Learning an Aesthetic Photo Cropping Cascade

Peng Wang[1] Zhe Lin[2] Radomír Měch[2]
[1]University of California, Los Angeles [2]Adobe Research

Abstract *Cropping is one of the most fundamental and common operations in image processing for improving the aesthetic quality of photographs. Instead of manually designing rules for cropping, in this paper, we propose a generative model that learns an aesthetic photo cropping cascade from a large database of well-composed images and a dataset containing images with crops generated by expert photographers. Specifically, this model includes cropping priori, intuitive likelihood, compositional likelihood and change likelihood. Our learning exploits a spatial pyramid saliency feature and a multi-level foreground segmentation. The inference is done by efficient subwindow search (ESS) [10] which is benefited from the bound at conditional distribution in the cascade. Additionally, for extracting attentional subjects and capturing scene composition, we design an iterative saliency method to model the saliency moving paths, which is beyond the typical saliency model predicting a single attentional region. Experiments show that our approach outperforms the state-of-the-art cropping methods by a large margin.*

1. Introduction

Cropping is the most common task in image editing for improving the aesthetic quality of a photo. In photography, aesthetic or composition rules essentially are defined for two specific purposes [9]. First, to improve the representation quality of the subjects, e.g. emphasizing the main subjects and display them in a better position and layout. Second, to remove undesired regions such as unrelated or distracting objects, as exemplified in Fig.1. In this work, we are interested in building an efficient computational model for photo cropping which achieves these goals automatically.

Automatic aesthetic image cropping system could help photographers and editors in improving efficiency in cropping tasks and provide professional cropping advice to novice photographers [19]. Additionally, an automatic shooting recommendation app could be developed for smart phones and digital cameras [28], etc.

Nevertheless, at present, solving the problem is quite challenging for several reasons. (1) It is not easy to accurately identify the subjects of human interest due to the variation in the light, object shape and position. (2) Though some of the aesthetics rules are presented in the photography books such as the rule of thirds, visual balance [9], they can not be encoded as a fixed formulation because of personal styles. (3) Collecting a large-scale, high-quality human annotations for learning this rules is very time costly due to the labelling difficulties. (4) Evaluating each enumerated cropping box is time consuming. Luckily, the research on image saliency has precipitated a dramatic progress in recent years [3, 21, 20], which helps a lot to handle scene variance to build robust cropping systems [11, 17, 27]. In this paper, we provides a strong improvement over previous strategies on handling these difficulties.

Figure 1. Two cropping examples for aesthetic composition.

1.1. Related works

Recently, with the popularity of photo editing for web images, various aesthetic cropping techniques have been proposed. In general, the first category of approaches is rule-based, focusing on formulating the general rules into an explicit formation. Zhang et.al [30] applied face cues for detect regions of interest, and an image is cropped based on the face alignment. Santella et.al [19] proposed a semi-automatic algorithm which simplified the main subjects detection by recording human eye movement. Then the rule of thirds and visual balance cues are formulated by the Euclidean distance. Luo et.al [14] proposed a subject belief map and generated crops by maximize the subject content inside with manually designed compositional constraints. Later, Liu et.al [11] introduced an automatic system by utilizing saliency strategies [7] and an energy function with three key factors influencing the aesthetic quality. However,

hand designed formulations lead to limited generalization power of these algorithms since rules may change with respect to the main subjects.

Later, for handling such problems, many learning-based methods have emerged. Nishiyama et.al [16] proposed to build a qualitative classifier to separate the generated cropping candidates with the features from saliency, edge and color features for learning the rules. The cropping candidates are then sorted with classification scores. However, the ranking of compositional cropping is not linearly related to the classification score. Park et.al [17] proposed to apply a Gaussian mixture model on a large database to generate intrinsic rules for composition. Then the best crops are chosen by maximizing the posterior probability. Similarly, Ni et al. [15] and Zhang et.al [29] proposed to learn a probability mixture model with image regional relationship features from a data set of online user-favored images, and use sampling for generating cropping candidates. Recently, Yan et.al [27] mentioned that it is critical to consider the changes from the original image as local measurements would fall into local optima. They collected a dataset with cropping candidates from experts, and trained a local regional exclusive model and a global change-based model to verify the cropping quality. However, due to labelling difficulties, they collected a relatively small training set which is limited for learning a robust model. Moreover, brute force sliding window with a single pre-pruning step makes it computationally expensive especially for high resolution aesthetic images.

