JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES
VOL. 2, ISSUE 2, PP. 22-32, DECEMBER 2021.
Available at: www.journal.research.fanap.com
DOI: https://10.22034/JAISIS.2021.307161.1035

AUTOMATIC IMAGE CROPPING AND SEMANTIC OBJECT SELECTION

Serveh Kakaei [1,*]

[1] *Department of Industrial Engineering, University of Kurdistan, Sanandaj, Iran*

ABSTRACT

Automatic image cropping is an important method for changing the visual quality of digital photos without using to tedious manual selection. Operation of cropping is common to the photographic, graphic design, film developing. At the process of cropping undesired regions cuts away. The performance of an intelligent image cropper highly depends on the ability to detect objects and speed in cutting operation. A lot of methods have been proposed to automate the cropping operation but the most important thing in object cropping is to identify the objects in the image. There are many techniques to detect objects in an image like object detection, image processing and image segmentation. Image segmentation divides a digital image into multiple segments and in this process one label assigns to every pixel in an image. The goal of segmentation is to simplify change the representation of an image into something that is more meaningful and easier to analyze. In this paper, ahistory of development in image segmentation was investigated based on discontinuity and similarity detection-based approaches. We investigated advantage and disadvantage of image segmentation and we studied new models in semantic segmentation and interactive object segmentation.

***KEYWORDS*:** Image Cropping, Object Selection, Image Segmentation, Semantic Segmentation

1. INTRODUCTION

One of the important tools for improving visual quality of digital photos is cropping. At the process of cropping undesired regions cuts away. Operation of cropping is common to the photographic, graphic design, film developing. In addition to helping the editors to save a lot of time, an automatic and efficient image cropping algorithm give them professional advice. A lot of methods have been proposed to automate the cropping operation. Generally, automatic image cropping techniques can be examined in two categories: attention-based and aesthetics-based approaches. The basis of the attention-based method is to place the crop window in an image over the most significant regions according to certain attention scores, e.g. saliency map (Suh et al., 2003; Stentiford, 2007), or by resorting to eye tracker (Santella et al., 2006), face detector (Zhang et al., 2005) to find the regions of interest. The main goal of aesthetics-based is to imitate human interpretation of the beauty of natural images. For a more detailed survey, the reader is referred to (Deng et al., 2017). Aesthetic quality analysis is viewed as a binary classification problem of predicting high- and low-quality images (Datta et al., 2006; Dhar et al., 2011; Luo et al., 2011). Extracting visual features and then employing various machine learning algorithms to predict photo aesthetic values is a common pipeline in this research area. Researchers design various features to capture the aesthetic properties of an image compliant with photographic rules or practices, such as the rule of thirds and visual balance. Yan et al. (2013) employed features which is designed

* *Corresponding Author, Email*: s.kakaei@eng.uok.ac.ir

RECEIVED: 10 SEPTEMBER, 2021; **ACCEPTED**: 7 NOVEMBER, 2021; **PUBLISHED ONLINE**: 9 NOVEMBER, 2021

Table 1. Well-known Methods of Automation Cropping

Method	Underlying Algorithms	Researchers
Attention-based Method	Saliency Map	Suh et al., 2003
		Stentiford, 2007
	Resorting to eye tracker	Santella et al., 2006
	Face Detector	Zhang et al. , 2005
Aesthetics-based Method	Binary Classification	Datta et al., 2006
		Dhar et al., 2011
		Luo et al., 2011
	Extract Feature	Yan et al., 2013
		Chen et al., 2010
		Su et al., 2012
Attention-based Method and Aesthetics-based Method	Neural Network Algorithm	Wang and Shen, 2017

to capture the changes between the original and cropped images. Some related researches get view finding by learning aesthetic features based on position relationships between regions (Cheng et al., 2010) or image decomposition (Su et al., 2012). Both attention and aesthetics approaches have been considered by other researchers (Wang & Shen, 2017). Based on deep learning, the image cropping problem was modeled like a cascade of attention box regression and aesthetic quality classification. Table 1 summarizes the most important research works in Automation Cropping.

There are several existing techniques which are used for object selection in images. One of these techniques is image segmentation which is partition a digital image into multiple segments and in this process one label assigns to every pixel in an image. The rest of the paper we investigate new models in image segmentation in order to more accurate detection of objects in the image. Section 2 presented about object selection techniques. The image segmentation is presented in detail and different segmentation techniques have been compared in Section 3 and conclusion is drawn in Section 4.

