Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230455

Review of automatic image cropping in practical application

Li Qiuyu

Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province, P.R.C., 116024, China.

1310650689@qq.com

Abstract. There is now a greater demand for photo Aesthetics and a desire for higher quality images. As a result there is also a greater demand for image cropping. Although there is a lot of work currently addressing this need, there has not been a comprehensive paper on the current state of development in this area until now. So this paper summarise the classical approach and some general datasets and some evaluation metrics. More importantly, this paper analyse the current challenges in the field and hope that future research will address these issues and lead to a better development of the field.

Keywords: Automatic image cropping, Aesthetics, Saliency.

1. Introduction

The word photography comes from the Greek words for light and painting, and as the name implies it is the use of light to make pictures. But it is more than just the process of recording images with professional equipment such as cameras and camcorders, and in many ways we should understand it on an artistic level. Nowadays, society has entered the digital age and with the computerisation of image processing, satellite methods of communication and the trend towards more colour in formats, photography is becoming more and more diverse and is being used in an increasing number of fields. These areas include fashion, industry, crafts, online life and many more. Photography is important in our lives. At the same time professional photography has high demands, as many factors need to be taken into account for a beautiful photograph. One is colour conversion and the other is image cropping.

It is important to optimise the colour of the cropped image and the colour conversion is used to make the final image more aesthetically pleasing. When we take photographs or acquire images, we often need to post-process their colours to meet the requirements of the scenario in which they will be used. This post-processing often requires the use of software such as PS and LR to manually mix the colours to produce the result. As you can imagine, this process is time consuming and requires some professionalism from the user. Therefore, some researchers have proposed to borrow some of the best ready-made examples and migrate their colour schemes directly to the original images, so that the colours of the original images can be edited to be consistent with the best examples in terms of colour style. This is the initial goal of the image colour migration study.

While the colour conversion is relatively simple, the more important and relatively more difficult part is the image cropping. This thesis focuses on image cropping.In photography, cropping, also known as secondary composition, is the process of taking an existing composition and adjusting it again as needed. The aim is to remove the objects that damage the overall structure of the picture and

DOI: 10.54254/2755-2721/4/20230455

adjust the subject to the ideal position, so that the subject stands out, the picture is simple and vivid, the proportions are reasonable and it has more aesthetic value. Simply put, cropping is the process of removing parts of the image to create a prominent or enhanced composition. Proper cropping can turn an otherwise bland image into an excellent piece of work. For arbitrarily captured images, the purpose of image cropping is to remove some irrelevant areas and find an aesthetically pleasing area. Perspective composition is one of the most important factors affecting the aesthetics of an image in image cropping. Nowadays, Adobe Photoshop, or "PS" for short, is known and often used in everyday life, and has powerful features that allow for complex and varied processing of images, such as adding brilliant effects to images, blending images and so on. Cropping. These images can be found all over the internet and in some apps, such as promotional images for websites, games, apps for new features, etc. They are very common in our lives.

Image cropping works by defining a rectangular area inside the input image as the final output and excluding content outside the selected area, which can be difficult to find the best area even for professionals. So professional image cropping can be very challenging. A great deal of money is spent on image processing on a photographic basis, and a great deal of time is spent on training the expertise of someone who does not have the basic knowledge to do this. Even with professional training, it can be difficult to achieve a professional aesthetic standard. It takes a lot of money, a lot of time to learn the techniques, and a lot of labour to produce images that meet aesthetic standards. However, if computers were used to process images, it would save more time, labour and would be cheaper. So if computers could be taught how to professionally crop images, it would be interesting to see how ordinary people could create beautiful photographs in their everyday lives.

Early image cropping techniques were largely dependent on attentional mechanisms. To find the primary item or the most informative, they generally used saliency detection [1-2]. However, cropping that is exclusively dependent on the cropping of the main object may not always produce results that are attractive to the eye. Methods of cropping that are aesthetic in nature [3-4]. Make use of an image's aesthetic elements or compositional guidelines to enhance the overall aesthetic elements or guidelines to improve the general image quality. These methods use hand-crafted features to assess the quality of candidate crops or rank them using a ranking model.

Image cropping is divided into aesthetic-based image cropping and saliency-based image cropping.