In this paper, different from [27], we developed a generative model-based aesthetic photo cropping system yielding the state-of-the-art results. Additionally, rather than just learning from human-labelled datasets, our system also exploits sparse auto-encoder [6] to discover the composition basis from the large amount of well-composed images, thus making it more robust than previous algorithms. More importantly, we organize the learning and inference in a cascade structure [23] for efficient optimization. We show that the cascade actually gives a weak upper-bound for brand-and-bound search, which could further accelerate the algorithm through efficient subwindow search (ESS) [10]. We call this Cascade ESS that potentially could be generalized to many optimization problems. Meanwhile, we notice that humans are detecting the main subjects in an image with a certain visual path [21], which indicates that the importance of subjects forms a sequence. Particularly, we design an iterative saliency estimation algorithm for capturing this information for composition. Last, we adopt multi-level segmentation for describing multiple subjects and propose several essential features for robust cropping results.

1.2. The framework

Fig. 2 shows an overview of our system which includes learning and inference stages. For learning, we firstly prepare two sets of image data , i.e., a large set of well-composed images $\mathcal{S}_1$ and a relatively small set of images cropped by experts $\mathcal{S}_2$. Then, a saliency map for each image is computed through our iterative strategy. After that, we extract the spatial pyramid saliency (SPS) feature which captures structural information of a saliency map. We first learn several key priori from data statistics. With the large image set $\mathcal{S}_1$, we construct a two-layer unsupervised learning structure. In the first layer, a sparse auto-encoder is adopted to learn a compact and effective rules basis, then a Gaussian Mixture Model (GMM) is stacked up to generate higher level rules for composition. At last in training stage, we learn changes between the original and cropped images with our proposed multi-level foreground and multiple attentional regions in each level of the foreground.

In the inference stage, by formulating it as a generative model, we perform an efficient search from a specific distribution depending on the image. As the best cropping box should achieve high posterior from the priori knowledge, compositional rules and cropping changes, this inference procedure could be performed in a cascaded manner. Specifically, we fit ESS in the learnt cascade. At each stage, the cascade gives a tighter upper-bound for ESS, resulting in smaller search space. Finally, we return several cropping boxes with the highest probability to the user.

2. Our approach

In this section, we introduce our model for learning and the attentional models for computing saliency maps and foregrounds will be presented in Sec.3. Formally, given an image $\mathbf{I}$, and its corresponding cropping cues, we wish to get a small set of cropping boxes $\mathcal{S}_{\mathbf{B}} = \{\mathbf{B}_i | \mathbf{B}_i = \{l, t, h, w\}, i = 1 \cdots m\}$ in which each of them crops the image into a new image $\mathbf{I}_b$ that has better visual quality than the original image. l, t, h, w denotes the left, top, height and with of the box. Specifically, the goal is to find the boxes with the highest probability given an aesthetic cropping model,

$$\begin{aligned}\mathbf{B}^* &= \arg\max_{\mathbf{B}} P(\mathbf{B}|\mathbf{I}, \theta) \propto \arg\max_{\mathbf{B}} P(\mathbf{I}|\mathbf{B}, \theta) P(\mathbf{B}|\theta) \\ &\propto P(\mathcal{C}|\mathbf{B}, \theta) P(\mathbf{B}|\theta) = \prod_{i=1}^{s} P(c_i|\mathbf{B}, \theta) P(\mathbf{B}|\theta) \qquad (1)\end{aligned}$$

where $\mathcal{C} = \{c_i\}_{i=1}^{s}$ is a set of independent cues that give the likelihood for a cropping box. $P(\mathbf{B}|\theta)$ performs as a priori knowledge from the dataset. Our model has three type of cues. $c_{in.}$ is a set of intuitive cues that is composed of cropping properties such as avoiding cutting through objects, etc. $c_{com.}$ is measuring the composition of the cropped image from our learnt compositional rules. $c_{ch.}$ is the change-based cues which measures the changes from $\mathbf{I}$ to $\mathbf{I}_b$. θ can be regarded as the model parameters we wish to learn. In our model, the complexity of computing each cues are different. Later, we would make benefits from such property