2. Overview of Objects Selection Techniques

How to select and crop part of a photo is important for graphic designers, so that they spend less time on that work. Hence, object selection is the most important step. Selection tools help to quickly select and easily edit images. There are several existing techniques which are used for object selection in images. These all techniques have their own importance. A common approach to select the objects in images is to require the user to provide mouse input to select the desired objects. The Magic Wand is an automatic selection tool which select pixels based on tone and color. Unlike other selection tools that select pixels in an image based on shapes or by detecting object edges. The Magic Wand tool works better at detecting objects of different colors (INCORP, 2002). Another method is Intelligent Scissors (Mortensen & Barrett, 1995) which is used for image segmentation and composition. Intelligent Scissors allow objects within digital images to be extracted quickly and accurately using simple gesture motions with a mouse. Intelligent Scissors works better to select a region defined by strong color changes at the edges. To select the desire region, the Scissors create some points and a continuous curve passing around the edges of the region. The path which is create by the tool may correspond

to the selected region. Another method is Graph Cut which is typically require the user to stroke over the object and background (Boykov & Jolly, 2001). They used the cost function that is general enough to include both region and boundary properties of segments. Consider $\mathcal{P}$ as an arbitrary set of data elements, $\mathcal{N}$ as some neighborhood system and {p, q} pairs of neighboring elements in $\mathcal{P}$. Given $A = (A_1, \dots, A_P, \dots, A_{|\mathcal{P}|})$ as abinary vector whose A_p determine assignments to pixels p in $\mathcal{P}$ and A_P specify either "object" or "background". Vector A is a segmentation and E(A) is the cost function.

$$\text{Min} \quad E(A) = \lambda . R(A) + B(A) \tag{1}$$

S.t:

$$R(A) = \sum_{p \epsilon \mathcal{P}} R_p(A_p) \tag{2}$$

$$B(A) = \sum_{\{p,q\} \in \mathcal{N}} B_{\{p,q\}} . \delta(A_p, A_q) \tag{3}$$

And

$$\begin{cases} \delta(A_p, A_q) = 1 \; if \quad A_p \neq A_q \\ \delta(A_p, A_q) = 0 \; \; Otherwise \end{cases} \tag{4}$$

The coefficient $\lambda \geq 0$ in (1) specifies a relative importance of the region properties term R(A) versus the boundary properties term B(A). The regional term R(A) assumes that the the individual penalties for assigning pixel p to "object" and "background", correspondingly Rp("object") and Rp("background"), are given. The goal is to compute the global minimum of (1) among all segmentations. Graph cutting algorithm is one of sub-branches combinatorial graph theory. Many researchers have applied this method to image and video segmentation and achieved good results.

GrabCut is another image segmentation method which is based on graph cuts. With draw a bounding box around the desired object to be segmented, the algorithm estimates the color distribution of the specified object and that of the background using a Gaussian mixture model. This is used to construct a Markov random field over the pixel labels, with an energy function that prefers connected regions having the same label, and running a graph cut based optimization to infer their values. As this estimate is likely to be more accurate than the original, taken from the bounding box, this two-step procedure is repeated until convergence. Deep neural networks are very effective in semantic segmentation, that is labeling each region or pixel with a class of objects/non-objects. Semantic segmentation plays an important role in image understanding and essential for image analysis tasks. Deep learning is a new field division of machine learning, which is rapidly growing with the pace making it very difficult to stay up to date, even to keep track of the works dealing with semantic segmentation. These works cover the development of new methods, improvements of existing methods, and their deployment in new application domains. A Survey by Zhu et al. (2016) covering a wide range of the papers and areas of semantic segmentation topics including, interactive methods, recent development in the super-pixel, object proposals, semantic image parsing, image co-segmentation, semi & weakly supervised, and fully supervised image segmentation. Deep learning based interactive segmentation methods have achieved remarkable performance with only a few user clicks. Long et al. proposed one of the first deep learning works for semantic image segmentation, using a fully convolutional network (Long et al., 2015). Chen et al. proposed a semantic segmentation algorithm based on the combination of convolutional neural networks and fully connected (Chen et al., 2014).