Aesthetics-based image cropping is a way to improve cropping by enhancing its aesthetic quality. Early approaches achieved this by using hand-crafted features and compositing rules. Numerous strategies have been put out to deal with image cropping in a data-driven manner by embracing the ongoing development of deep learning and its application in this field. A target detection system with two networks [5], one for producing candidate crops and the other for aesthetic assessment, was employed by Wei et al. [6]. And the aesthetic preferences of people were recorded by constructing an aesthetic dataset.

The saliency-based method is oriented on maintaining the most crucial material in the finest crops [7-8]. Wang and Shentrained a network to create a network that encompasses the most salient parts of candidate crops using a saliency dataset [9]. The concept of establishing an initial visual saliency rectangle that contains the most crucial objects was also put forward by the Lu et al. [10].

Although a very large number of different methods have been proposed for image cropping, there has not been a comprehensive article on these methods. This paper will provide an introduction to them. In the second part, the classical methods are presented; in the third part, resources are presented, including online resources, datasets and so on; in the fourth part, the challenges of image cropping are presented, including in the commercial, research and human cognitive domains.

2. Classical methods

2.1. Aesthetic-based image cropping

The characteristics of image cropping make training an efficient cropping model a challenging task. The amount of data annotated in the current database is not enough to train a reliable cropping model because, on the one hand, image cropping requires excellent photographic knowledge and experience

DOI: 10.54254/2755-2721/4/20230455

and is very expensive. On the other hand, the search space for image cropping is very large, with millions of candidate images for each image.

To solve these problems, it is now popular to operate the image cropping module under a convolutional neural network (CNN) architecture. the CNN architecture for image cropping model learning is divided into two main parts, the feature extraction part and the image cropping part.

Zeng Hui, Li Lida, Cao Zisheng, and Zhang Lei's grid-anchored image-based planting strategy is the currently suggested extremely traditional approach. It decreases the number of eligible crops from millions of different species to no more than 100 [11]. A new picture cropping database was created with thorough annotation of each reference image using this information as well. This database contains 106,860 annotated candidate crops, while defining more reliable metrics to evaluate the performance of the learned cultivation model. In addition, they designed an efficient image cropping module under a convolutional neural network (CNN) architecture, with the learned cropping model running at 125 FPS.

This work offers an efficient and portable learning framework for cropping models while being aware of the unique properties of image cropping. The model's overall design, which can be shown in Fig. 1, is made up of two modules: one for feature extraction and the other for image cropping.

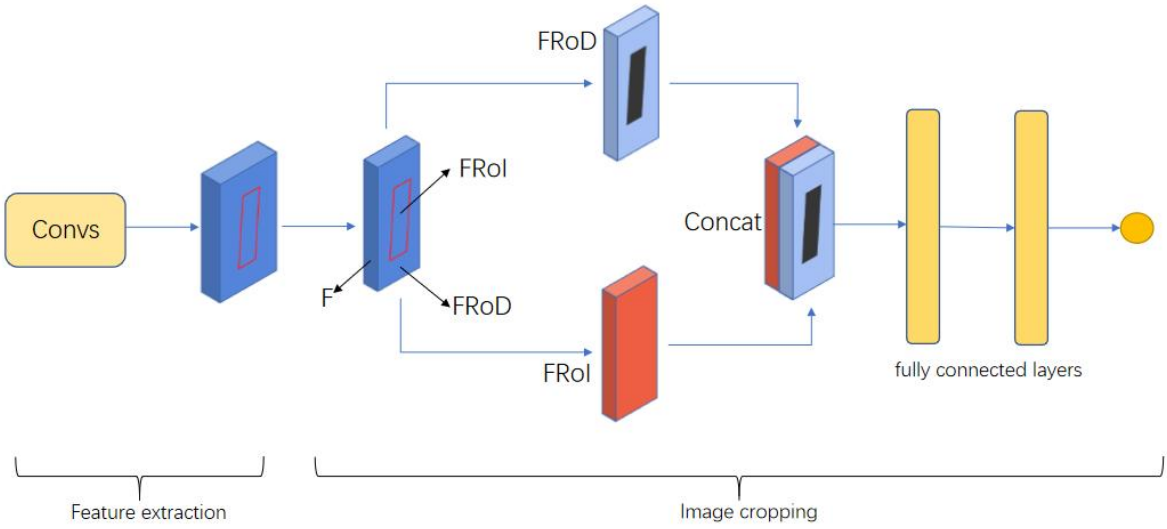

Figure 1. The proposed CNN architecture for image cropping model learning.(F: The entire feature mapping output from the feature extraction module; RoI: the region of interest; RoD：Information to be discarded at the time of cropping).

