(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 February 2020 (20.02.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/034663 A1

(51) International Patent Classification:
G06K 9/78 (2006.01)

(21) International Application Number:
PCT/CN2019/084959

(22) International Filing Date:
29 April 2019 (29.04.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/717,931 13 August 2018 (13.08.2018) US

(71) Applicant: **THE HONG KONG POLYTECHNIC UNIVERSITY** [CN/CN]; Hung Hom, Kowloon, Hong Kong (CN).

(72) Inventors: **ZENG, Hui**; The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (CN). **ZHANG, Lei**; The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (CN). **LI, Lida**; The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (CN).

(74) Agent: **BEIJING PAT SF INTELLECTUAL PROPERTY AGENCY CO. LTD**; 1110-11, A, Floor 11, Chengjian Plaza, 18 Beitaipingzhuang Road, Haidian District, Beijing 100088 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,

(54) Title: GRID-BASED IMAGE CROPPING

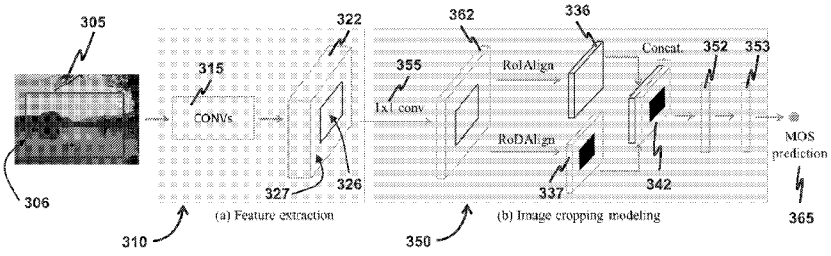


FIG. 3

(57) Abstract: A source image is automatically cropped by generating crop candidates from the source image and selecting a cropped image from the crop candidates. The crop candidates are generated by gridding the source image to form an image grid having grid units each having a grid anchor, selecting two corner regions in the image grid, and creating an individual crop candidate having two corner points respectively selected from two pluralities of grid anchors in the two corner regions. It shrinks the number of crop candidates generated while reducing a likelihood of obtaining similar crop candidates. A suitability score of using the individual crop candidate as the cropped image is determined by utilizing feature information in both region of interest (RoI) and region of discard (RoD). The RoI is a region of feature map containing feature information of the individual crop candidate and the RoD is the remaining region without the RoI.

WO 2020/034663 A1

[Continued on next page]

MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

Grid-Based Image Cropping

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/717,931, filed on August 13, 2018, the disclosure of which is incorporated by reference herein in its entirety.

LIST OF ABBREVIATIONS

[0002]	BDE	Boundary displacement error.
[0003]	CNN	Convolutional neural network.
[0004]	GAICD	Grid Anchor based Image Cropping Database.
[0005]	IoU	Intersection-over-union
[0006]	MOS	Mean opinion score.
[0007]	RoD	Region of discard.
[0008]	RoI	Region of interest.
[0009]	SRCC	Spearman's rank-order correlation coefficient.

LIST OF REFERENCES

[0010] There follows a list of references that are occasionally cited in the specification. Each of the disclosures of these references is incorporated by reference herein in its entirety.

[1] L.-Q. Chen, X. Xie, X. Fan,W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. A visual attention model for adapting images on small displays. In *Multimedia systems*, 9(4):353–364, 2003.

[2] 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*, pages 226–234, 2017.

[3] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma. Learning to compose with professional photographs on the web. In *ACM Multimedia*, pages 37–45, 2017.

[4] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *ACM Multimedia*, pages 291–300, 2010.

[5] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by adversarial learning. In *2018 ACM Multimedia Conference*, arXiv preprint arXiv:1707.05251, 2017.

[6] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. In *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[7] C. Fang, Z. Lin, R. Mech, and X. Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM Multimedia*, pages 1105–1108, 2014.

[8] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y. M. Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. In *IEEE Transactions on Multimedia*, vol. 20, issue 8, Aug. 2018.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] P. J. Huber et al. Robust estimation of a location parameter. In *The annals of mathematical statistics*, 35(1):73–101, 1964.

[12] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. In *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[13] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016.

[14] D. Li, H. Wu, J. Zhang, and K. Huang. A2-RL: Aesthetics aware reinforcement learning for image cropping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8193–8201, 2018.

[15] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, pages 469–478, 2010.

[16] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213, 2011.

[17] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang. Waterloo exploration database: New challenges for image quality assessment models. In *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017.

[18] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, pages 2232–2239, 2009.

[19] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415, 2012.

[20] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *ACM Multimedia*, pages 669–672, 2009.

[21] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM SIGCHI*, pages 771–780, 2006.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[23] F. Stentiford. Attention based auto image cropping. In *ICVS Workshop on Computation Attention & Applications*, 2007.

[24] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM symposium on User interface software and technology*, pages 95–104, 2003.

[25] W. Wang and J. Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, 2017.

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

[27] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In *CVPR*, pages 971–978, 2013.

[28] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe. Weakly supervised photo cropping. In *IEEE Transactions on Multimedia*, 16(1):94–107, 2014.

[29] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen. Probabilistic graphlet transfer for photo cropping. In *IEEE Transactions on Image Processing*, 22(2):802–815, 2013.

[30] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma. Auto cropping for digital photographs. In *ICME*, 2005.

BACKGROUND

Field of the invention

[0011] The present invention generally relates to image processing. Particularly, the present invention relates to a method for automatically cropping a source image based on grids.

Description of related art

[0012] Cropping plays a key role in recomposing taken photos. By removing unwanted region of a photo or readjusting the arrangement of the content, cropping can make a photo much more visually pleasing. Manually cropping by users is currently the most popular and reliable way in most of the camera devices and image processing applications. However, manually cropping photos is time-consuming especially when the number is large, and requiring the users to know the rules of photography composition. A reliable and efficient automatic cropping system can significantly improve the user experience and benefit billions of users.

[0013] Existing image cropping methods can be divided into three categories according to major drives of these methods.

[0014] The first category is directed to attention-driven methods. Earlier methods are mostly attention driven, aiming to identify the major subject or the most informative region of an image. Most of them [1], [18], [23], [24] resort to a saliency detection algorithm (e.g. [12]) to get an attention map of an image, and search a cropping window with the highest attention value. Some methods also employ face detection [30] or gaze interaction [21] to find the important region of an image.

[0015] The second category is directed to aesthetic-driven methods. The aesthetic-driven methods improve the attention-based methods by emphasizing the overall aesthetic quality of an image. These methods [4], [7], [15], [20], [27]-[29], [30] usually design a set of handcrafted features to characterize the widely recognized aesthetic properties or composition rules. Some methods further design quality measures [15], [30] to evaluate the quality of crop candidates, while some resort to training an aesthetic discriminator [4], [20]. The release of two cropping databases [7], [27] facilitates the training of discriminative cropping models. However, the handcrafted features are not strong enough to accurately predict image aesthetics [6].

[0016] The third category is directed to data-driven methods. Most recent methods are data-driven, which train an end-to-end CNN model for image cropping. However, limited by insufficient number of annotated training samples, many methods in this category [2], [5], [6], [8], [14], [25] adopt a general aesthetic classifier trained from image aesthetic databases

such as the ones described in [19] and [16] to help cropping. However, a general aesthetic classifier trained on full images may not be able to reliably evaluate the crops within one image [3], [26]. An alternative strategy is to use pairwise learning to construct more training data [3], [26]. However, annotation of ranking pairs is also very expensive because of the subjective nature of image cropping. Recently, Wei et al. [26] constructed a large scale comparative photo composition dataset using an efficient two-stage annotation protocol, which provides a good training set for pairwise learning. Unfortunately, pairwise learning cannot provide adequate evaluation metrics for image cropping.

[0017] Usually, traditional automatic cropping methods usually first detect the salient objects or regions of an image using some hand-crafted features, then match a lot of crop candidates to a set of pre-defined cropping rules. Both the handcrafted features and manually defined rules have limited capabilities to represent various and complicated composition styles. Data-driven methods together with advanced learning algorithms such as deep neural network provide a more general and powerful solution. However, manually annotating sufficient and reliable data to train a deep neural network is very expensive. Models trained on limited number of images are not stable and have poor generalization capability. Efficiency is another important issue for practical applications. Previous cropping methods search hundreds of thousands of pixel-wise crop candidates, which is very time-consuming.

[0018] There is a need in the art for a more-efficient cropping method.

SUMMARY OF THE INVENTION

[0019] The present invention provides a method for automatically cropping a source image to yield a cropped image. The method comprises the step of generating plural crop candidates from the source image and the step of selecting the cropped image from the generated crop candidates.

[0020] A first aspect of the present invention is to provide embodiments for generating the crop candidates from the source image. A grid anchor based approach is disclosed and is used to generate the crop candidates. Generating the crop candidates using the disclosed grid anchor based approach shrinks the number of crop candidates from millions to less than one hundred in certain situations, thus making an image cropping system based on this approach very efficient. Redundant and unnecessary crop candidates are also discarded, making cropping results more stable.

[0021] In generating the crop candidates from the source image, a preselected portion of the source image is gridded to form an image grid having $M \times N$ grid units. The preselected

portion may be the whole source image or a part thereof. An individual grid unit has a grid anchor for identifying the individual grid unit. The grid anchor is located on or inside the individual grid unit. In one option, the grid anchor is selected to be a center of the individual grid unit.

[0022] After the image grid is formed, a first corner region of size $m \times n$ grid units and a second corner region of size $m' \times n'$ grid units in the image grid are selected. The first and second corner regions are diagonally-opposite to each other. Furthermore, m, m', n and n' are selected such that $M - m - m'$ and $N - n - n'$ are positive. Thereby, first and second pluralities of grid anchors identifying respective grid units in the first corner region and in the second corner region, respectively, are created.

[0023] After the first and second corner regions are selected, an individual crop candidate is created as a rectangular portion of the source image having a first anchor point and a second anchor point. The second anchor point is diagonally-opposite to the first anchor point. Hence, the crop candidates are generated via selecting a plurality of different anchor-point pairs each consisting of respective first and second anchor points. The first and second anchor points in each anchor-point pair are respectively selected from the first and second pluralities of grid anchors rather than from a first plurality of image pixels in the first corner region and a second plurality of image pixels in the second corner region. As a result, a likelihood of obtaining similar crop candidates among the generated crop candidates is advantageously reduced.

[0024] Preferably, the first and second anchor points are selected under a first constraint that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the source image. Optionally, the first constraint may be modified such that the area of the individual crop candidate is not less than the predetermined proportion of a whole area of the image grid.

[0025] Preferably, the first and second anchor points are selected under a second constraint that an aspect ratio of the individual crop candidate is between a predetermined lower limit and a predetermined upper limit inclusively. The aspect ratio is a ratio of a width of the individual crop candidate to a height thereof.