Another popular family of deep models for image segmentation is based on the convolutional encoder-decoder architecture. Most of the DL-based segmentation works use encoder-decoder models. We group these works into two categories, encoder-decoder models for general segmentation, and for medical image segmentation. Noh et al. published an early paper on semantic segmentation based on deconvolution. Their model consists of two parts, an encoder using convolutional layers adopted from the VGG 16-layer network and a deconvolutional network that takes the feature vector as input and generates a map of pixel-wise class probabilities. The deconvolution network is composed of deconvolution and unpooling layers, which identify

pixel-wise class labels and predict segmentation masks (Noh et al., 2015). In another promising work known as SegNet, Badrinarayanan et al. proposed a convolutional encoder-decoder architecture for image segmentation (Badrinarayanan et al., 2017).

A semantic segmentation model was proposed by Liang et al. based on the Graph Long Short-Term Memory (Graph LSTM) network which is a generalization of LSTM from sequential data or multidimensional data to general graph-structured data (Liang et al., 2016). Using a combination of CNN which encodes the image and LSTM which encodes its natural language description (Hu et al., 2016) a semantic segmentation algorithm was developed by Hu et al. on the basis of natural language expression. The Pyramid Scene Parsing Network which is a multi-scale network to better learn the global context representation of a scene was developed by Zhao et al. Using a residual network as a feature extractor, with a dilated network (Zhao et al., 2017) the input image provides various patterns. based on a Laplacian pyramid which employs skip connections from higher resolution feature maps and multiplicative gating to successively refine segment boundaries reconstructed from lower-resolution maps, a multi-resolution reconstruction architecture was developed by Ghiasi and Fowlkers (Ghiasi & Fowlkes, 2016). a deep convolutional neural network model named DeepLab was proposed by Chen et al. from Google. The proposed method which is combined with fully connected conditional random fields and is (Chen et al., 2017). another interesting segmentation problem with increasing popularity is panoptic segmentation, and there are already several interesting works in the field, including Panoptic Feature Pyramid Network (Kirillov et al., 2019), attention-guided network for Panoptic segmentation (Li et al., 2019), Seamless Scene Segmentation (Porzi et al., 2019), panoptic deeplab (Cheng et al., 2019), unified panoptic segmentation network (Xiong et al., 2019), efficient panoptic segmentation.

The Faster R-CNN architecture which employs a region proposal network in order to present bounding box candidates, is developed to detect objects. In fact, other extensions of the regional convolutional network (R-CNN) such as Fast R-CNN, Faster R-CNN, Maksed-RCNN as well as the regional convolutional network itself have all been successful in object detection applications. In order to extract the bounding box coordinates and the class of the object, the features from these proposals are computed by a RoIPool and a Region of Interest (RoI) itself is extracted by the region proposal network. In order to focus on the instance segmentation problem i.e. the task of simultaneously performing object detection and semantic segmentation (Ren et al., 2015) some of R-CNN extension have been used widely. a Mask R-CNN was proposed by He et al. in an extension of this model. While simultaneously generating a high-quality segmentation mask for each instance, this model efficiently detects objects in an image (He et al., 2017). Using semantic and direction features based on Faster R-CNN and refining object detection Chen et al. developed an instance segmentation model. Three outputs i.e. box detection, semantic segmentation, and direction prediction (Chen et al., 2018) are produced by this model. There are various other publications reports on image segmentation methods which is described in detail in the next section. There are several datasets for training image segmentation models like PASCAL VOC, PASCAL-10K, ThinObject-5K, COCO and other datasets which is used in papers in next sections. After the training model, its performance should be examined. Object detection metrics serve as a measure to assess how well the model performs on an object detection task. Precision, Recall, Intersection-over-Union (IoU), F1 Score are object detection metrics. In most competitions, IoU is used for segmentation. IoU metric determines how many objects were detected correctly and how many false positives were generated.

3. IMAGE SEGMENTATION

In order to accurately select the object, a lot of tedious user interactions are required in each one of these methods, because of the complexity of natural scenes, overlapping objects and background color distributions, and complicated object boundaries. The significant and challenging process of image processing is called Image segmentation and it is employed in order to divide an image into meaningful parts with similar features and properties. Segmentation is mainly aimed at simplification i.e. representing an image into meaningful and easily analyzable way. Image segmentation is the necessary first step of image analysis. It is aimed at dividing an image into several parts/segments with similar features or attributes. Medical image analysis, autonomous vehicles, video surveillance, and augmented reality are among the many applications in which segmentation plays an essential role. The earliest methods, such as thresholding (Otsu, 1979), histogram based bundling, region growing (Nock & Nielsen, 2004), k-means clustering (Dhanachandra et al., 2015), watersheds