In this work, a feature extraction module is first pre-trained on the CNN model. In order for the cropping module to analyze the image composition, the feature extraction program must retain a high enough spatial resolution.

The whole feature map result from the feature extraction module has been represented by the letter F in the image cropping module, where it is then sent. Additionally, in picture cropping, the data that needs to be removed is referred to as the discard region (RoD) and the region of interest (RoI) is known as the region of interest. The terms FRoI and FRoD, respectively, stand for the feature maps in RoI and RoD. In this paper, we first use RoIAlign to convert FRoI into an FA component with a fixed spatial resolution. FRoD removes FRoI from F, i.e. sets the FRoI value in F to zero, and then performs RoDAlign on FRoD, i.e. uses the same bilinear interpolation as RoIAlign to obtain FARoD with the same spatial resolution as FARoI. The FARoI and FARoD are then subjected to a concat operation to join them into an aligned feature map along the channel dimension, which will contain the RoI and RoD information. The two completely connected layers then receive the combined feature map, and the process is completed with a MOS prediction.

DOI: 10.54254/2755-2721/4/20230455

Since experimenting with the results, it was discovered that A2-RL and VFN only accomplished performance that was comparable to the benchmark L. This was due to the fact that A2-RL had a common aesthetic classifier supervise it during training, whereas VFN's ranking pair lacked a high degree of reliability. VEN was able to outperform VFN thanks to the use of a more dependable ranking pair. As a result of being overseen by VEN's predictions, VPN fared somewhat worse than VEN. Contrarily, the method suggested in this study outperforms VEN to a large extend because the annotation approach uses richer clipping information than the pairwise ranking comments used by VEN, enabling the model to be trained with clipping modules more successfully.

2.2. Saliency-based image cropping

One of the fundamental methods for altering images is cropping, which aims to improve the composition overall, remove distracting elements from the image, and improve the visual/aesthetic perception. However, obtaining a suitable composition for a cropped image and obtaining better visual quality is very difficult, identifying the main object as well as the subject of a given image requires extensive domain knowledge and sophisticated skills, while the aesthetic evaluation of a cropped image is highly subjective and the solution space is very large from the large number of cropping candidate areas present in the image.

Regression networks are currently widely used since experimental evidence has demonstrated that they can help to achieve large aesthetic quality cropped images using numerous individual stages.

As a result, Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin et al. propose a deep learning-based structure to learn object composition from images with high aesthetic quality, in which the anchor region is discovered by a Gaussian kernel convolutional neural network (CNN), and that this initially detected region is then fed into a light-weighted regression network [12].

The suggested method, as shown in Figure 2, differs from conventional image growing techniques in that it uses a regression network to directly map individual regions to the final output region rather than explicitly or implicitly generating and evaluating various candidate growing regions. Therefore, in this framework, there is no extensive evaluation of multiple candidates and the data will only be streamed through the network once, which greatly improves efficiency while maintaining accuracy.