[0026] The first and second constraints may be used together in the selection of anchor-point pair. It is also possible that only one of the constraints is used.

[0027] A second aspect of the present invention is to provide embodiments for selecting the cropped image from the generated crop candidates. An approach of using feature

information contained in both RoI and RoD in evaluating an individual crop candidate is adopted. This approach is shown to be more effective in evaluating the generated crop candidates and enables the image cropping system to handle more complicated scenes possibly appeared in the source image.

[0028] In selecting the cropped image from the generated crop candidates, a suitability score of using the individual crop candidate as the cropped image is determined. The determination of the suitability score is repeated until respective suitability scores for all the crop candidates are determined. A preferred crop candidate having a maximum suitability score among the crop candidates is selected to be the cropped image.

[0029] The suitability score of using the individual crop candidate as the cropped image is determined as follows.

[0030] A whole feature map of the source image is obtained, where the whole feature map contains feature information of the source image. Preferably, the whole feature map is obtained by processing the source image with a plurality of convolutional layers of a CNN.

[0031] The whole feature map is separated into a first feature map containing feature information of a RoI and a second feature map containing feature information of a RoD. The RoI is a region of the whole feature map containing feature information of the individual crop candidate. The RoD is a remaining region of the whole feature map without the RoI.

[0032] The first feature map and the second feature map are resized to form a resized first feature map and a resized second feature map, respectively. In one embodiment, bilinear interpolation is used to resize each of the first and second feature maps. The resized first and second feature maps are concatenated to form a combined feature map such that the combined feature map contains feature information of both the RoI and the RoD. If spatial resolutions of the resized first and second feature maps are different, before concatenation, each of the resized first and second feature maps is processed with a fully connected layer of the CNN for adjusting the resized first and second feature maps to have a same spatial resolution.

[0033] The suitability score is determined according to the combined feature map. Advantageously, feature information of both the RoI and the RoD is utilized in determining the suitability score.

[0034] Other aspects of the present invention are disclosed as illustrated by the embodiments hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0035] FIG. **1** provides an example for illustrating local redundancy in image cropping, indicating that small local changes, such as shifting and scaling, on the cropping window of an acceptable crop (as shown in the bottom-right one in FIG. **1**) are very likely to output acceptable crops too.

[0036] FIG. **2** illustrates the grid anchor based formulation of image cropping in accordance with an exemplary embodiment of the present invention, where M and N are the numbers of grid units used for gridding the source image, and m and n define the adopted range of grid anchors for facilitating content preservation.

[0037] FIG. **3** depicts a CNN architecture for image cropping model learning in accordance with an exemplary embodiment of the present invention.

[0038] FIG. **4** depicts a flow diagram showing exemplary steps used for cropping a source image to yield a cropped image, where the steps include generating crop candidates from the source image and selecting the cropped image from the generated crop candidates.

[0039] FIG. **5** depicts a flow diagram showing exemplary steps used for generating the crop candidates from the source image.

[0040] FIG. **6** depicts a flow diagram showing exemplary steps used for selecting the cropped image from the generated crop candidates, where the steps include determining a suitability score of using a crop candidate as the cropped image.

[0041] FIG. **7** depicts a flow diagram showing exemplary steps used for determining the suitability score of using an individual crop candidate as the cropped image.

[0042] FIG. **8** provides examples for illustrating different choices of image grid coverage on the source image.

[0043] FIG. **9** depicts, as examples, two choices of locations for setting up grid anchors of grid units, where the grid anchors are used as anchor points for defining crop candidates.

[0044] FIG. **10** illustrates two arrangements of locating first and second corner regions in an image grid.

DETAILED DESCRIPTION

[0045] As used herein, “a crop” is synonymous to a cropped image, and means an image cropped from a source image. It is also used herein that “a crop candidate” means an image cropped from a source image, where this cropped image is intended to contend with other crop candidates for selection as a finalized cropped image for the source image according to certain selection criteria.

[0046] Disclosed herein are (1) a novel grid-based automatic cropping system which is very efficient (shrinking the cropping candidates from millions to less than one hundred), flexible (applicable to images taken by any camera devices with any resolution), and stable in various scenes, and (2) a method for evaluating the cropping candidates by evaluating both the RoI and RoD for each cropping candidate.

[0047] Section 1 elaborates embodiments regarding grid anchor based image cropping. Section 2 describes embodiments regarding cropping model learning by a CNN. Section 3 provides experimental results. By condensing the details disclosed in Sections 1-3, the present invention is detailed in Section 3.

1. Grid anchor based image cropping

[0048] Image cropping has a high degree of freedom and there is no unique optimal crop for a given image. There are two practical requirements of a good image cropping system. Firstly, a reliable cropping algorithm should be able to return acceptable results for different settings (e.g., aspect ratio and resolution) rather than one single output. Secondly, the cropping system should be lightweight and efficient to run on devices without powerful computational resource. With these considerations, we develop a grid anchor based formulation for practical image cropping, and consequently construct a new benchmark under this formulation.

1.1. Grid anchor based formulation

[0049] Given an image with resolution $H \times W$, a crop candidate can be defined using its top-left corner (x_1, y_1) and bottom-right corner (x_2, y_2), where $1 \leq x_1 < x_2 \leq H$ and $1 \leq y_1 < y_2 \leq W$. It is easy to calculate that the number of potential crop candidates is $H(H-1)W(W-1)/4$, which is a huge number even for an image of size 100×100. Fortunately, by exploiting the following three properties and requirements of image cropping, the searching space can be significantly reduced, making automatic image cropping a tractable problem.

[0050] ***Local redundancy:*** Image cropping is naturally a problem with local redundancy. As illustrated in FIG. **1**, a set of similar and acceptable crops can be obtained in the neighborhood of a good crop candidate by shifting and/or scaling the size of the cropping widow. Intuitively, we can remove the redundant crop candidates by defining crops on image grid anchors rather than dense pixels. The grid anchor based formulation developed herein is

illustrated in FIG. **2**. We construct an image grid **210** with $M \times N$ grid units on the source image. The image grid **210** is uniform in the sense that all the grid units of are same size. A grid unit **230** is a rectangular block of minimal size in the image grid **210**. Furthermore, the grid unit **230** is selected such that the grid unit **230** is not as small as one image pixel of the source image so as to achieve that the resultant grid anchors are not as dense as the image pixels. We also define the corners (x_1, y_1) **221** and (x_2, y_2) **222** of one crop on the centers of grid units. The two corners **221**, **222** serve as the anchors to generate a representative crop in the neighborhood. Such a formulation largely reduces the number of cropping candidates from $H(H-1)W(W-1)/4$ to $M(M-1)N(N-1)/4$, which can be several orders smaller.

[0051] ***Content preservation:*** Generally, a good crop should preserve the major content of the source image [7]. Cropping should not dramatically change the subject of the source image and should preserve a large proportion of the source image. Therefore, the cropping window should not be too small in order to avoid discarding the image content too much. To this end, we constrain that the anchor points (x_1, y_1) **221** and (x_2, y_2) **222** selected to form a crop are located in two regions with $m \times n$ grid units on the top-left and bottom-right corners of the source image, respectively, as illustrated in FIG. **2**. It further reduces the number of crops from $M(M-1)N(N-1)/4$ to $m(m-1)n(n-1)$.

[0052] The smallest possible crop **240** generated by the scheme developed herein covers about a ratio of $(M-2m)(N-2n)/(MN)$ in area of the source image, which may still be too small to preserve enough image content. Preferably and advantageously, the area of potential crops is further constrained to be no smaller than a certain proportion of the whole area of source image, i.e.

$$S_{crop} \geq \lambda S_{Image}, \tag{1}$$

where S_{crop} and S_{Image} represent the areas of crop and source image, respectively, and

$$\lambda \in \left(\frac{(M-2m)(N-2n)}{MN}, 1 \right) \tag{2}$$

is a predetermined proportion.

[0053] ***Aspect ratio consideration:*** Because of the standard resolution of imaging sensors and displays, most people have been accustomed to the popular aspect ratios such as 16:9, 4:3 and 1:1. Crop candidates which have very different aspect ratios from these popular ones are inconvenient to display and can make people feel uncomfortable. Preferably and advantageously, the aspect ratio of acceptable crop candidates satisfies the following condition:

$$\alpha_1 \leq \frac{W_{crop}}{H_{crop}} \leq \alpha_2, \quad (3)$$

where W_{crop} and H_{crop} are the width and height of a crop. The variables α_1 and α_2 define the range of aspect ratio and we set them to be $\alpha_1 = 0.5$ and $\alpha_2 = 2$ to cover most of the common aspect ratios. With EQN. 3, the final number of crop candidates is less than $m(m-1)n(n-1)/2$.

1.2. Grid anchor based cropping database

[0054] The grid anchor based formulation as developed herein reduces the number of cropping candidates from H^2W^2 to less than $m(m-1)n(n-1)/2$. It enables us to annotate all of the crop candidates for each image. To make the annotation cost as low as possible, we first make a small scale subjective study to find the smallest $\{M, N, m, n\}$ that ensure at least 3 acceptable crops for each image. We collected 100 natural images and invited five volunteers to participate in this study. We set $M = N$ (selected from {16, 14, 12, 10}) and $m = n$ (selected from {5, 4, 3}) to reduce possible combinations. In EQN. 1, λ is set to 0.5. After the tests, we found that $M = N = 12$ and $m = n = 4$ could lead to a good balance between cropping quality and annotation cost. Finally, the number of crop candidates is successfully reduced to be ≤ 90 for images with any resolution. Note that the setting of these parameters mainly aims to reduce annotation cost for training. In the testing stage, it is possible to use a finer image grid to generate more crop candidates.

[0055] With the above settings, we have constructed a GAICD. We first crawled ~50,000 images from the Flickr website. Considering that many images uploaded to Flickr already have good composition, we manually selected ~1,000 images whose composition could be obviously improved, as well as ~200 images with proper composition to ensure the generality of the GAICD. The selected images covered a variety of scenes, weather and lighting conditions. For each image, our annotation toolbox automatically generated all of the crop candidates in an ascending/descending order of aspect ratios. There were 106,860 crop candidates of the 1,236 images in total. The annotators were required to rate the candidates at five scores (from 1 to 5) which represent "bad," "poor," "fair," "good," and "excellent."

[0056] A total of 19 annotators passed our test on photography composition and participated into the annotation. The annotators were either experienced photographers from photography communities or students from the art department of two universities. Each crop was annotated by seven different subjects. The MOS was calculated for each crop candidate

as its ground-truth quality score. We have found that for 94.25% crop candidates in our dataset, the standard deviations of their rating scores are smaller than 1, which confirms the annotation consistency under our grid anchor based formulation.