Table 2. Comparison of various segmentation techniques

Segmentation Technique	Description	Advantages	Disadvantages
Thresholding Method	It focuses on the histogram peaks of the image in order to find specific threshold values.	It is the simplest method of segmenting images and it does not require previous information.	It does not consider spatial details. It is very dependent on the appropriate threshold values.
Edge-Based Method	Edge based segmentation methods are the structural techniques based on discontinuity detection.	It performs well among objects with contrast and especially on the background and the objects of the image. Also, edge-based algorithms are usually less complex.	It does not perform well among objects with many edges. Another disadvantage is noise sensitivity.
Region-Based Method	Grow regions by recursively including the neighboring pixels that are similar and connected to the seed pixel.	Region growing techniques are generally better in noisy images where edges are difficult to detect.	It is an expensive method in terms of time and memory.
Clustering Method	The clustering process divisions the assorting into classes from a set of objects with similar characteristics.	Fuzzy clustering method implements soft clustering in which a pixel has a membership or a degree in association with each of the clusters.	Determining its membership function is not easy.
Watershed Method	It operates on the basis of topological interpretation.	The results of watershed methods are more stable, the identified boundaries are continuous.	It often needs preprocessing to work well Often needs preprocessing to work well and has complex calculations of gradients.
PDE-Based Method	It works based on calculations of differential equations.	It is the fastest way. It is the best method in terms of time.	It has complex calculations.
ANN-Based Method	ANN are used for recognition as well as segmentation of images.	One of the best advantages of neural networks is their graceful degradation in the presence of noise. Also, their ability to be used in real-time applications and the ease of implementing them.	There are storage and time to learn issues during training.

(Najman & Schmitt, 1994), and the more advanced ones such as active contours (Kass et al., 1988), graph cuts (Boykov et al., 2001), conditional and Markov random fields (Plath et al., 2009), and sparsity based (Minaee & Wang, 2019; Starck et al., 2005) methods have been developed in the literature. By specifying a brief description of every method each with its advantages and disadvantages (Gurusamy et al., 2013). Table.2 represents a comparison between various segmentation techniques.

However, recently deep learning models have provided a new generation of image segmentation models which led to astonishing enhancement of the performance. This means that these newly developed models are able to achieve the highest accuracy rates on popular benchmarks leading to a paradigm shift in the field.

In Fig. 1 image segmentation outputs of a deep learning model are presented. There can be two types of image segmentation approaches: discontinuity detection and similarity detection. They are based on properties of image.

3.1. Semantic segmentation

Every object in an image is automatically segmented according to semantic segmentation approaches (Liu et al., 2011). Semantic segmentation methods which do not focus on the object of interest solve a much large problem than needed. These methods require a lot of pre-labeled data as well as a predetermined label set which might be without the desired object label. If the label set contains the desired object, it still may not be found in the image. Many require training classifiers for every label, which for a sufficiently large label set requires excessive computation. Recently, however, deep learning models have developed a new generation of image segmentation models with excellent performance improvements leading to a paradigm shift in the field. Employing neural networks, Semantic segmentation approaches have been drastically enhanced. A sample graphical representation of object selection with semantic segmentation has been shown in Fig.2.

The rise of digital cameras, cell phone cameras, and the computing power, which is getting faster as GPUs become general purpose computing tools has led to the increasing availability of data and subsequently the considerable progress of neural networks. Deep neural networks are exponentially viable in semantic segmentation, which means labeling each region or pixel with a class of objects or non-objects. An increasingly growing field division of machine learning is deep learning, in fact it is growing with the pace making it very difficult to stay up to date, even to keep track of the works dealing with semantic segmentation. Some surveys and review papers have addressed advancements and innovations about deep learning and semantic segmentation. A study was conducted by Zhu et al. which covers a wide range of the papers and areas of semantic segmentation topics including, interactive methods, recent development in the super-pixel, object proposals, semantic image parsing, image cosegmentation, semi & weakly supervised, and fully supervised

Fig. 1. Segmentation results on sample images.

image segmentation (Zhu et al., 2016). A taxonomy of segmentation algorithms and overview of completely automatic, passive, semantic segmentation algorithms are presented by Thoma (2016). As for scene understanding in the context of the autonomous driving, a review of neural network based semantic segmentation was presented by Niemeijer et al. (2017). A survey of semantic segmentation approaches, i.e., region based, FCN based and weakly supervised approaches was presented by Guo et al. (Guo et al., 2018). The strengths, weaknesses and major challenges in image semantic segmentation were summarized in their studies.