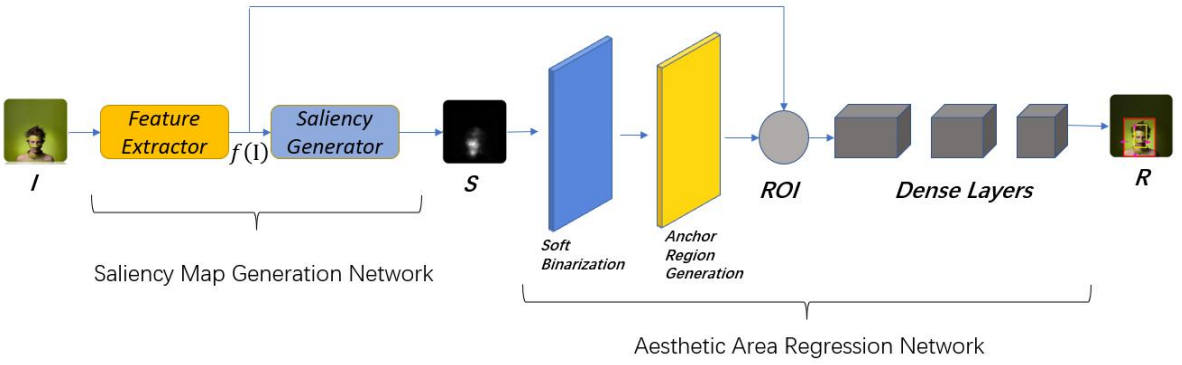

Figure 2. Architecture of the proposed saliency map detection and aesthetic area regression network.

Saliency Map Generation Network and Aesthetic Area Regression Network make up the proposed structure in this paper.

The salicency map is produced by a modified u-type network in the saliency map generation network. the detailed structure of the u-type network is shown in Figure 3. During the upsampling process, the feature mapping from the convolutional layer to the deconvolutional layer is gradually merged. The u-encoder network's is made up of four fundamental building blocks, each built by stacking two convolutional layers and a maximum pooling layer. Similar to this, a decoder is built from four basic blocks, where two deconvolutional layers and one upsampling layer are created for each basic block. The encoder is used to extract the features from the input image and create the

DOI: 10.54254/2755-2721/4/20230455

saliency mapping based on the decoder. The feature mapping from the input image is duplicated and attached directly to the equivalent block in the decoder with the same feature size.

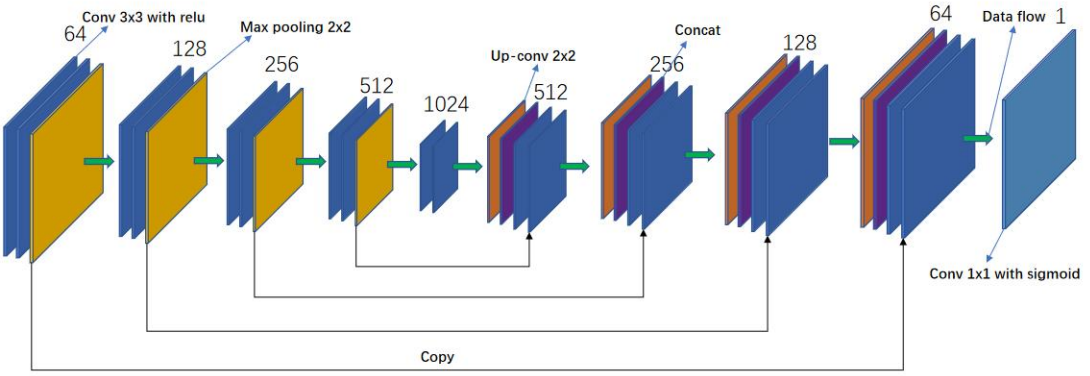

Figure 3. The U-Shaped network implemented in this work for saliency map detection

The suggested Aesthetic Area Regression Network is utilized to determine the association between objects of interest and areas of high aesthetic quality based on the saliency map that has been observed. Aesthetic Area Regression Network consists of Soft Binarization Layer, Anchor Region Generation Layer, Dense Layers, and Aesthetic Area Representation.To address the planting The Soft Binarization Layer is used to improve the quality of the object of interest in the saliency map in order to address the sensitivity of the system to outliers in the saliency map. Dense Layers are used to map the final cropping window of the anchor region with three fully connected layers for better visual quality, and Aesthetic Area Representation is utilized to explain the connection between high-quality aesthetic regions and discovered anchor regions.

Figure 4 displays the outcomes of the multiple planting from the assessment set, with the orange boxes denoting the anticipated optimal planting window by the proposed approach. These cropped photographs demonstrate that the cropped images attain superior aspect ratio and composition than the original images. With the suggested procedure, successful image cropping was accomplished.