1.3. Evaluation metrics

[0057] The dense annotations of multiple crops for each image enable us to define more reasonable metrics to evaluate cropping performance compared to the IoU or BDE used in previous databases [2], [7], [27]. We define two new metrics for our database. The first one is average SRCC. The SRCC has been widely used to evaluate the rank correlation between the MOS and model's predictions in image quality and aesthetic assessment [13], [17]. Denote the MOS and model's predictions by $\mathbf{g}$ and $\mathbf{p}$. The SRCC is defined as

$$SRCC(\mathbf{g}, \mathbf{p}) = cov(r_{\mathbf{g}}, r_{\mathbf{p}})/\sigma_{r_{\mathbf{g}}}\sigma_{r_{\mathbf{p}}}, \quad (4)$$

where: $r_{\mathbf{g}}$ and $r_{\mathbf{p}}$ are the rank indexes of $\mathbf{g}$ and $\mathbf{p}$, respectively; and $cov(.)$ and $\sigma(.)$ denote covariance and standard deviation, respectively. The average SRCC is computed by

$$\overline{SRCC} = \frac{1}{T}\sum_{i=1}^{T} SRCC(\mathbf{g}_i, \mathbf{p}_i), \quad (5)$$

where $\mathbf{g}_i$ and $\mathbf{p}_i$ denote the MOS and predictions of all crops in i-th image, and T is the number of testing images.

[0058] Considering that users may care more about whether the returned crops are acceptable or not than an accurate ranking order of all crop candidates, we define a more intuitive metric for practical cropping applications, which we call "return K of top-N accuracy" ($Acc_{K/N}$). For the i-th image associated with a set of pre-defined crop candidates C_i, we suppose that crops, whose MOS rank top-N, are acceptable for users and a cropping model can return at most K ($K \leq N$) crops. Let $S_i(N)$ and $\{c_{i1}, c_{i2}, \cdots, c_{iK}\}$ denote the top-N crop subset and the returned K crops. Furthermore, $Acc_{K/N}$ is defined as

$$Acc_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{j=1}^{K} True\big(c_{i,j} \in S_i(N)\big), \quad (6)$$

where $True(*) = 1$ if * is true, otherwise $True(*) = 0$. We set N to either 5 or 10, and evaluate $K = 1,2,3,4$ ($K > 4$ may lead to unacceptable user experience) for both $N = 5$ and $N = 10$, resulting in two sets of accuracy indexes. Considering that each single index may not be very stable, we further average the multiple indexes in each set, leading to two average accuracy metrics: $\overline{Acc}_{/5}$ and $\overline{Acc}_{/10}$.

2. Cropping model learning

[0059] Limited by insufficient training data, most previous cropping methods focused on how to leverage additional aesthetic databases [3], [5], [25] or how to construct more training pairs [2], [26], paying limited attention to how to design a suitable network for image cropping itself. They usually adopt the standard CNN architecture widely used in object detection. Our GAICD database provides a better platform with much more annotated samples for model training. By carefully considering the special properties of image cropping and its differences from object detection, we design an effective and lightweight module for cropping model learning. The overall architecture is shown in FIG. **3**, where the architecture includes one general feature extraction module **310** and one image cropping module **350**.

2.1. Feature extraction

[0060] As in many previous works [2], [3], [5], [6], [8], [14], [25], [26], we truncate one pre-trained CNN model (e.g. VGG16 [22] or ResNet50 [10]) as the feature extraction module **310**. The spatial arrangement of content and location of objects in an image plays a key role in image composition. For example, the "rule of thirds", which is the most commonly used composition rule, suggests to place important compositional elements at certain locations of an image. Therefore, the feature extraction module **310** needs to preserve sufficient spatial resolution for evaluating image composition in the following cropping module **350**. Truncating at shallower layers can preserve larger spatial resolution but the output feature map **322** may not have enough receptive field to describe large objects in images. (The output feature map **322** is generated by processing a source image **305** by a plurality of convolutional layers **315**. For reference, a cropped image **306** to be considered is also shown on the source image **305**.) Since, as far as we know, no previous work has addressed the effects of due to the aforementioned truncation for the feature extraction module **310**, we conduct extensive experiments to decide the most cost-effective layer to truncate the standard CNN models for image cropping.

2.2. Modeling both the RoI and RoD

[0061] The first significant difference between image cropping and object detection is that object detection only focuses on the RoI (viz., the region of a feature map containing feature information of the cropped image **306** where the feature map contains feature

information of the source image **305**), while cropping also needs to consider the discarded information (i.e. the RoD, the remaining region of the feature map excluding the RoI). On one hand, removing distracting information can significantly improve the composition. On the other hand, cropping out important region can dramatically change or even destroy an image. A CNN model is unavailable to such important information if only the RoI is considered, while additionally modeling the RoD can effectively solve this problem.

[0062] Let F denote the whole feature map **322** outputted by the feature extraction module **310**. It consists of a first feature map **326** in RoI and a second feature map **327** in RoD, denoted by F_{RoI} and F_{RoD}, respectively. We first employ the RoIAlign routing [9] to transform F_{RoI} **326** into F_{RoI}^{A} **336** which has fixed spatial resolution $s \times s$. The F_{RoD} **327** is constructed by removing F_{RoI} **326** from F **322**, namely, setting the values of F_{RoI} **326** to zeros in F **322**. Then RoDAlign (using the same bilinear interpolation as RoIAlign) is performed for F_{RoD} **327**, leading to F_{RoD}^{A} **337**, which has the same spatial resolution as F_{RoI}^{A} **336**. Then F_{RoI}^{A} **336** and F_{RoD}^{A} **337** are concatenated along the channel dimension as one aligned feature map (referred to as a combined feature map **342**) that contains the feature information in both RoI and RoD. The combined feature map **342** is fed into two fully connected layers **352**, **353** for final prediction of a MOS **365**.

2.3. Reducing the channel dimension

[0063] Another difference between image cropping and object detection is that the former does not need to accurately recognize the category of different objects, which allows us to significantly reduce the channel dimension of the feature map. In practice, we find another way of generating F_{RoI}^{A} **336** and F_{RoD}^{A} **337** from F **322** such that we are able to reduce the channel dimension of the feature map **322** (output by the VGG16 model) from 512 to 8 using 1×1 convolution **355** without sacrificing much performance to form a reduced-dimension feature map **362** that is subsequently transformed to F_{RoI}^{A} **336** and F_{RoD}^{A} **337**. The low channel dimension makes the image cropping module **350** very efficient and lightweight. More details can be found in Section 3.

[0064] The Huber loss [11] is employed to regress the MOS of crop candidates:

$$\mathcal{L} = \begin{cases} (g-p)^2, when\ |g-p| \le \delta \\ \delta\left(|g-p| - \frac{1}{2}\delta\right), otherwise, \end{cases} \quad (7)$$

where g and p respectively denote the groundtruth MOS and prediction of one crop, and δ is fixed at 1 throughout the experiments.

3. Experiments

3.1. Implementation details

[0065] We randomly selected 200 images from our GAICD as the testing set and the remaining 1,036 images (containing 89,519 annotated crops in total) are divided into 1,000 images for training and 36 images for validation. In the training stage, our model takes one image and 64 randomly selected crops of it as a batch to input. In the testing stage, the trained model evaluates all the generated crops of one image and output the candidates with the highest prediction scores. To improve the training and testing efficiency, the short side of input images is resized to 256 and the other side is adjusted to the nearest multiple of 32 (if necessary). The standard Adam optimizer with the default parameters was employed to train our model for 40 epoches. Learning rate was fixed at $1e^{-4}$ throughout our experiments. We randomly adjusted the contrast and saturation of the source images for data augmentation in the training stage. The MOS were normalized by removing their mean and dividing by their standard deviation across the database.

3.2. Feature extraction module **310**

[0066] We first conducted a set of experiments to determine the appropriate feature extraction module on two pre-trained models (VGG16 [22] and ResNet50 [10]). For each model, we truncated at five different layers, which cover various stride and receptive field, and evaluated their effects on cropping performance. The image cropping module (including both the RoI and RoD) was fixed for all cases for fair comparison. The truncating layer (tlayer), stride (str), receptive field (rf), parameter size (par) and cropping performance ($\overline{SRCC}$, $\overline{Acc}_{/5}$ and $\overline{Acc}_{/10}$) for each module are reported in Table 1.

Table 1.

Model	Tlayer	Str	rf	Par (Mbit)	$\overline{SRCC}$	$\overline{Acc}_{/5}$	$\overline{Acc}_{/10}$
	C4_1	8	60	11.1			
	C4_2	8	76	20.1			
vgg16	C5_1	16	132	38.1	0.735	46.9	64.4

	C5_2	16	164	47.1			
	Pool5	32	212	56.1	0.683	41.1	57.6
resnet50	C3_2	8	67	3.4			
	C3_4	8	99	5.6			
	C4_3	16	195	19.9			
	C4_6	16	291	32.7			
	C5_1	32	355	55.8			

[0067] We can make three observations from Table 1. First, for both the VGG16 and ResNet50 models, using too large strides (e.g., 32) results in very poor accuracy. It is because overly downsampling the feature map leads to loss of important spatial information for image cropping. Specifically, for the input image of resolution 256×256, downsampling with stride 32 results in feature maps of size 8×8, and consequently the feature map of a crop candidate may only have a spatial resolution of 4×4, which is insufficient to generate accurate crops. Both stride 8 and stride 16 seem to be acceptable under our input image size. Second, a too small receptive field leads to dissatisfactory performance too. Increasing the receptive field can significantly improve the cropping accuracy at a cost of deeper architecture and more parameters. The performance plateaus when the receptive field is increased to more than half of the image size. It is worth noting that the above observations on stride and receptive field are based on certain input image size (short side equal to 256 in our experiments), which may provide a good reference for other input sizes. Finally, the VGG16 models generally outperform the ResNet50 models. It may be because the ResNet50 models are easier to overfit on our database. We thus choose the VGG16 model (truncated at conv5_1 layer) as the feature extraction module in the following experiments.

3.3. Image cropping module **350**

[0068] We then conducted ablation study on the developed image cropping module, including the effects of parameter size, RoI and RoD.