Another survey studied recent progress in semantic segmentation with CNN's (Geng et al., 2018). In recent years, convolutional neural networks are leading the way in most computer vision algorithms. Geng et al. used Pascal VOC 2012 image datasets to semantic segmentation challenge to combine some popular underlying referential methods to create effective methods as components to enhance CNNs for the segmentation of specific semantic objects. The first successful application of convolutional neural network was developed by LeCun et al (LeCun et al., 1998). An architecture named LeNet5 was introduced by them to read zip codes, digits, and extract features at multiple locations in the image. Then, a large deep convolutional neural network (AlexNet) considered as one of the most influential publications in the field (Krizhevsky et al., 2012) was released by Alex Krizhevsky. AlexNet which is a deeper and wider version of the LeNet is used in order to learn complex objects and object hierarchies. ZFNet which is a fine-tuning of the AlexNet structure (Zeiler & Fergus, 2014) was presented by Zeiler and Fergus. A technique for visualizing feature maps at any layer in the network model was proposed by them. A multi-layered deconvolutional network is used by this technique in order to project the feature activations back to the input pixel space. a Network-In-Network model was proposed by Lin et al. based on micro neural networks, a multilayer perceptron encompassing several completely connected layers with nonlinear activation functions (Lin et al., 2013). A thorough study of the literature at the time of this writing covering a broad spectrum of pioneering works for semantic and instance-level segmentation, including fully convolutional pixel-labeling networks, encoder-decoder architectures, multi-scale and pyramid based approaches, recurrent networks, visual attention models, and generative models in adversarial settings was provided by Minaee et al. They investigated the similarity, strengths and challenges of these deep learning models, examined the most widely used datasets, reported performances (Minaee et al., 2021).

A hybrid of semantic segmentation and object detection is Instance segmentation. Instance segmentation is a computer vision challenging that requires the prediction of object instances and their per-pixel segmentation mask. Thus, it can be said instance segmentation is a task of finding simultaneous solution to object detection as well as semantic segmentation. Hafiz and Bhat studied on the background of instance segmentation and its evolution, techniques and popular datasets. The revolution of notable techniques in instance segmentation began with RCNN, Fast R-CNN, Multipath Network, Faster R-CNN. After these techniques, Mask R-CNN and TensorMask invented. TensorMask uses sliding window technique which

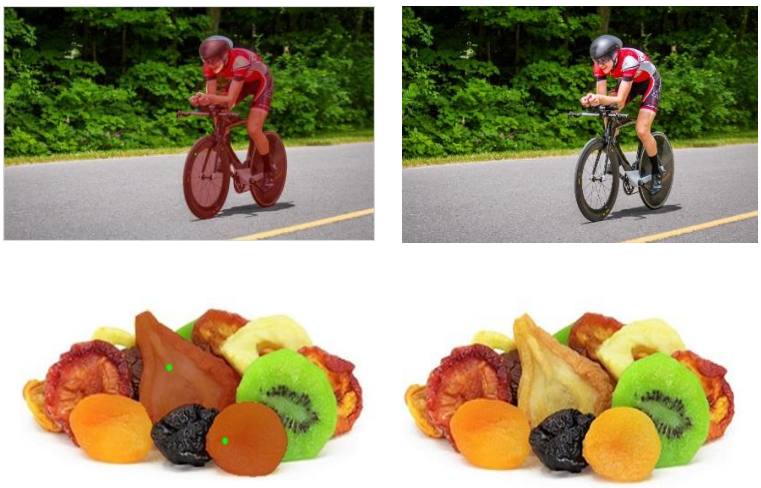

Fig. 2. Selecting Objects in Semantic Segmentation

generate predictions of bounding-box with the help of a dense, regular grid have advanced rapidly and proven popular (Hafiz & Bhat, 2020). Chen et al investigate the paradigm of dense sliding-window instance segmentation. They treated dense instance segmentation as a prediction task over 4D structured tensors and They provided a general framework called TensorMask that explicitly captured novel operators on 4D structured tensors (Chen et al., 2019).