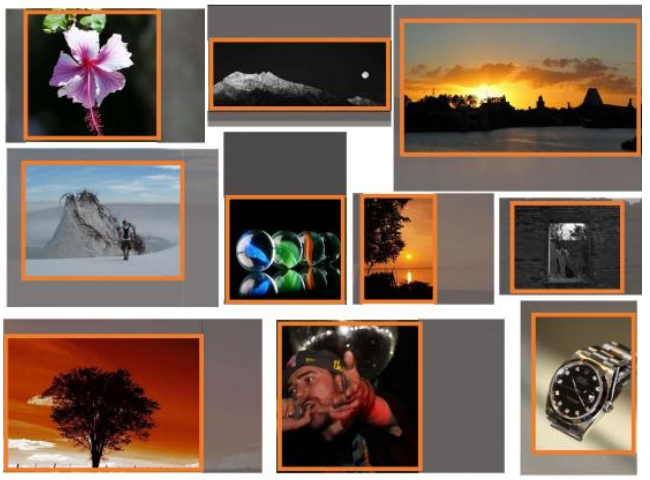

Figure 4. Cropping rectangle produced by the proposed system.

In terms of both IoU and BDE metrics, the algorithm presented in this research performs better than previous cutting-edge cropping techniques. As the method finds the final planting area by means of a neural network based on a hidden relationship between the object of interest and a region of high

DOI: 10.54254/2755-2721/4/20230455

aesthetic quality, this avoids the need for iterative evaluation of multiple candidate planting areas, and also achieves a higher processing efficiency compared to other methods.

3. Dataset

In this section we present some widely used datasets used for training and evaluation.

3.1. Training Data

3.1.1. AVA Dataset. AVA Dataset is a database for aesthetic quality assessment, including 250,000 photographs [13]. For each photograph, there is a series of ratings, as well as semantic level labels, of which there are 60 categories, and photographic style, that is, the style of the photograph, which has 14 categories.

The AVA Dataset has 3 categories of annotations: Aesthetic annotations, Semantic annotations and Photographic style annotations.

In Aesthetic annotations, each image has a number of people voting on it, with the number of votes ranging from 78 to 549. roughly 210 votes are cast for each image. The range of votes is 0 to 9, with a higher score indicating a higher-quality photograph. The visual quality increases as the score rises. And the markers include not only professional photographers and photographers, but also photography enthusiasts, so that it seems more universal.

Semantic annotations have 66 textual tags. 150,000 photographs have two tags, while 200,000 images have just one.

Photographic style annotations relate to photographic aesthetics. The descriptions start from 3 main directions: light, colour and composition, and end up with 14 attributes.

AVA is not the first aesthetic quality database, nor will it be the last, but it is still the largest aesthetic data set. The proportion of scores between 2 and 8 is over 99.77%, so the proportion of 0s and 9s is so low that there is no need to worry about scores being too outrageous. For scores close to 5, which show a clear Gaussian distribution, the combined performance suggests that all votes are largely reaching a unanimous conclusion. The AVA dataset is reliable.

An example of the AVA dataset is shown in Figure 5, which illustrates the bounding boxes and action annotations in the example frames. Each bounding box corresponds to one gestural action (orange), 0–3 object interactions (red), and 0–3 social interactions (blue).