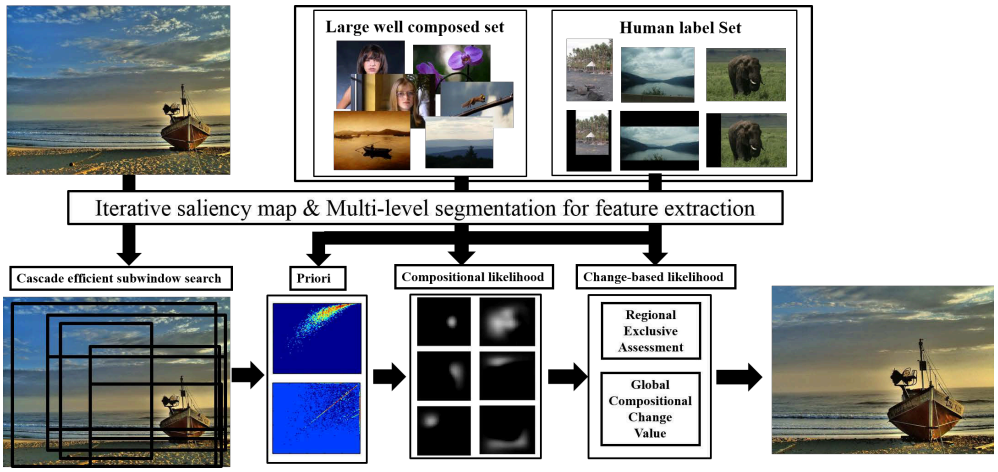


Figure 2. The overview of our aesthetic photo cropping cascade.

to makes the inference efficient.

2.1. Learning the cascade for image composition

In this section, we will first introduce the modelling and learning of the cues.

Learn the priori. As in Eqn.(1), we can first generate cropping boxes with the priori knowledge. We construct our prior distribution $P(\mathbf{B}|\theta) = P(\mathbf{B}|\mathcal{S}_2)$, where $\mathcal{S}_2$ is the given human labelled datasets. One can directly learn this distribution by factorizing it into $P(l,t,h,w|\mathcal{S}_2) = P(l|h,\mathcal{S}_2)P(t|w,\mathcal{S}_2)P(h,w|\mathcal{S}_2)$ [18]. To make the priori more informative, we define the probability into $P(\gamma,a,f|\mathcal{S}_2)$, where $\gamma = h'/w'$ is the aspect ratio, where $h' = h/\hat{h}$ is the portion of the cropping box height to the image height, and it is the same for $w' = w/\hat{w}$. $a = h' \times w'$ is the proportion of cropping box area to the image area. $f = \text{Fg}(\mathbf{B})$ indicates the portion of included foreground region (Sec. 3) inside the cropping box, which could be computed efficiently using integral image. For multi-level foreground, we use the average for this foreground priori. Then we have the probability defined as $P(\gamma|\mathcal{S}_2)P(a,f|\mathcal{S}_2)$ in order to apply these knowledge. The distributions from our database are shown in Fig. 3, where we show $P(h,w|\mathcal{S}_2)$ for indicating $P(\gamma|\mathcal{S}_2)$. We further add a Gaussian Kernel to smooth the probability density for better generality.

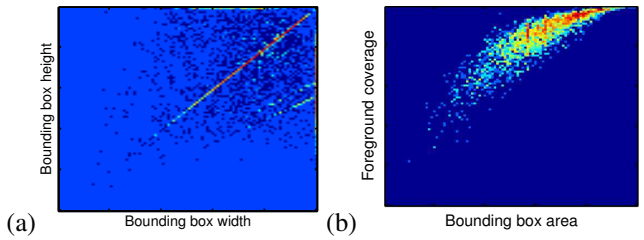


Figure 3. Learnt priori distributions of crops properties. (a) Normalized Width w.r.t Height (Normalized by respective image width and height). The diagonal line indicates aspect ratio of crops by experts is close to that of the original image. (b) Bounding box area w.r.t Foreground coverage.

Learn the intuition. Our intuitive cues in $P(c_{in.}|\mathbf{B})$ are composed of three aspects, 1. Human avoidance. Human is the major object class in natural images. We embed the well performed human head [26] and body detector [4] which are fast computing. Formally, we define $P(c_{\text{Hum.}}|\mathbf{B}) = f(O(\mathbf{B},\mathbf{H})/\text{Area}(\mathbf{H}))$, where $\mathbf{H}$ is a detected human bounding box. O indicates the overlapping area, $\text{Area}(\mathbf{H}))$ is the area of $\mathbf{H}$ and f is a distribution function that indicates the probability density given the overlapping score $O(\mathbf{B},\mathbf{H})/\text{Area}(\mathbf{H})$. We find that 98% ground truth crops have the overlapping score larger than 0.9, or smaller than 0.1, which indicates a very good property for us to use.