3.2. Interactive Object Segmentation

Interactive image segmentation task is aimed at extracting a high-quality segmentation mask delineating the object of interest using only a few user clicks. Deep learning based methods have proved successful in this task (Jang and Kim, 2019; Lin et al., 2020; Sofiiuk et al., 2020). In case there are any mistakes in the predictions, this task enables refining the prediction with further interaction inputs. These inputs usually come in the form of user clicks or strokes (Le et al., 2018; Liew et al., 2017; Maninis et al., 2018; Xu et al., 2016) or bounding boxes (Jain and Grauman, 2013; Lempitsky et al., 2009; Xu et al., 2017). This kind of input places restrictions on the location of the related object. Recent advances in deep learning allows such methods to select familiar objects with a small amount of input. Instead, systems which use language based input to drive the selection have been proposed (Liu et al., 2017; Rupprecht et al., 2018; Yu et al., 2018b). A neural network employs natural language phrases to infer high-level attribute information about what the related object looks like that can then be used to select the objects.

Each one of these interaction approaches may still fall short and require additional and excessive user interact while great improvements have been made in interactive selection. For instance, in order to infer the target object given only spatial constraints click based methods are necessary and thus they are usually trained to select entire objects. However, the region of user interest may instead be an object part or a combination of multiple objects. Click based methods also generally assume accurate input, but with mobile devices it can be difficult for users to accurately click on objects, especially given that the users finger is occluding the object of interest. Accurately segmenting a target of interest with a few clicks is a significant challenge.

On the other hand, language-driven segmentation methods learn the overall appearance of objects and must infer their location. Ambiguities such as whether the target is an object, object part, or collection of objects are overcome by a language phrase naturally and easily. A phrase can also provide rough spatial information. Moreover, speech is a natural and preferred interface which is easier than precise touching on a small phone screen for mobile devices like smartphones. However, in many cases an object name and rough location is not enough to produce a desired result and thus it is much easier for a user to directly click on the object. To verbally articulate some required corrections which is not related to an entire semantically meaningful region be difficult. Furthermore, it is troublesome to have labels and training data for all possible objects because of the long tail distribution of objects in images. The strengths and weaknesses of click based and phrase-based inputs are complementary with clicks giving hard spatial constraints and phrases giving high-level attribute information. The amount of user interactions required to accurately select related objects might be reduced by an effective combination of these inputs.

There are also some language-based segmentation methods like Hu et al. They proposed an end-to-end recurrent convolutional network model to encode the expression into a vector representation, extract a spatial feature map representation from the image. They used ReferIt datasets for train the model and evaluated the performance of their model on test dataset with IoU metrics (Hu et al., 2016) Another study by (Liu et al., 2017). They proposed convolutional multimodal LSTM to learn word-to-image interaction. They Trained model with ImageNet, and DeepLab-101 datasets and evaluated their model with four datasets Google-Ref, UNC, UNC+, and ReferItGame on IoU and Precision@X where Precision@X means the percentage of images with IOU higher than X. But these researchs only provide an initial result and cannot further refine the result to correct mistakes. Early methods depend on low-level features, such as color similarity or boundary properties (Kass et al., 1988; Mortensen & Barrett, 1995). For example (Boykov & Jolly, 2001; Li et al., 2004; Rother et al., 2004) take graphical models. Boykov and Jolly proposed a new technique for N-dimensional images in general purpose interactive segmentation. In order to provide hard constraints for segmentation, the user marks certain pixels as “object” or “background” and They used Graph cuts to find the globally optimal segmentation of the N-dimensional image. Rother et al extend the graph-cut approach in three respects and showed that the

proposed method outperforms competitive tools for moderately difficult examples. Grady (2006) used random walker and are based on geodesic approaches (Bai and Sapiro, 2007; Criminisi et al., 2008; Price et al., 2010). However, low-level features are not robust and thus excessive user interactions are necessary.

4. CONCLUSION

The performance of an intelligent image cropper highly depends on the ability to detect objects and speed in cutting operation. There are many methods to detect objects in an image. One of them is Image segmentation. Image segmentation divides a digital image into multiple segments and in this process one label assigns to every pixel in an image. The goal of segmentation is to simplify change the representation of an image into something that is more meaningful and easier to analyze. Algorithms which are based on neural network and deep learning methods are more accurate in detecting and eliminating noise in objects in the image. In this paper, a history of development in image segmentation was investigated based on discontinuity and similarity detection-based approaches. We investigated advantage and disadvantage of image segmentation and studied new models in semantic segmentation and interactive object segmentation.