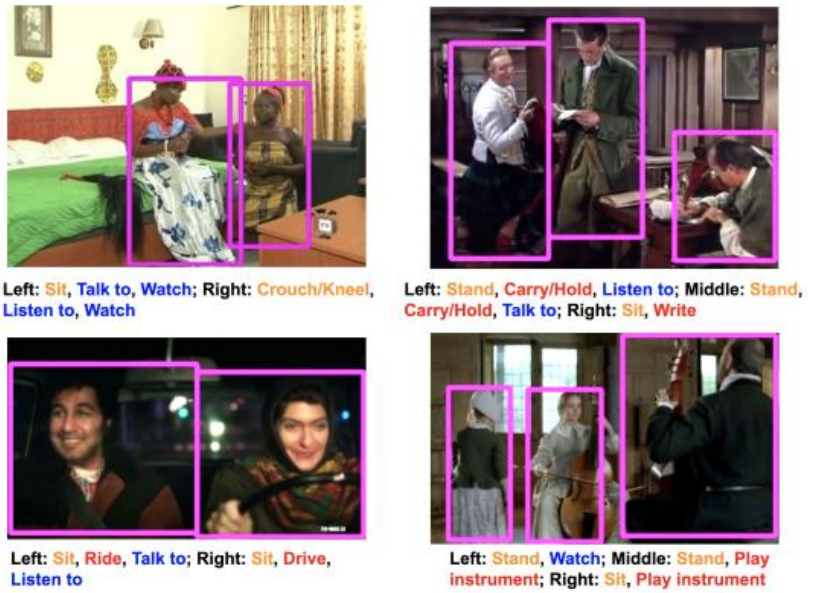

Figure 5. Examples of AVA datasets.

DOI: 10.54254/2755-2721/4/20230455

3.1.2. SALICON Dataset. The SALICON dataset was built in 2015 by Jiang et al. at the National University of Singapore [14]. The dataset, which is currently the largest in the field of image human eye focus detection, consists of 20,000 photos that were chosen from the Microsoft COCO dataset. The dataset makes use of Amazon Mechanical Turk, a crowdsourcing tool that enables annotators to click with a mouse on the region of their focus, in place of an eye-tracking device to record eye movement data. Nevertheless, Tavakoli et al. noted that there was still a sizable discrepancy between the actual eye movement data recorded by the oculograph and the eye movement data recorded by the mouse, so using the eye movement data recorded in various ways as training samples to train the model, the end result was that different training samples would have a different impact on the model's final performance. The results and the relative performance of the models using mouse-recorded eye-movement data are not consistent with the results of the tests on real eye-movement data. Nevertheless, given the large size of the SALICON dataset, it is widely used by the current mainstream deep learning-based saliency detection models. The SALICON dataset exposes the eye-movement data from the training (10,000) and validation (5,000) sets, but retains the eye-movement data from the test set (5,000).

An example of a prediction using the SALICON dataset is shown in Figure 6, while the prediction results for the focus position are shown on the right.

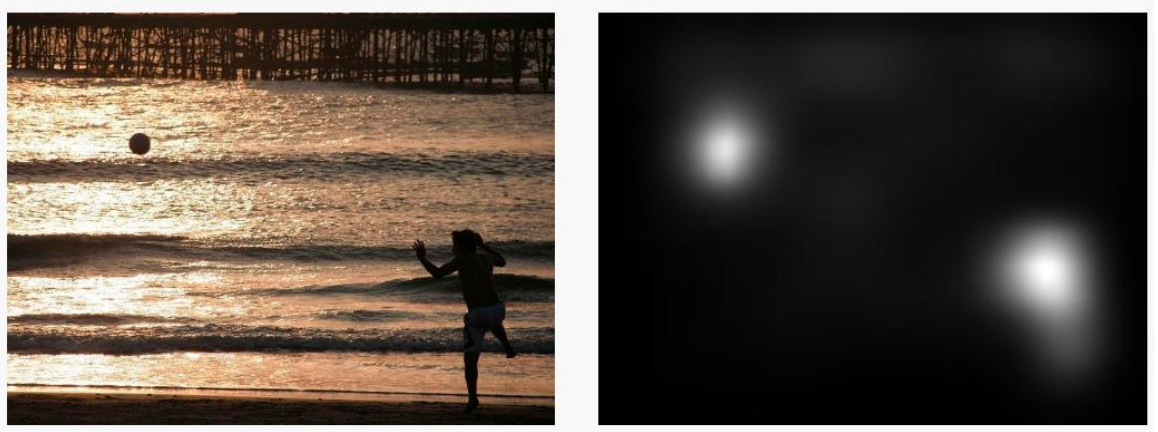

Figure 6. Examples of SALICON datasets.

3.2. Evaluation Data and Metrics

3.2.1. CUHK Cropping Dataset. This dataset was released by the laboratory of Xiaogu Tang at the Chinese University of Hong Kong [15], and was manually cropped by experienced photographers, with a total of 950 images. The 950 images comprise a total of 7 categories of images, of which 134 are animal, 136 are architecture, 133 are human, 140 are landscape, 136 are night, 138 are plant and 133 are still life. Three photographers will assign labels to each shot, with each set of crop factors consisting of x,y positions in the top left and bottom right corners.