[0069] ***Parameter size:*** There are two key parameters in the image cropping module: spatial resolution ($s \times s$) of the aligned feature map and channel dimension (cdim) after dimension reduction. Table 2 below reports the cropping performance of using different

$s \times s$ and cdim. The number of filters (nfilter) was fixed at 512 for the followed fully connected layers. The VGG16 model (truncated at conv5_1) is employed as the feature extraction module for all cases. The parameter size (par (Mbit)) of the image cropping module (including two fully connected layers with $s \times s \times (s * cdim) \times 512$ and $1 \times 1 \times 512 \times 512$ kernels) is reported for each case. We first found that small s (e.g. 3 or 5) resulted in obviously worse performance. It again proves the importance of sufficient spatial information for image cropping. The selection $s = 9$ seems to be an appropriate choice since further increasing the value will not bring obvious improvements. The channel dimension of feature maps can be significantly reduced for the problem of image cropping. As can be seen from Table 2, the performance is still reasonable even if we reduce the channel dimension to 1 (note that VGG16 output 512 channels of feature maps). The low channel dimension makes the developed image cropping module both efficient and lightweight. In the following experiments, we chose 8 as the reduced channel dimension which has a good tradeoff between cost and efficacy. Under this setting, the whole image cropping module has only 3.53 Mbits parameters.

Table 2.

$s \times s$	Cdim	nfilter	Par (Mbit)	$\overline{SRCC}$	$\overline{Acc}_{/5}$	$\overline{Acc}_{/10}$
3×3	8	512	1.28	0.689	42.4	58.9
5×5	8	512	1.78	0.711	44.6	61.5
7×7	8	512	2.53	0.725	45.4	63.1
9×9	8	512	3.53	0.735	46.6	65.5
11×11	8	512	4.78	0.736	46.8	65.6
9×9	32	512	11.13	0.733	46.4	65.3
9×9	16	512	6.06	0.736	46.8	65.8
9×9	8	512	3.53	0.735	46.6	65.5
9×9	4	512	2.27	0.731	45.9	65.1
9×9	2	512	1.63	0.719	45.1	64.1
9×9	1	512	1.32	0.706	43.8	62.6

[0070] ***RoI and RoD:*** We made an ablation study on the role of RoI and RoD. The results of using only RoI, only RoD and both of them are reported in Table 3 below. As can be seen, modeling only the RoD results in very poor accuracy, modeling only the RoI performs much better, while simultaneously modeling the RoI and RoD achieves the best cropping accuracy in all cases. This corroborates our analysis that cropping needs to consider both the RoI and RoD.

Table 3.

Module	$\overline{SRCC}$	$\overline{Acc}_{/5}$	$\overline{Acc}_{/10}$
RoD	0.557	29.8	44.5
RoI	0.706	44.8	62.9
RoI+RoD	0.735	46.6	65.5

4. Present invention

[0071] The present invention is obtained and further developed according to the details and embodiments disclosed above in Sections 1-3.

[0072] The present invention provides a method for automatically cropping a source image to yield a cropped image. FIG. **4** depicts a flowchart showing exemplary steps of the disclosed method. In a step **410**, plural crop candidates are generated from the source image. After the crop candidates are generated, the cropped image is selected from the generated crop candidates in a step **420**.

[0073] In a first aspect of the present invention, embodiments of the step **410** are provided. FIG. **5** depicts a flowchart showing exemplary steps used for generating the crop candidates from the source image. As shown in FIG. 5, exemplarily the step **410** includes steps **510**, **520** and **530**.

[0074] In the step **510**, a preselected portion of the source image is gridded to form an image grid having $M \times N$ grid units. The grid units are of same size and each grid unit is rectangular in shape. Optionally, a square grid unit may be used. The size and shape of the individual grid unit are determined by those skilled in the art according to practical situations, such as the predicted size of object of interest in the source image, and the computation power available for performing the image-cropping task.

[0075] An individual grid unit has a grid anchor for identifying the individual grid unit. The grid anchor is located on or inside the individual grid unit. In a convenient choice, the grid anchor is selected as a center of the individual grid unit, and this grid anchor is located inside the individual grid unit. In another choice, the grid anchor is a top-left corner of the individual grid unit, so that this grid anchor is located on the individual grid unit, and equivalently, located on a boundary of the individual grid unit. For illustration, FIG. **8** depicts that (a) top left corners **831** of grid units are used as grid anchors, and (b) grid unit centers **832** are used as grid anchors. Those skilled in the art will appreciate that other choices of location in the individual grid unit are possible. As mentioned above, the size of the individual grid unit is not selected to be one image pixel of the source image in order that the resultant grid anchors are not as dense as the image pixels.

[0076] Although Section 1.1 mentions that the whole source image is gridded so that the image grid and the source image have the same size, the present invention is not limited only to the case that the preselected portion of the source image is the whole source image. In some practical situations, it may be advantageous to grid only a certain portion of the source image. For example, if it is known beforehand that an object of interest never appears in a certain region of the source image, such as a region near the image boundary, then the preselected portion of the source image may be selected to be the whole source image excluding the aforementioned certain region. As examples, FIG. **9** shows three image grids **921**, **922**, **923** each obtained by gridding only a central region of a source image **910**.

[0077] In the step **520**, a first corner region of size $m \times n$ grid units and a second corner region of size $m' \times n'$ grid units in the image grid are selected. The first and second corner regions are diagonally-opposite to each other. In one choice, also shown in FIG. **10** as a first arrangement **1010**, the first and second corner regions are located at a top-left corner and a bottom-right corner, respectively, of the image grid. Alternatively, as shown in FIG. **10** as a second arrangement **1020**, the first and second corner regions are located at a bottom-left corner and a top-right corner, respectively, of the image grid. Furthermore, m, m', n and n' are selected such that $M - m - m'$ and $N - n - n'$ are positive. This condition is imposed to ensure that any two anchor points (namely, two diagonally-opposite corner points of a rectangle that defines a crop candidate) respectively selected from the first and second corner regions form a unique crop candidate such that no another pair of two anchor points selected from the two corner regions can form this crop candidate. In one special case, the first and second corner regions are of equal size, i.e. $m = m'$ and $n = n'$. This special case is

considered in Section 1.1. However, the present invention is not limited only to this special case. It is possible that $m \neq m'$ or $n \neq n'$. By establishing the first and second corner regions in the image grid, first and second pluralities of grid anchors identifying respective grid units in the first corner region and in the second corner region, respectively, are created.

[0078] In the step **530**, the crop candidates are generated. Particularly, an individual crop candidate is created as a rectangular portion of the source image having a first anchor point and a second anchor point. The individual crop candidate is created in a manner as mentioned in Section 1.1. If (x_1, y_1) and (x_2, y_2) are respectively the first and second anchor points, viz., the coordinates of image pixels on the source image, then the individual crop candidate is formed as a rectangular portion of the source image having four corners at image-pixel positions (x_1, y_1), (x_1, y_2), (x_2, y_2) and (x_2, y_1). For convenience, denote that the first and second anchor points form an anchor-point pair. The crop candidates are generated via selecting a plurality of different anchor-point pairs. Note that the first and second anchor points in each anchor-point pair are respectively selected from the first and second pluralities of grid anchors, but not from a first plurality of image pixels in the first corner regions and a second plurality of image pixels in the second corner region. Advantageously, a likelihood of obtaining similar crop candidates among the generated crop candidates is reduced as explained in Section 1.1 regarding local redundancy.

[0079] As elaborated in Section 1.1 regarding content preservation, preferably the first and second anchor points are selected under a first constraint that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the source image. The first constraint is given by EQN. 1 above. The predetermined proportion is given by λ of EQN. 1. To take into account that the first and second corner regions may be of different size, EQN. 2 may be generalized to

$$\lambda \in \left(\frac{(M - m - m')(N - n - n')}{MN}, 1 \right) \tag{7}$$

for enabling those skilled in the art to select a usable value of λ.

[0080] It is possible that not the whole source image is gridded in forming the image grid in the step **510**. In this situation, the first constraint may be modified to that the first and second anchor points are selected under a constraint that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the image grid. Under this modified first constraint, EQN. 7 is still applicable whereas S_{Image} in EQN. 1 is replaced by the whole area of the image grid.

[0081] As elaborated in Section 1.1 regarding aspect ratio consideration, preferably the first and second anchor points are selected under a second constraint that an aspect ratio of the individual crop candidate is between a predetermined lower limit and a predetermined upper limit inclusively. The aspect ratio is a ratio of a width of the individual crop candidate to a height thereof. The second constraint is reflected in EQN. 3, in which the predetermined lower and upper limits are α_1 and α_2, respectively. As one practical choice mentioned in Section 1.1, the predetermined lower and upper limits may be selected to be $\alpha_1 = 0.5$ and $\alpha_2 = 2$, respectively.

[0082] The first and second constraints may be used together in the selection of anchor-point pair. It is also possible that only one of the constraints is used. Similarly, if the image grid does not occupy the whole source image, the modified first constraint and the second constraint may or may not be used together in the selection of anchor-point pair.

[0083] When compared to generating crop candidates on the pixel-wise manner, the grid anchor based approach via executing the steps **510**, **520** and **530** can shrink the number of crop candidates by several orders of magnitude.

[0084] In a second aspect of the present invention, embodiments of the step **420** are provided. In the step **420**, the cropped image is selected from the crop candidates generated in the step **410**. Although the crop candidates are preferably and advantageously generated through executing the steps **510**, **520**, **530**, in the present invention it is not intended that the crop candidates used in executing the step **420** are generated only from the steps **510**, **520**, **530**. The sequence of the steps **510**, **520**, **530** only forms certain preferred embodiments of the step **410**. It is possible that the crop candidates are generated by another embodiment of the step **410**.

[0085] FIG. **6** depicts a flow diagram showing exemplary steps used in the step **420** for selecting the cropped image from the generated crop candidates. First, a suitability score of using an individual crop candidate as the cropped image is determined, or estimated, in a step **610**. The suitability score is to quantify whether the individual crop candidate is suitable to be used as the cropped image. Particularly, the suitability score is a degree of suitability that the individual crop candidate is suitable to be used as the cropped image. One example of the suitability score is a MOS. In a step **620**, the step **610** is repeated until all the generated crop candidates are processed such that each of the generated crop candidates has a respective suitability score determined. After all the suitability scores for the crop candidates are obtained, a preferred crop candidate having a maximum suitability score among the

suitability scores determined for the crop candidates is identified in a step **630**. The preferred crop candidate is selected as the cropped image.

[0086] An exemplary embodiment of the step **610** is illustrated as follows with an aid of FIGS. **3** and **7**. FIG. 7 depicts a flow diagram showing exemplary steps used for determining the suitability score of using the individual crop candidate as the cropped image. As shown in FIG. 7, exemplarily the step **610** includes steps 7**10**, **720**, **730**, **735**, **740**, **745**, **750** and **760**. Advantageously, the steps 7**10**, **720**, **730**, **735**, **740**, **745**, **750** and **760** are conveniently implemented by a CNN. FIG. **3**, which depicts a CNN model, is occasionally referenced.

[0087] In the step **710**, a whole feature map of the source image is obtained. The whole feature map contains feature information of the source image. Details of the step **710** are elaborated in Section 2.1. In one CNN implementation as shown in FIG. **3**, the whole feature map **322** is obtained by processing the source image **305** with a plurality of convolutional layers **315**.