REFERENCES

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on pattern analysis and machine intelligence, 39*(12), 2481-2495.

Bai, X., & Sapiro, G. (2007). *A geodesic framework for fast interactive image and video segmentation and matting.* Paper presented at the 2007 IEEE 11th International Conference on Computer Vision.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence, 23*(11), 1222-1239.

Boykov, Y. Y., & Jolly, M.-P. (2001). *Interactive graph cuts for optimal boundary & region segmentation of objects in ND images.* Paper presented at the Proceedings eighth IEEE international conference on computer vision. ICCV 2001.

Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., & Adam, H. (2018). *Masklab: Instance segmentation by refining object detection with semantic and direction features.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on pattern analysis and machine intelligence, 40*(4), 834-848.

Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., & Chen, L.-C. (2019). Panoptic-deeplab. *arXiv preprint arXiv:1910.04751*.

Cheng, B., Ni, B., Yan, S., & Tian, Q. (2010). *Learning to photograph.* Paper presented at the Proceedings of the 18th ACM international conference on Multimedia.

Criminisi, A., Sharp, T., & Blake, A. (2008). *Geos: Geodesic image segmentation.* Paper presented at the European Conference on Computer Vision.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). *Studying aesthetics in photographic images using a computational approach.* Paper presented at the European conference on computer vision.

Deng, Y., Loy, C. C., & Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine, 34*(4), 80-106.

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science, 54*, 764-771.

Dhar, S., Ordonez, V., & Berg, T. L. (2011). *High level describable attributes for predicting aesthetics and interestingness.* Paper presented at the CVPR 2011.

Geng, Q., Zhou, Z., & Cao, X. (2018). Survey of recent progress in semantic image segmentation with CNNs. *Science China Information Sciences, 61*(5), 051101.

Ghiasi, G., & Fowlkes, C. C. (2016). *Laplacian pyramid reconstruction and refinement for semantic segmentation.* Paper presented at the European conference on computer vision.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on pattern analysis and machine intelligence, 28*(11), 1768-1783.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval, 7*(2), 87-93.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask r-cnn.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Hu, R., Rohrbach, M., & Darrell, T. (2016). *Segmentation from natural language expressions.* Paper presented at the European Conference on Computer Vision.

INCORP, A. (2002). Adobe Photoshop User Guide.

Jain, S. D., & Grauman, K. (2013). *Predicting sufficient annotation strength for interactive foreground segmentation.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Jang, W.-D., & Kim, C.-S. (2019). *Interactive image segmentation via backpropagating refinement scheme.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, *1*(4), 321-331.

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). *Panoptic segmentation.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

Le, H., Mai, L., Price, B., Cohen, S., Jin, H., & Liu, F. (2018). *Interactive boundary prediction for object selection.* Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Lempitsky, V., Kohli, P., Rother, C., & Sharp, T. (2009). *Image segmentation with a bounding box prior.* Paper presented at the 2009 IEEE 12th international conference on computer vision.

Li, Y., Chen, X., Zhu, Z., Xie, L., Huang, G., Du, D., & Wang, X. (2019). *Attention-guided unified network for panoptic segmentation.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Li, Y., Sun, J., Tang, C.-K., & Shum, H.-Y. (2004). Lazy snapping. *ACM Transactions on Graphics (ToG)*, *23*(3), 303-308.

Liang, X., Shen, X., Feng, J., Lin, L., & Yan, S. (2016). *Semantic object parsing with graph lstm.* Paper presented at the European Conference on Computer Vision.

Liew, J., Wei, Y., Xiong, W., Ong, S.-H., & Feng, J. (2017). *Regional interactive image segmentation networks.* Paper presented at the 2017 IEEE international conference on computer vision (ICCV).

Lin, M., Chen, Q., & Yan, S. (2013). Network in network, CoRR abs/1312.4400. *URL http://arxiv. org/abs/1312.4400*.

Lin, Z., Zhang, Z., Chen, L.-Z., Cheng, M.-M., & Lu, S.-P. (2020). *Interactive image segmentation with first click attention.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., & Yuille, A. (2017). *Recurrent multimodal interaction for referring image segmentation.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Liu, C., Yuen, J., & Torralba, A. (2011). Nonparametric scene parsing via label transfer. *IEEE Transactions on pattern analysis and machine intelligence*, *33*(12), 2368-2382.

Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully convolutional networks for semantic segmentation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Luo, W., Wang, X., & Tang, X. (2011). *Content-based photo quality assessment.* Paper presented at the 2011 International Conference on Computer Vision.

Maninis, K.-K., Caelles, S., Pont-Tuset, J., & Van Gool, L. (2018). *Deep extreme cut: From extreme points to object segmentation.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on pattern analysis and machine intelligence*.

Minaee, S., & Wang, Y. (2019). An ADMM approach to masked signal decomposition using subspace representation. *IEEE transactions on image processing*, *28*(7), 3192-3204.

Mortensen, E. N., & Barrett, W. A. (1995). *Intelligent scissors for image composition.* Paper presented at the Proceedings of the 22nd annual conference on Computer graphics and interactive techniques.

Najman, L., & Schmitt, M. (1994). Watershed of a continuous function. *Signal Processing*, *38*(1), 99-112.

Niemeijer, J., Pekezou Fouopi, P., Knake-Langhorst, S., & Barth, E. (2017). *A Review of neural network based semantic segmentation for scene understanding in context of the self driving car.* Paper presented at the Student Conference on Medical Engineering Science.

Nock, R., & Nielsen, F. (2004). Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, *26*(11), 1452-1458.

Noh, H., Hong, S., & Han, B. (2015). *Learning deconvolution network for semantic segmentation.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62-66.

Plath, N., Toussaint, M., & Nakajima, S. (2009). *Multi-class image segmentation using conditional random fields and global classification.* Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning.

Porzi, L., Bulo, S. R., Colovic, A., & Kontschieder, P. (2019). *Seamless scene segmentation.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Price, B. L., Morse, B., & Cohen, S. (2010). *Geodesic graph cut for interactive image segmentation.* Paper presented at the 2010 IEEE computer society conference on computer vision and pattern recognition.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*, 91-99.

Rother, C., Kolmogorov, V., & Blake, A. (2004). " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (ToG)*, *23*(3), 309-314.

Rupprecht, C., Laina, I., Navab, N., Hager, G. D., & Tombari, F. (2018). *Guide me: Interacting with deep networks.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., & Cohen, M. (2006). *Gaze-based interaction for semi-automatic photo cropping.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in computing systems.

Sofiiuk, K., Petrov, I., Barinova, O., & Konushin, A. (2020). *f-brs: Rethinking backpropagating refinement for interactive segmentation.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Starck, J.-L., Elad, M., & Donoho, D. L. (2005). Image decomposition via the combination of sparse representations and a variational approach. *IEEE transactions on image processing*, *14*(10), 1570-1582.

Stentiford, F. (2007). *Attention based auto image cropping.* Paper presented at the International Conference on Computer Vision Systems: Proceedings (2007).

Su, H.-H., Chen, T.-W., Kao, C.-C., Hsu, W. H., & Chien, S.-Y. (2012). Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, *14*(3), 833-843.

Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. (2003). *Automatic thumbnail cropping and its effectiveness.* Paper presented at the Proceedings of the 16th annual ACM symposium on User interface software and technology.

Thoma, M. (2016). A survey of semantic segmentation. CoRR abs/1602.06541 (2016). In.

Wang, W., & Shen, J. (2017). *Deep cropping via attention box prediction and aesthetics assessment.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., & Urtasun, R. (2019). *Upsnet: A unified panoptic segmentation network.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Xu, N., Price, B., Cohen, S., Yang, J., & Huang, T. (2017). Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243*.

Xu, N., Price, B., Cohen, S., Yang, J., & Huang, T. S. (2016). *Deep interactive object selection.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Yan, J., Lin, S., Bing Kang, S., & Tang, X. (2013). *Learning the change for automatic image cropping.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., & Berg, T. L. (2018). *Mattnet: Modular attention network for referring expression comprehension.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks.* Paper presented at the European conference on computer vision.

Zhang, M., Zhang, L., Sun, Y., Feng, L., & Ma, W. (2005). *Auto cropping for digital photographs.* Paper presented at the 2005 IEEE International Conference on Multimedia and Expo.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). *Pyramid scene parsing network.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Zhu, H., Meng, F., Cai, J., & Lu, S. (2016). Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, *34*, 12-27.