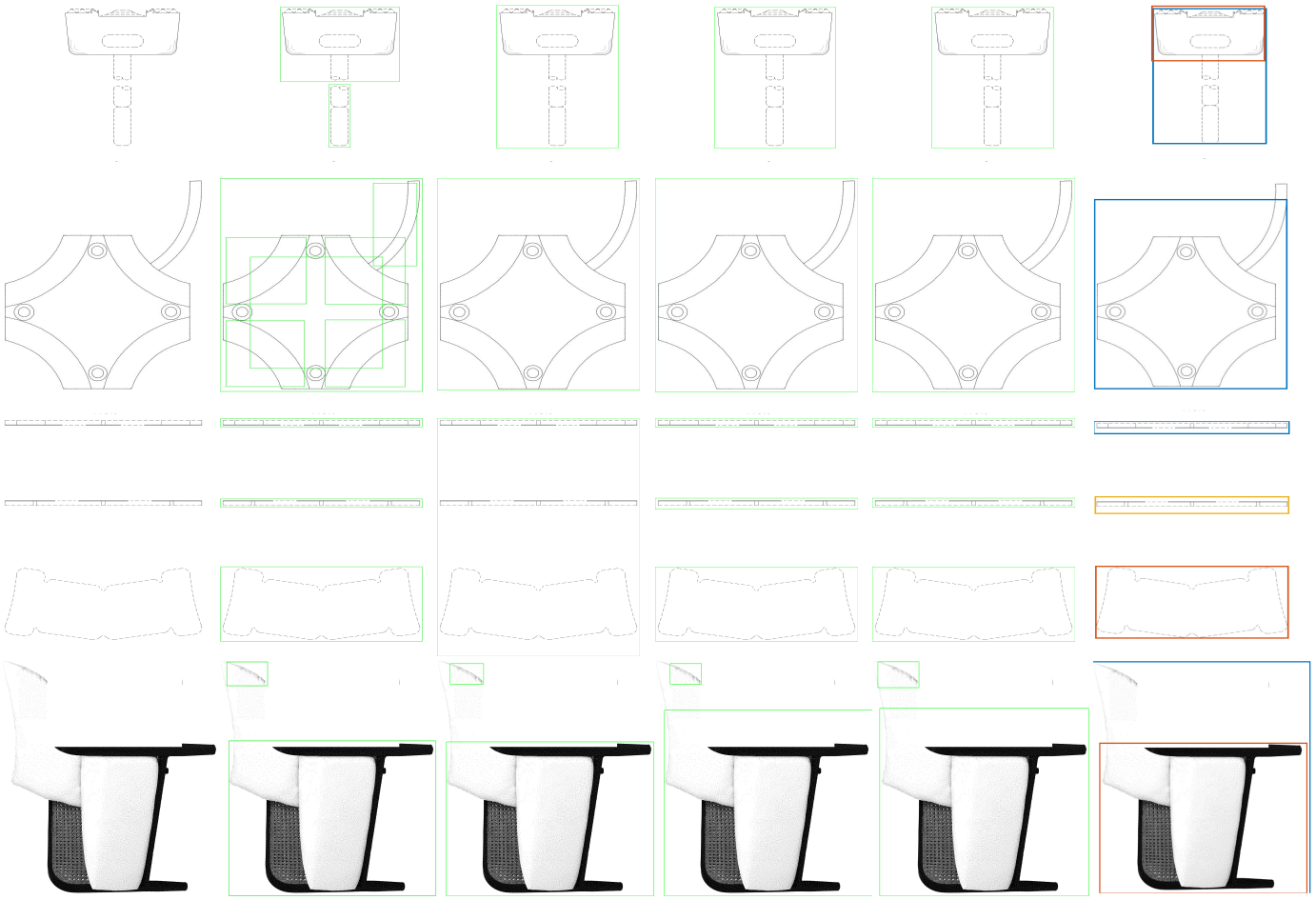

Figure 4: Models performance when applied to the test (Image without caption). From left to right, the first columns shows input images , 2nd columns shows point-shooting output, 3rd, 4th for U-Net, HR-Net. 5th and 6th columns are for the Transformer methods i.e. MedT and DETR respectively.

Figure 5: A figure with a single subfigure, at which all models failed. From left to right: input figure, followed by results of point-shooting, U-Net, HR-Net, MedT, and DETR.

mined correct if IOU is greater than an empirical threshold of 0.7. To verify consistency, we also perform qualitative evaluation by visually inspecting predicted and ground truth segmentations. The manual inspection is consistent with automatic inspection with an agreement rate of 98%.

In general, deep learning-based methods perform better than point-shooting methods, except for DETR. Figure 2 shows a case in which deep learning method succeeded but point-shooting failed. In certain cases, the point-shooting method produced the correct segmentation map but deep learning-based methods failed. There are a few challenging cases, in which all methods failed (Figure 5). This occurred when a subfigure contains isolated fragments, which were treated as individual objects.

Table 1: Segmentation model evaluation. Each model was timed on segmenting 200 figures. Runtime is in seconds.

Models	Automatic [1]	Manual [2]	Runtime
Point-shooting	92.5%	92.5%	1035
U-Net	90.5%	91.5%	15
HR-Net	96.0%	96.5%	18
MedT	**97.0%**	**97.0%**	**29**
DETR	90.0%	91.0%	1396

[1] Automatic Evaluation Accuracy.
[2] Manual Verification Accuracy.

In conclusion, we compared heuristic and deep learning methods on the task of segmenting technical drawings in US patents. Both heuristic and deep learning-based models achieve over 90% accuracy. Interestingly, though we trained using data containing noisy labels, generated using the point shooting method, the deep learning models still captured the right features and outperformed the point shooting method. The CNN-based model (e.g., HR-Net) under-performs the transformer model by a small margin. We attribute this to the gated attention mechanism in the transformer model, which captured the long-range relations between pixels.

References

Bai, M.; and Urtasun, R. 2017. Deep watershed transform for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Clark, C.; and Divvala, S. 2016. PDFFigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 143–152. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Lee, P.-s.; West, J. D.; and Howe, B. 2017. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 4(1): 117–129.

Minaee, S.; Boykov, Y. Y.; Porikli, F.; Plaza, A. J.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rane, C.; Subramanya, S. M.; Endluri, D. S.; Wu, J.; and Giles, C. L. 2021. ChartReader: Automatic Parsing of Bar-Plots. In *22nd International Conference on Information Reuse and Integration for Data Science, IRI 2021, Virtual*. IEEE.

Ren, M.; and Zemel, R. S. 2017. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6656–6664.

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

Taschwer, M.; and Marques, O. 2018. Automatic separation of compound figures in scientific articles. *Multimedia Tools and Applications*, 77(1): 519–548.

Tsutsui, S.; and Crandall, D. J. 2017. A data driven approach for compound figure separation using convolutional neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 533–540. IEEE.

Valanarasu, J. M. J.; Oza, P.; Hacihaliloglu, I.; and Patel, V. M. 2021. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.