2. Image symmetry. For an image with strong symmetry, our experts keep the symmetry as shown in the right (a) of Fig. 4. Due to limited number of training images, we code it as intuitive knowledge through symmetry detection. Specifically, our algorithm first proposes several symmetry axis candidates $\{a_i\}_{i=1}^{N_a}$ evenly around the center axis. Each axis separates the image into two parts, and the SIFT [13] descriptors on each parts are computed. We match the two sets of SIFT and find the match points $\mathbf{p}_1$ and $\mathbf{p}_2$ that are located in a similar position in the two parts, i.e. $\|\mathbf{p}_1 - \mathbf{p}_2\| \leq Th_d$. A symmetry axis is found once the match is significant. At last, our symmetry probability is defined on image symmetry properties,

$$P(c_{Sym.}|\mathbf{B},\mathbf{I}) = \begin{cases} 1 & if\ \ \mathbb{1}(Sym(\mathbf{B},a)) \,\&\, \mathbb{1}(Sym(\mathbf{I},a)) \\ 0 & if\ \ \neg\mathbb{1}(Sym(\mathbf{B},a)) \,\&\, \mathbb{1}(Sym(\mathbf{I},a)) \\ 0.5 & if\ \ \mathbb{1}(Sym(\mathbf{I},a)) \end{cases}$$

where $\mathbb{1}(Sym(\mathbf{I},a))$ is an indicator function which indicates the image $\mathbf{I}$ is symmetric with respect to the axis a. Then our intuition cues probability are a joint distribution of these cues, i.e. $P(c_{in.}|\mathbf{B}) = P(c_{Hum.}|\mathbf{B})P(c_{Sym.}|\mathbf{B})$.

Learn the composition. Our compositional model is a mixture of probability density built based on the well-composed image set $\mathcal{S}_1$ with the spatial pyramid saliency

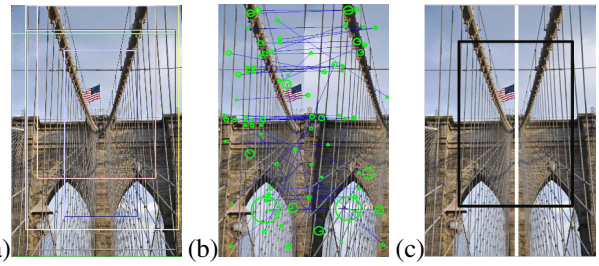


Figure 4. Symmetry detection. (a) Original image with ground truth crops. (b) Spatial matching. (c) Our detected symmetry axis and a preferred cropping box.

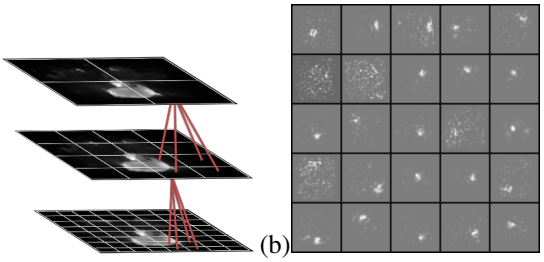


Figure 5. (a) Our spacial pyramid saliency (SPS) feature. (b) The basis for composition learnt from our well-composed database with 32×32 resolution.

(SPS) feature. In particular, as shown in Fig. 5(a), the SPS is computed by partitioning the saliency map (in Sec. 3) into a hierarchy of regular grids i.e. $\{p_i\}_{i=1}^{m}$, and we concatenate the feature from m levels into a final feature vector, i.e. $\mathbf{f}_{sps} = [\mathbf{f}_{p_1}^T..\mathbf{f}_{p_m}^T]^T$. By separating saliency in a hierarchical way, our feature captures the spatial attentional information for learning. In addition, through bottom-up aggregation, the computation is very fast by using integral image. In our experiments, the pyramid of the feature are $\{4\times4,\ 8\times8,\ 16\times16,\ 32\times32\}$, resulting in a feature with the dimension of 1360.