[0088] The whole feature map obtained in the step **710** contains feature information of a RoI and of a RoD. The RoI is a region of the whole feature map containing feature information of the individual crop candidate (i.e. the cropped portion of the source image). The RoD is a remaining region of the whole feature map without the RoI.. In the step **720**, the whole feature map is separated into a first feature map and a second feature map. The first feature map contains feature information of the RoI. The second feature map contains feature information of the RoD. As shown in FIG. **3**, the whole feature map **322** consists of the first feature map **326** and the second feature map **327**. Implementation details of separating the first and second feature maps **326**, **327** from the whole feature map **322** are provided in Section 2.2.

[0089] Afterwards, the first feature map and the second feature map are resized to form a resized first feature map in the step **730** and a resized second feature map in the step **735**, respectively. Details of the steps **730** and **735** are elaborated in Section 2.2. In one embodiment, each of the first and second feature maps is resized by bilinear interpolation.

[0090] If the spatial resolutions of the resized first and second feature maps are the same, the resized first and second feature maps are directly concatenated to form a combined feature map in the step **750**. Note that the combined feature map contains feature information of both the RoI and the RoD. As shown in FIG. **3**, the first feature map **336** and the second feature map **337** both under the same spatial resolution are directly concatenated to form the combined feature map **342**. If, on the other hand, the spatial resolutions of the resized first and second feature maps are not the same, the resized first and second feature maps are

adjusted to align in spatial resolution in the steps **740** and **745**, respectively before concatenation proceeds in the step **750**. In one embodiment, each of the resized first and second feature maps is processed by a fully connected layer in a CNN (not shown in FIG. **3**) for adjusting the resized first and second feature maps to have a same spatial resolution. In order to avoid the steps **740** and **745**, therefore, it is preferable to resize the first and second feature maps in the steps **730** and **735** to have the same spatial resolution.

[0091] In the step **760**, the suitability score is determined according to the combined feature map. Feature information of both the RoI and the RoD contained in the combined feature map is advantageously used in determining the suitability score. In one embodiment, one or more fully connected layers in a CNN are used for determining the suitability score from the combined feature map. As an example shown in the CNN implementation of FIG. **3**, the two fully connected layers **352**, **353** are used to process the combined feature map **342** to give a resultant MOS.

[0092] The embodiments disclosed herein may be implemented using computing devices, such as computers, computing servers, general purpose processors, specialized computing processors, digital signal processors, processors specialized in computing convolution products for images, programmable logic devices and field programmable gate arrays (FPGAs), where the computing devices are configured or programmed according to the teachings of the present disclosure. Computer instructions or software codes running in the computing devices can readily be prepared by practitioners skilled in the software or electronic art based on the teachings of the present disclosure.

[0093] The present invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The present embodiment is therefore to be considered in all respects as illustrative and not restrictive. The scope of the invention is indicated by the appended claims rather than by the foregoing description, and all changes that come within the meaning and range of equivalency of the claims are therefore intended to be embraced therein.

CLAIMS

What is claimed is:

1. A method for automatically cropping a source image to yield a cropped image, the method comprising the step of generating plural crop candidates from the source image and the step of selecting the cropped image from the generated crop candidates, wherein the step of generating the crop candidates from the source image comprises the steps of:

   gridding a preselected portion of the source image to form an image grid having $M \times N$ grid units, an individual grid unit having a grid anchor for identifying the individual grid unit, the grid anchor being located on or inside the individual grid unit;

   selecting a first corner region of size $m \times n$ grid units and a second corner region of size $m' \times n'$ grid units in the image grid, the first and second corner regions being diagonally-opposite to each other, m, m', n and n' being selected such that $M - m - m'$ and $N - n - n'$ are positive, whereby first and second pluralities of grid anchors identifying respective grid units in the first corner region and in the second corner region, respectively, are created; and

   creating an individual crop candidate as a rectangular portion of the source image having a first anchor point and a second anchor point diagonally-opposite thereto so as to generate the crop candidates via selecting a plurality of different anchor-point pairs each consisting of respective first and second anchor points, wherein the first and second anchor points in each anchor-point pair are respectively selected from the first and second pluralities of grid anchors rather than from a first plurality of image pixels in the first corner regions and a second plurality of image pixels in the second corner region for reducing a likelihood of obtaining similar crop candidates among the generated crop candidates.

2. The method of claim 1, wherein the first and second anchor points are selected under a constraint that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the source image.

3. The method of claim 1, wherein the first and second anchor points are selected under a constraint that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the image grid.

4. The method of claim 1, wherein the first and second anchor points are selected under a constraint that an aspect ratio of the individual crop candidate is between a predetermined lower limit and a predetermined upper limit inclusively, the aspect ratio being a ratio of a width of the individual crop candidate to a height thereof.

5. The method of claim 1, wherein the first and second anchor points are selected under first and second constraints, the first constraint being that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the source image, the second constraint being that an aspect ratio of the individual crop candidate is between a predetermined lower limit and a predetermined upper limit inclusively, the aspect ratio being a ratio of a width of the individual crop candidate to a height thereof.

6. The method of claim 1, wherein the first and second anchor points are selected under first and second constraints, the first constraint being that an area of the individual crop candidate is not less than a predetermined proportion of a whole area of the image grid, the second constraint being that an aspect ratio of the individual crop candidate is between a predetermined lower limit and a predetermined upper limit inclusively, the aspect ratio being a ratio of a width of the individual crop candidate to a height thereof.

7. The method of any of claims 1-6, wherein the grid anchor of the individual grid unit is selected to be a center of the individual grid unit.

8. The method of any of claims 1-6, wherein $m = m'$ and $n = n'$.

9. The method of any of claims 1-6, wherein the first and second corner regions are located at a top-left corner and a bottom-right corner, respectively, of the image grid.

10. The method of any of claims 1-6, wherein the first and second corner regions are located at a bottom-left corner and a top-right corner, respectively, of the image grid.

11. The method of claim 1, wherein the preselected portion of the source image is the whole source image.

12. The method of any of claims 1-6, wherein the step of selecting the cropped image from the generated crop candidates comprises the steps of:

for an individual crop candidate, determining a suitability score of using the individual crop candidate as the cropped image;

repeating the step of determining the suitability score until respective suitability scores for all the crop candidates are determined; and

selecting a preferred crop candidate having a maximum suitability score among the crop candidates as the cropped image;

wherein the step of determining the suitability score of using the individual crop candidate as the cropped image comprises the steps of:

obtaining a whole feature map of the source image, wherein the whole feature map contains feature information of the source image;

separating the whole feature map into a first feature map containing feature information of a region of interest (RoI) and a second feature map containing feature information of a region of discard (RoD), wherein the RoI is a region of the whole feature map containing feature information of the individual crop candidate, and the RoD is a remaining region of the whole feature map without the RoI;

resizing the first feature map and the second feature map to form a resized first feature map and a resized second feature map, respectively;

concatenating the resized first and second feature maps to form a combined feature map such that the combined feature map contains feature information of both the RoI and the RoD; and

determining the suitability score according to the combined feature map for utilizing feature information of both the RoI and the RoD in determining the suitability score.

13. The method of claim 12, wherein each of the first and second feature maps is resized by bilinear interpolation.

14. The method of claim 12, wherein the whole feature map is obtained by processing the source image with a plurality of convolutional layers of a convolutional neural network (CNN).

15. The method of claim 14, wherein the step of determining the suitability score of using the individual crop candidate as the cropped image further comprises the step of:

before concatenation, processing each of the resized first and second feature maps with a fully connected layer of the CNN for adjusting the resized first and second feature maps to have a same spatial resolution.

16. A method for automatically cropping a source image to yield a cropped image, the method comprising the step of generating plural crop candidates from the source image and the step of selecting the cropped image from the generated crop candidates, wherein the step of selecting the cropped image from the generated crop candidates comprises the steps of:

for an individual crop candidate, determining a suitability score of using the individual crop candidate as the cropped image;

repeating the step of determining the suitability score until respective suitability scores for all the crop candidates are determined; and

selecting a preferred crop candidate having a maximum suitability score among the crop candidates as the cropped image;

wherein the step of determining the suitability score of using the individual crop candidate as the cropped image comprises the steps of:

obtaining a whole feature map of the source image, wherein the whole feature map contains feature information of the source image;

separating the whole feature map into a first feature map containing feature information of a region of interest (RoI) and a second feature map containing feature information of a region of discard (RoD), wherein the RoI is a region of the whole feature map containing features of the individual crop candidate, and the RoD is a remaining region of the whole feature map without the RoI;

resizing the first feature map and the second feature map to form a resized first feature map and a resized second feature map, respectively;

concatenating the resized first and second feature maps to form a combined feature map such that the combined feature map contains feature information of both the RoI and the RoD; and

determining the suitability score according to the combined feature map for utilizing feature information of both the RoI and the RoD in determining the suitability score.

17. The method of claim 16, wherein each of the first and second feature maps is resized by bilinear interpolation.

18. The method of claim 16, wherein the whole feature map is obtained by processing the source image with a plurality of convolutional layers of a convolutional neural network (CNN).

19. The method of claim 18, wherein the step of determining the suitability score of using the individual crop candidate as the cropped image further comprises the step of:

before concatenation, processing each of the resized first and second feature maps with a fully connected layer of the CNN for adjusting the resized first and second feature maps to have a same spatial resolution.