Figure 7 shows an example of the original and cropped image.

DOI: 10.54254/2755-2721/4/20230455

Figure 7. Examples of CUHK Cropping datasets.

3.2.2. Flickr cropping dataset. This dataset was first collected by researchers from Flickr with 31,888 images [16], then workers were hired on the Amazon annotation platform AMT to filter out the inappropriate images. The remaining images were cropped by a group of photographers to produce 10 cropped versions of each image, which were then sent to the AMT platform for the annotators to choose the good and bad.

For absolute annotations, each flicker_photo_id corresponds to a set of annotation values crop. In addition to absolute annotations, this dataset also contains relative annotation results, that is, a single image is annotated with two crops, allowing subjects to choose which one they prefer, and therefore includes two annotations, crop0, crop1, vote_for_0 and vote_for_1, which are the results of subjects voting for the 0th crop box and the 1st crop box respectively, with 1 indicating that there is a vote.
Figure 8 depicts a crop pair generating example. The source image is on the left, and the four equivalent crop pairings are on the right. To avoid having too much irrelevant content, saliency maps were used to produce each crop pair window at random. Secondly, the aesthetic preference relationships between the crop pairs were determined by the results of the AMT worker rankings.

Figure 8. Examples of Flickr cropping datasets.

Variations of pixel accuracy and IoU are frequently used as evaluation metric criterion in picture segmentation. It indicates the amount of pixels that would have belonged to class I but are anticipated to be in class j. There are a total of k+1 classes (from to, which contains an empty class or backdrop). That is, the genuine number is denoted by whereas the false positive and false negative values are, respectively, and signify the true number.

DOI: 10.54254/2755-2721/4/20230455

Pixel accuracy is defined as the number of pixels in the correct prediction category as a proportion of the total number of pixels and can be calculated using pixel_accuracy = number of correct pixels predicted / total number of pixels predicted. This is calculated as follows:

$$PA = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{i=0}^{k} P_{ij}} \#(1)$$

The percentage of pixels properly classified in each category, or mean pixel accuracy (MPA), is obtained by taking the mean value of pixel accuracy for each category and is calculated as follows:

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}} \#(2)$$

Intersection over Union is the intersection ratio between the detection frame and the true frame for target detection, and the intersection ratio between the prediction mask and the true mask is calculated for image segmentation.

The mean intersection-to-merge ratio (mean IOU), or mIOU for short, is the intersection of the predicted and actual regions divided by the merge of the predicted and actual regions, and is expressed as the proportion of the merge of the true values for each category and the anticipated outcomes of the model, with the sum being computed as the average, calculated as follows:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \#(3)$$

The Frequency Weighted Intersection over Union (FWIoU) is a boost to MIoU, a method that sets weights for each class based on its frequency of occurrence and is calculated as follows:

$$FWIoU = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} P_{ij} P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \#(4)$$

4. Challenge

There are many challenges in the implementation of aesthetically based image cropping.

Because it is based on aesthetic image cropping, the results obtained are subjective. Different people have different cultures, different backgrounds and different preferences, so each person has a different assessment of aesthetics. Because of these differences, it is very difficult to find a way of cropping so that the image obtained after cropping can be approved by different people. For example, as shown in Figure 9, where a tree is shown, different people have different views on how to cropping an image, as shown by the different coloured cropping boxes in the figure. Some people will prefer to place it in a symmetrical position, but others will prefer to place it at a golden mean. Therefore, it is very difficult to evaluate and explore more accurately the aesthetic modelling that can be used as a baseline.

DOI: 10.54254/2755-2721/4/20230455

Figure 9. Picture of a tree.

It is very difficult to get an aesthetic evaluation from the public, and it is also very difficult to make a dataset based on aesthetic cropping artificially. To create a dataset artificially, it is necessary to collect data, which is faced with different aesthetic standards of different people, as different people have different preference, which makes it difficult to collect training data to all people's preference, and if only some specific people's preference is collected, it will lead to all training data to be used for training. If only certain people's preference is collected, it will lead to a certain bias in all the models trained with this training data, so how to collect this data is not simple. The challenge is how to collect this data and to ensure that the training data gives results that are acceptable to most people.