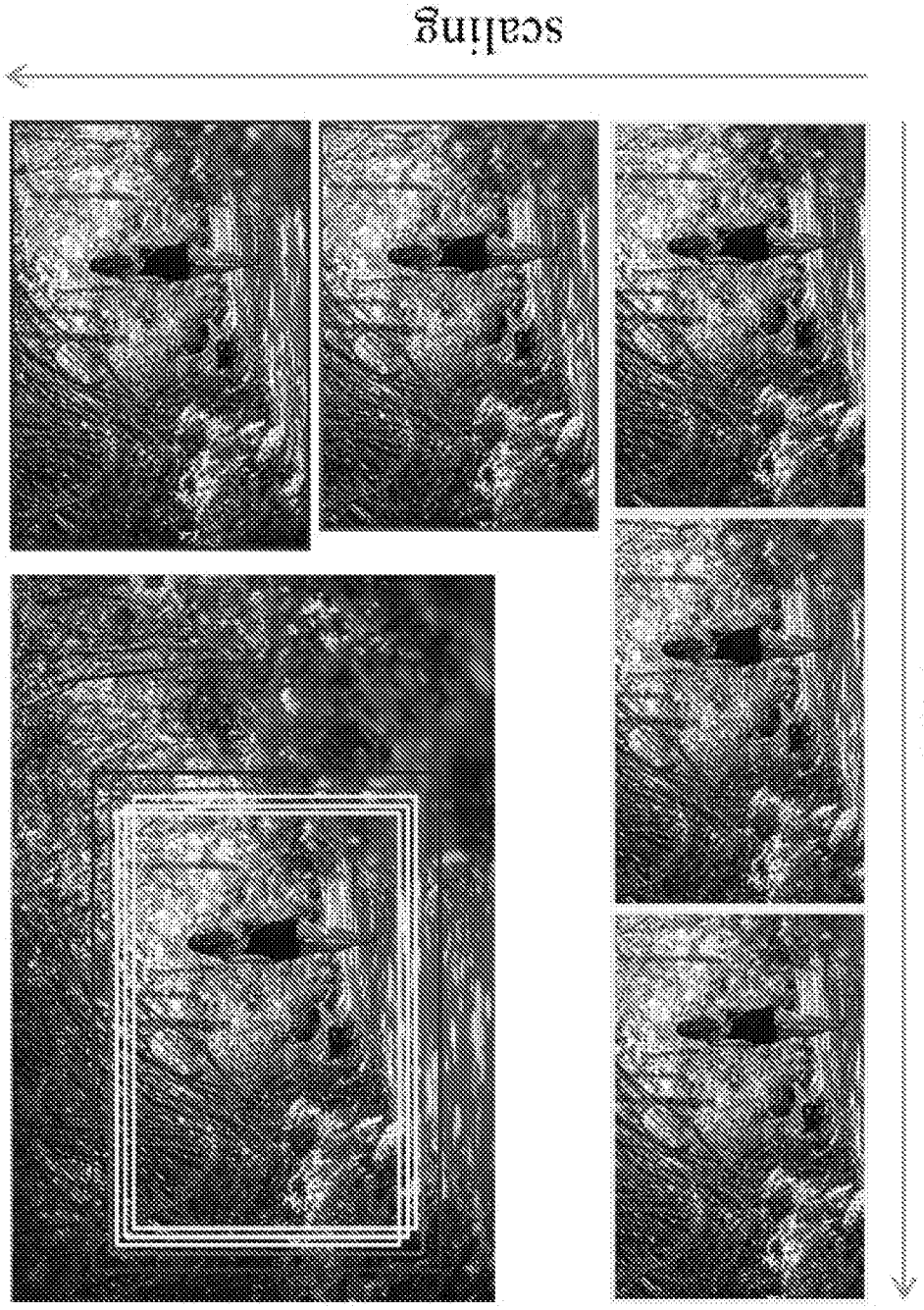


FIG. 1

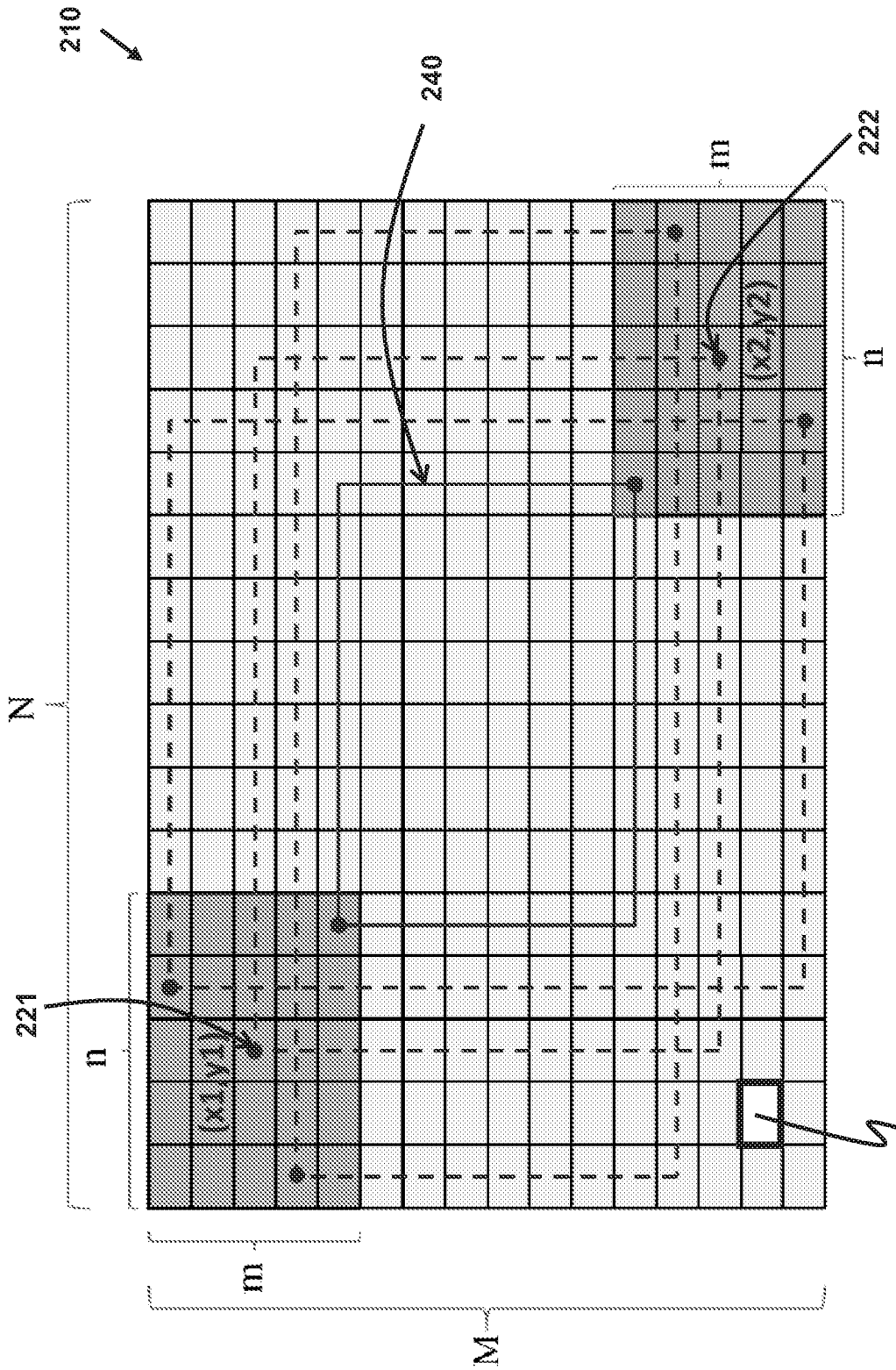


FIG. 2

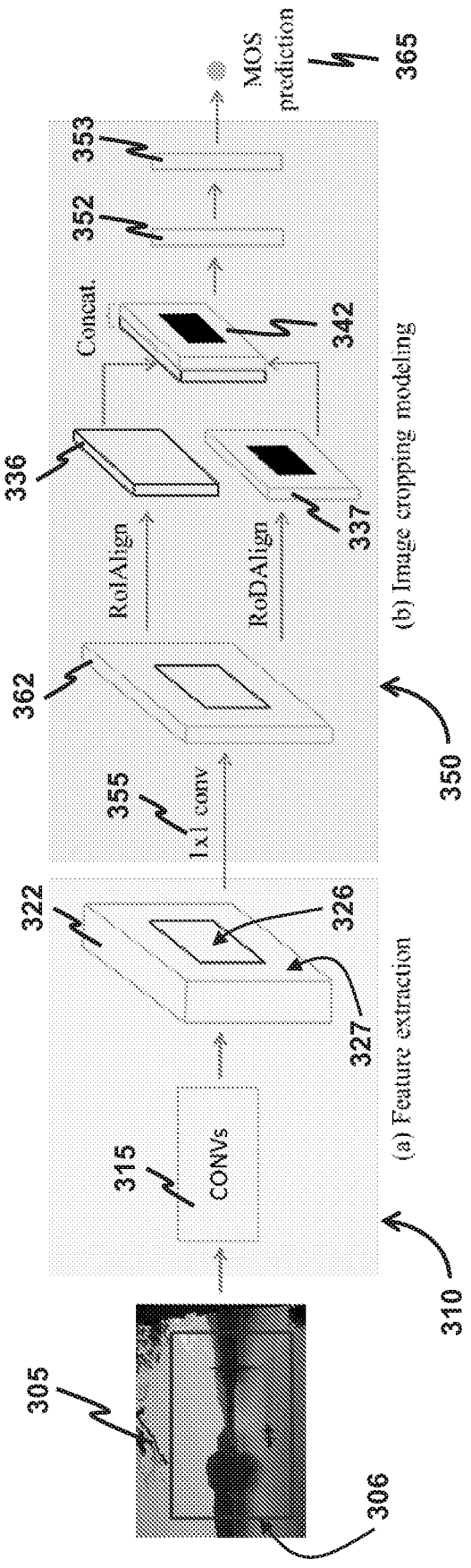

305
306
310
315
CONVs
322
326
327
(a) Feature extraction
355
1x1 conv
362
RoIAlign
RoDAlign
336
337
Concat
342
352
353
MOS prediction
365
(b) Image cropping modeling
350