When image cropping it is also very difficult to ensure that the cropped parts play an important role in the image. If important parts are cropped out, it can sometimes result in a picture that is much less effective than the photographer intended. In general, it is a good idea to crop out any meaningless borders and backgrounds to emphasise the subject matter of the image and to make it more aesthetically pleasing. However, if there is something in the cropped area that is important to some people and some teams, it can take away some of the meaning of the image. As shown in Figure 10, the image shows a group of people in a vast desert, but if the cropping is done as shown in the yellow box, the result is a group of people in the desert, which does not reflect the vast desert and does not show the full picture of the desert for the viewer, so the cropping is obviously not very successful and the meaning of the image is greatly reduced.

Figure 10. Picture of a desert.

DOI: 10.54254/2755-2721/4/20230455

Also, if two or more important parts are very far apart or have very different proportions in the diagram, a cropping operation may only yield one of the important parts. As shown in Figure 11, there is both a tower and a UFO in the diagram, and because they are very far apart and the tower is a much larger part of the diagram than the UFO, there are two cropping results in the diagram, which would crop out a very important part of the diagram, which would not be in keeping with the aesthetic view of the two being one. So it is a great challenge to ensure that the aesthetic standards are met while retaining the meaning of the image.

Figure 11. Picture of a tower.

Many images now follow a certain ratio, such as 4:3, 16:9, etc. After the image cropping operation, it is a challenge to adjust the size, or to cropping to a certain size, so that the cropped image looks better.

5. Conclusion

This paper present a relatively comprehensive review of the existing classical networks in the field, provide specific datasets and resources that can be used for training, and provide some evaluation criteria that can be widely applied in the field. More importantly, this paper analyse some of the challenges that need to be addressed in this area.

References

[1] Lu P, Zhang H, Peng X, et al. Aesthetic guided deep regression network for image cropping[J]. Signal Processing: Image Communication, 2019, 77:1-10.

[2] Zhang J, Fan D P, Dai Y, et al. UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders[C]// 2020.

[3] Tu Y, Niu L, Zhao W, et al. Image Cropping with Composition and Saliency Aware Aesthetic Score Map. [C]// National Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (AAAI), 2020.

[4] Cornia M, Pini S, Baraldi L, et al. Automatic Image Cropping and Selection Using Saliency: An Application to Historical Manuscripts[J]. Springer, Cham, 2018.

[5] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In ECCV, 21 – 37. Springer.

[6] Wei, Z.; Zhang, J.; Shen, X.; Lin, Z.; Mech, R.; Hoai, M.; and Samaras, D. 2018. Good view hunting: Learning photo composition from dense view pairs. In CVPR, 5437 – 5446.

DOI: 10.54254/2755-2721/4/20230455

[7] Lu, P.; Zhang, H.; Peng, X.; and Jin, X. 2019a. An end-to-end neural network for image cropping by learning composition from aesthetic photos. arXiv preprint arXiv:1907.01432.
[8] Li, X.; Li, X.; Zhang, G.; and Zhang, X. 2019. Image aesthetic assessment using a saliency symbiosis network. Journal of Electronic Imaging 28(2):023008.
[9] Wang, W., and Shen, J. 2017. Deep cropping via attention box prediction and aesthetics assessment. In ICCV, 2186–2194.
[10] Lu, P.; Zhang, H.; Peng, X.; and Peng, X. 2019b. Aesthetic guided deep regression network for image cropping. Signal Processing: Image Communication.
[11] Zeng H, Li L, Cao Z, et al. Reliable and efficient image cropping: A grid anchor based approach[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5949-5957.
[12] Lu P, Zhang H, Peng X, et al. An end-to-end neural network for image cropping by learning composition from aesthetic photos[J]. arXiv preprint arXiv:1907.01432, 2019.
[13] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A largescale database for aesthetic visual analysis. In CVPR, 2012.
[14] Ming J, Huang S, Duan J, et al. SALICON: Saliency in Context[C]// Computer Vision & Pattern Recognition. IEEE, 2015.
[15] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In CVPR, 2013.
[16] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T.Chen, and B.-Y. Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In WACV, 2017.