FIG. 3

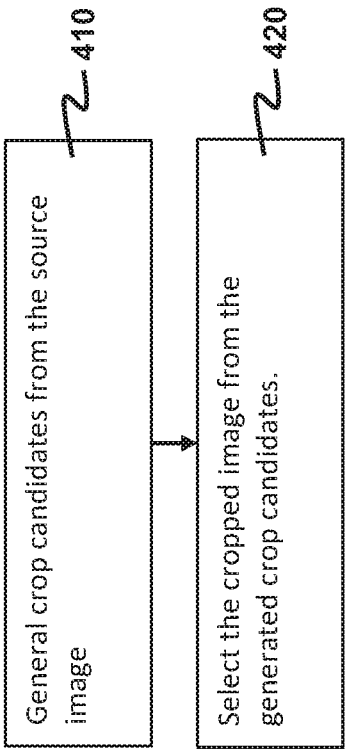

410
General crop candidates from the source image
420
Select the cropped image from the generated crop candidates.

FIG. 4

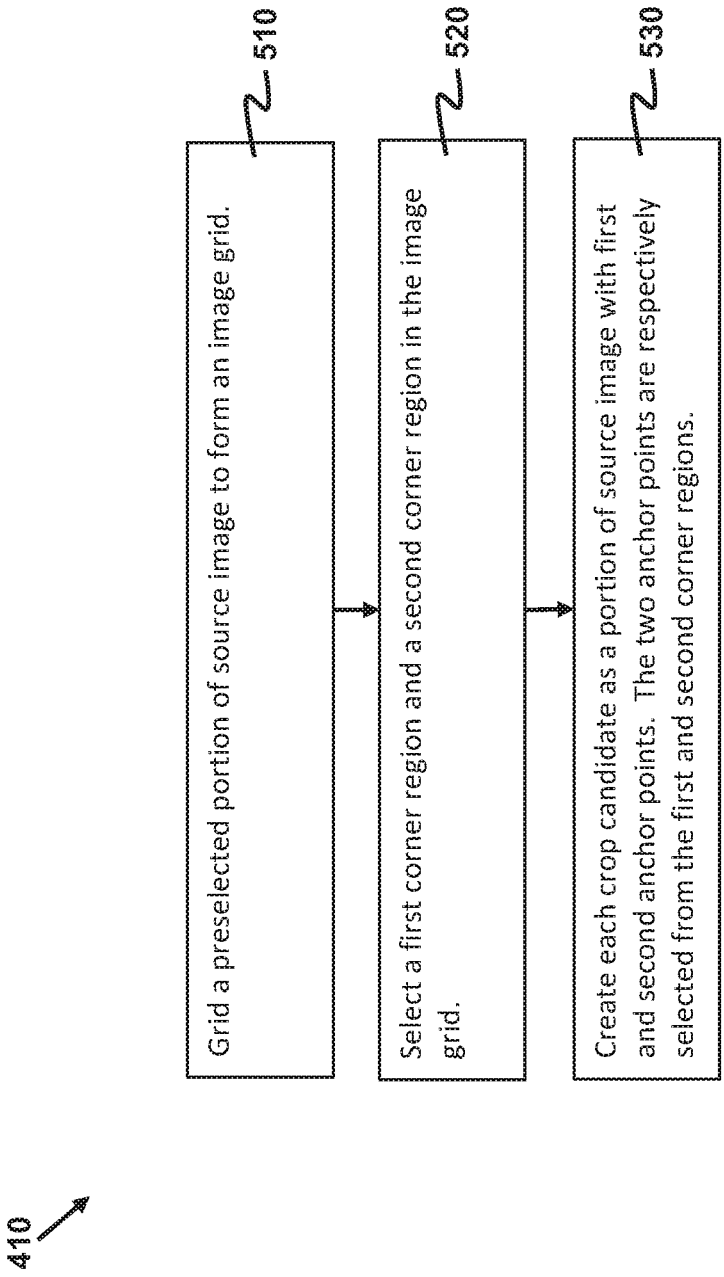

410
510
Grid a preselected portion of source image to form an image grid.
520
Select a first corner region and a second corner region in the image grid.
530
Create each crop candidate as a portion of source image with first and second anchor points. The two anchor points are respectively selected from the first and second corner regions.

FIG. 5

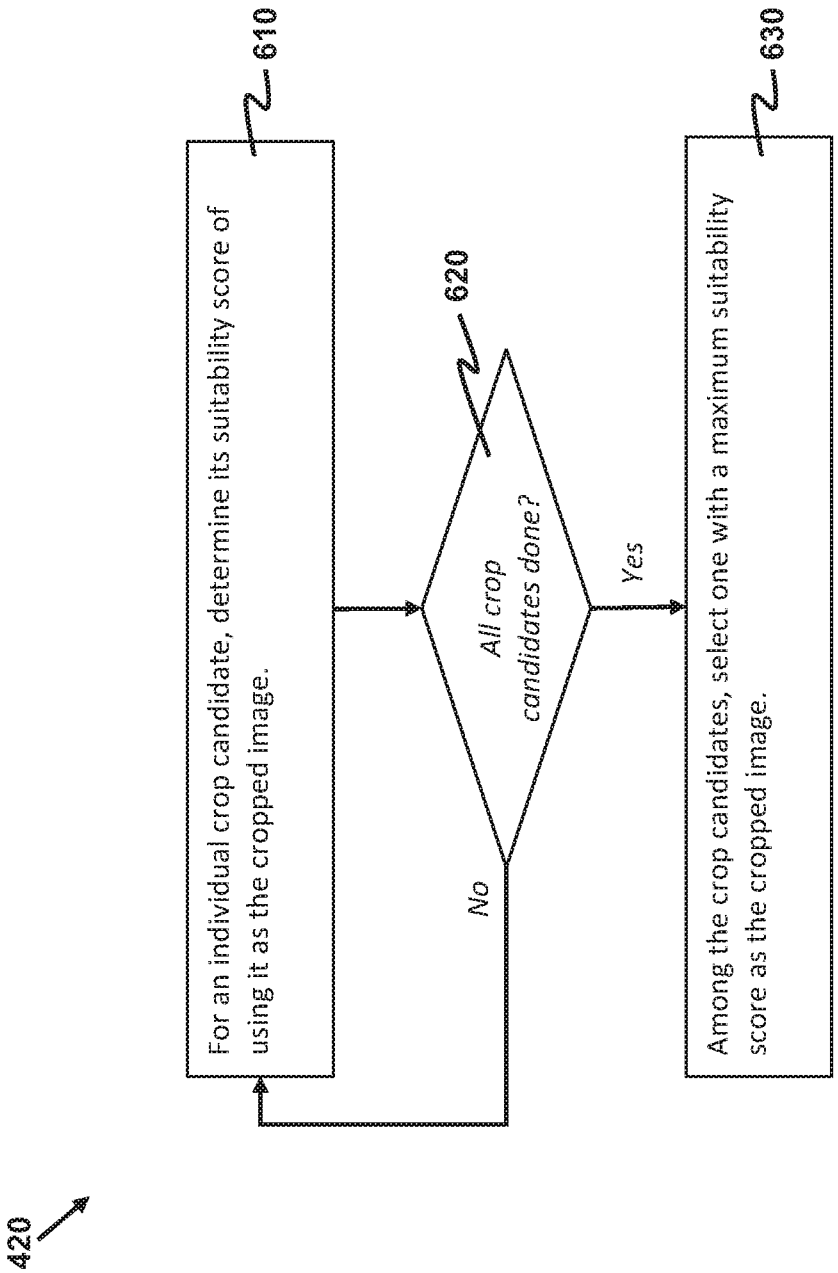

420
610
For an individual crop candidate, determine its suitability score of using it as the cropped image.
620
All crop candidates done?
No
Yes
630
Among the crop candidates, select one with a maximum suitability score as the cropped image.

FIG. 6

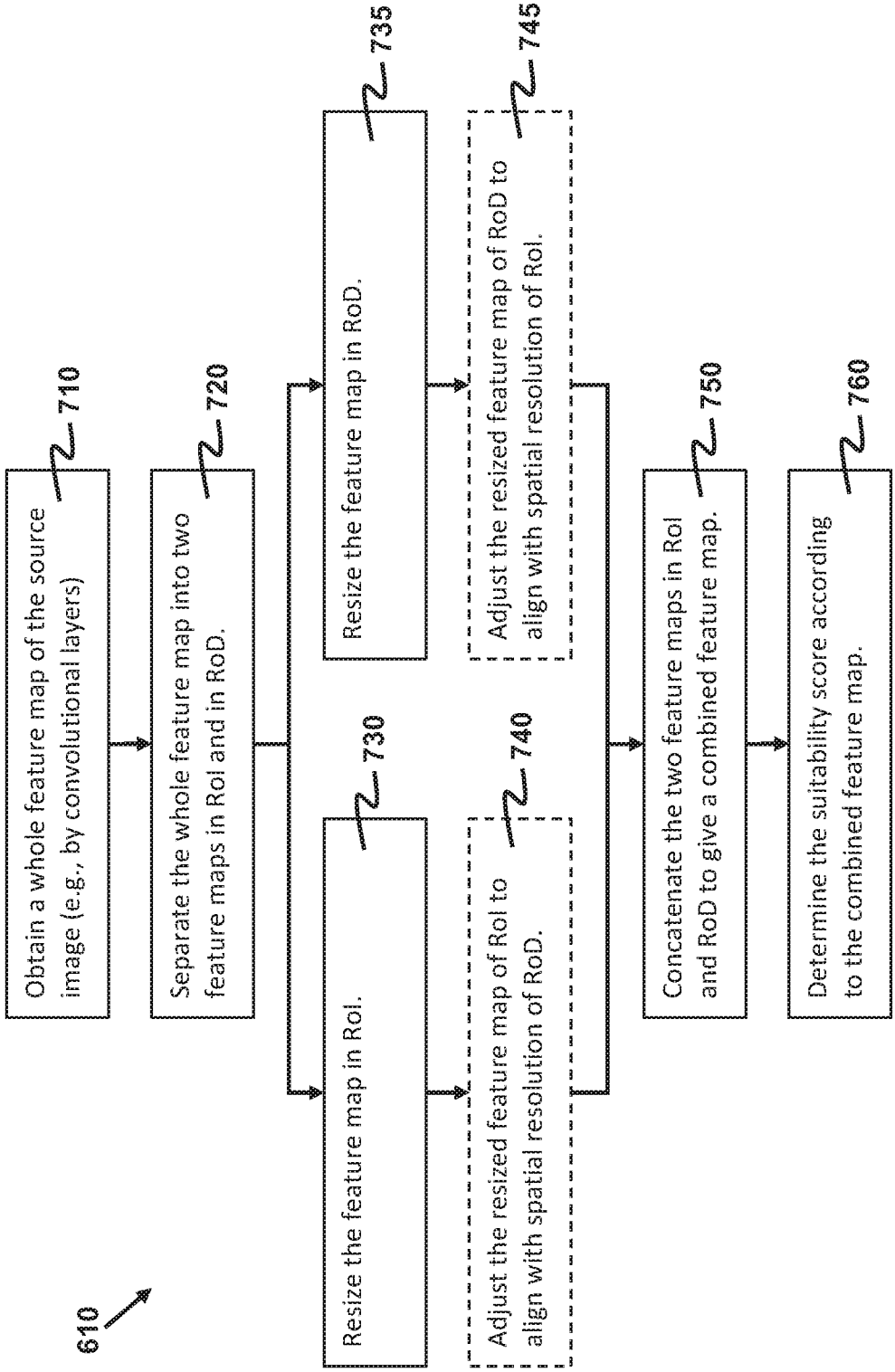

610
Obtain a whole feature map of the source image (e.g., by convolutional layers)
710
Separate the whole feature map into two feature maps in RoI and in RoD.
720
Resize the feature map in RoI.
730
Resize the feature map in RoD.
735
Adjust the resized feature map of RoI to align with spatial resolution of RoD.
740
Adjust the resized feature map of RoD to align with spatial resolution of RoI.
745
Concatenate the two feature maps in RoI and RoD to give a combined feature map.
750
Determine the suitability score according to the combined feature map.
760

FIG. 7

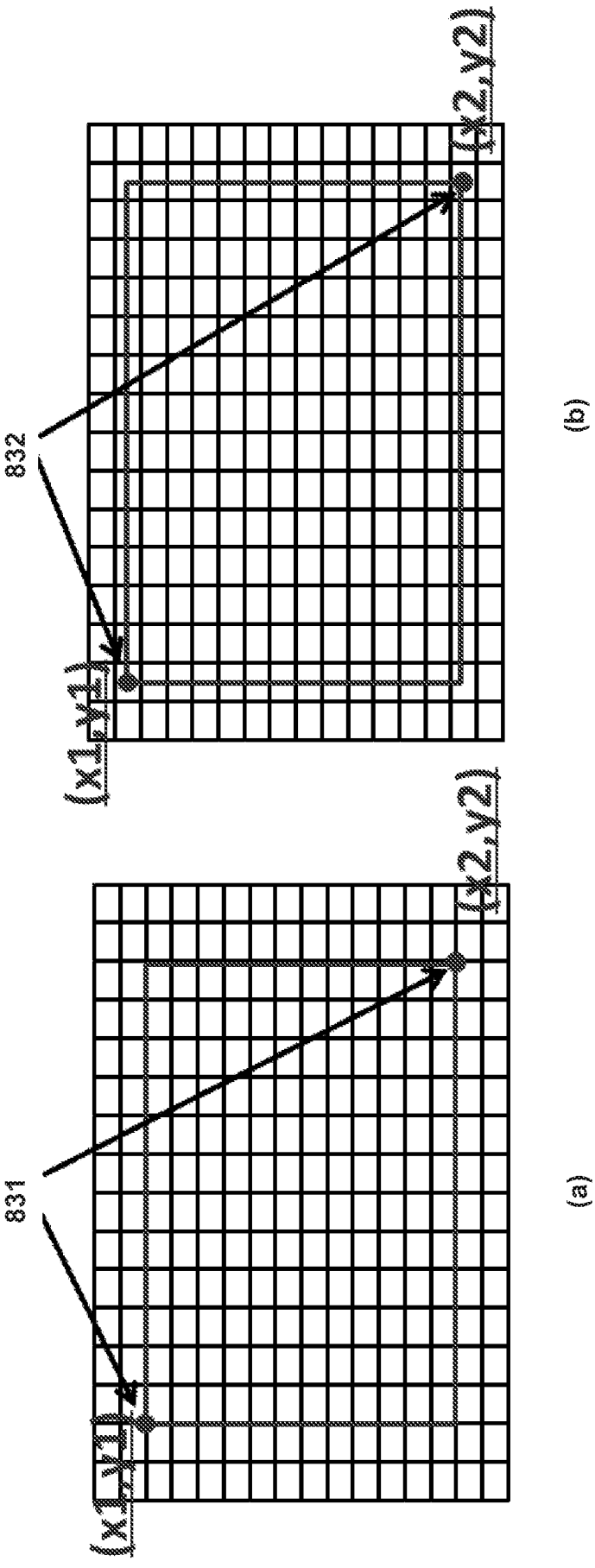


FIG. 8

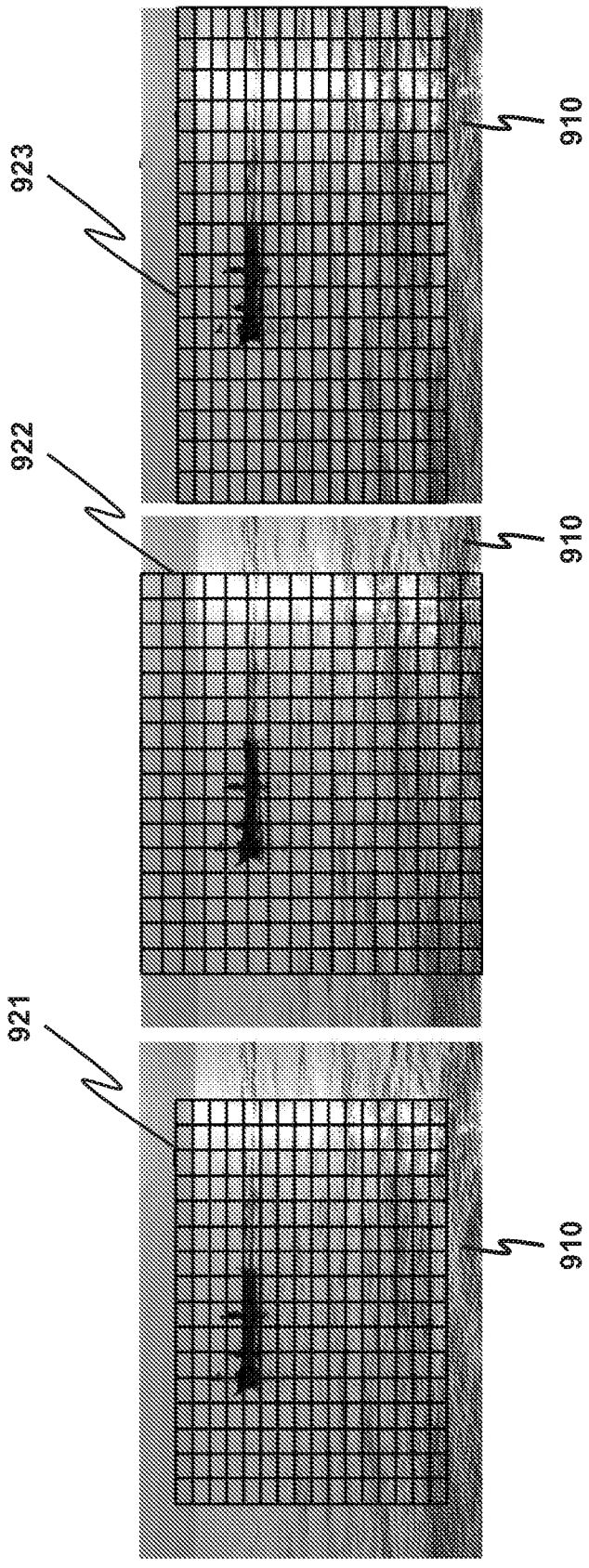


FIG. 9

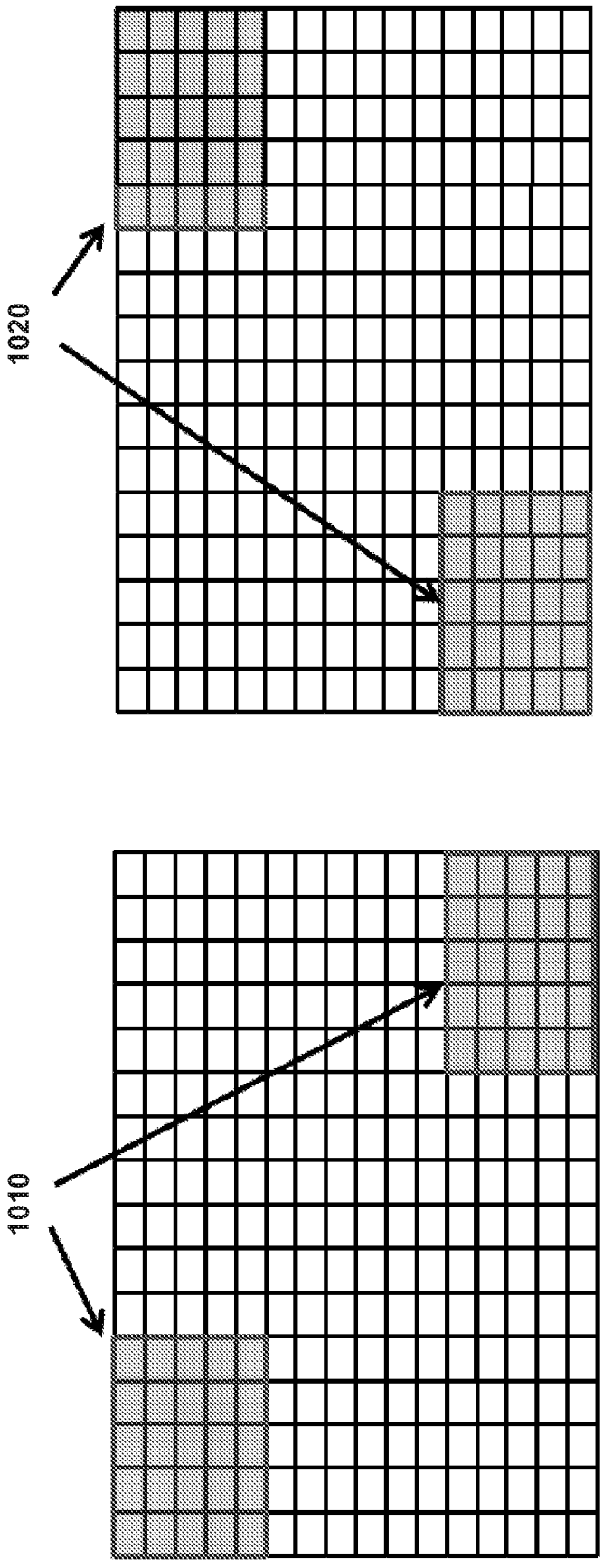


FIG. 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/084959

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/78(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS,SIPOABS,CNKI, IEEE, CNTXT, DWPI, PATENTICS: crop+, image, select, corner, region, grid, pair?, candidates, size, similar+, second, first, rectangular

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2016207875 A1 (PHOTOMYNE LTD) 29 December 2016 (2016-12-29) the whole document	1-19
A	CN 102736837 A (EWAUTO BEIJING VIDEO TECHNOLOGY CO LTD) 17 October 2012 (2012-10-17) the whole document	1-19
A	CN 101167363 A (EUCLID DISCOVERIES LLC) 23 April 2008 (2008-04-23) the whole document	1-19

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 July 2019	**09 August 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**ZHANG,Guihua**
Facsimile No. **(86-10)62019451**	Telephone No. **86-(10)-62411839**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/084959

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2016207875	A1	29 December 2016	US	10198629	B2	05 February 2019
				US	2017083762	A1	23 March 2017
				US	2018211107	A1	26 July 2018
				US	2019171876	A1	06 June 2019
				US	2017357851	A1	14 December 2017
				US	9754163	B2	05 September 2017
				US	9928418	B2	27 March 2018
CN	102736837	A	17 October 2012		None		
CN	101167363	A	23 April 2008	EP	1878256	A1	16 January 2008
				KR	101216161	B1	27 December 2012
				KR	20080002915	A	04 January 2008
				CA	2590869	A1	05 October 2006
				EP	1878256	A4	08 June 2011
				JP	2010259087	A	11 November 2010
				WO	2006105470	A1	05 October 2006
				AU	2006230545	B2	28 October 2010
				JP	5065451	B2	31 October 2012
				CA	2590869	C	21 January 2014
				JP	4573895	B2	04 November 2010
				AU	2006230545	A1	05 October 2006
				JP	2008537391	A	11 September 2008
				CN	101167363	B	07 July 2010