However, directly using the high dimensional feature for one mixture model, e.g. GMM as previous work [16, 17], results in a high computational cost and a model with limited generalization power. Instead of learning the model in a flat way, we build a two layer model to encode the feature into a compact form. In the first layer, the sparse auto-encoder [6] is adopted to reduce the dimensionality and linearise the original feature vectors. It learns a compact and regularized basis of composition with good generalization power. Using such a method, our learnt basis intrinsically represent the rules from the well composed images such as the rule of the thirds and object balance, and a well composed image should be a sparse combination of these rule basis. We show the learnt basis in finest SPS level from the sparse auto-encoder in Fig. 5(b). After coding, the i_{th} dimension of new compact features are represented by $\mathbf{f}_c(i) = sigmoid(\mathbf{W}_i{}^T\mathbf{f}_{sps} + b_i)$. We set the number of basis to 25 in our experiments.

After the coding, in the second layer, we further built a GMM model M_c on top of the features learnt from the sparse auto-encoder for capturing higher level compositional rules and estimating a probability density with a likelihood as $p(M_c|\mathbf{f}_{c,\mathbf{B}})$. This can be applied to measure how well each input image fits to our learnt compositional rules, where $\mathbf{f}_{c,\mathbf{B}}$ is the coded feature inside cropping box $\mathbf{B}$. Then, our compositional likelihood can directly take the output of the distribution as, $P(c_{com.}|\mathbf{B}, \theta_{com.}) = p(M_c|\mathbf{f}_{c,\mathbf{B}}))$, where $\theta_{com.}$ is the learnt parameters from the sparse auto-encoder and the GMM model.

Learn the changes from the original image. For learning the changing cues, we refer to Yan et.al [27], while enhancing the approach over both exclusive and compositional change cues. In [27], to extract exclusive cues, it finds a single foreground ($\mathcal{F}$) and background ($\mathcal{B}$). Then, features for each segmented region is extracted with respect to the $\mathcal{F}$ and $\mathcal{B}$. Formally, the feature of a region can be written as $\mathbf{f}_{\mathcal{F},\mathcal{B}}^r = [D(r,\mathcal{F}), D(r,\mathcal{B}), \mathbf{f}_r^T]^T$, where $D(r,\mathcal{F})$ is the minimum geodesic distance from the region r to $\mathcal{F}$, and $\mathbf{f}_r$ is independent regional features. After that, a regression is learnt measuring how much a region should be excluded from a cropping box. A good cropping box should include the regions with low elusive score. To get the compositional change, several designed compositional cues, such as the horizon line, the foreground position etc., are extracted within each cropping box, and a regression is learnt for predicting composition changes from the original images.

However, the assumption of single foreground is limited in practice. In our paper, we consider the general cases with multi-foreground hypothesis. For exclusive cues, we first decompose the image into superpixels [25, 12], i.e. $\{r_i\}_{i=1}^{i=N_R}$. We compared the two superpixels, resulting in [25] gives better results and [12] gives better speed. Then, multi-level foreground segmentation (Sec. 3) are generated by utilizing our saliency map (Sec. 3), resulting in L level foreground segments. We compute a feature vector for each superpixel by concatenating features from the multi-level segmentation. Formally, the feature of a superpixel can be written as $\mathbf{f}_r = [\mathbf{f}_{r1}^T, \cdots, \mathbf{f}_{rL}^T]^T$, where $\mathbf{f}_{rl}$ is the feature respecting the foreground and background regions in level l. For foreground region in each level, we detect K_l connected components as foreground segmentation, i.e. $\{\mathcal{F}_{il}\}_{i=1}^{K_l}$.For the background, we find K_b background regions $\{\mathcal{B}_{il}\}_{i=1}^{K_b}$ in the image through building a kernel support vector machine (KSVM) between the regions outside and inside of human cropping over image sets $\mathcal{S}_2$. We use three type of features for distinguish, i.e. $[\text{Pos}(r), \text{Bd}(r), \text{Sal}(r)]$, where the $\text{Pos}(r)$ is the center position of the region, $\text{Sal}(r)$ is the average saliency within the region. $\text{Bd}(r) = \text{BourndryLen}(r)/\text{Len}(r)$, where $\text{BourndryLen}(r)$ is the length of region contour that is connect with image border and $\text{Len}(r)$ is contour length of the

region. We keep the top three if there is a lot of predictions, while taking the region with the highest score if there is no prediction.

With the foreground and background regions at hand, the feature for each superpixel is computed as:

$$\mathbf{f}_{rl} = \begin{cases} [\min_i D(r, \mathcal{F}_{il}),\ \min_j D(r, \mathcal{B}_{jl})),\ \mathbf{f}_r^T]^T \\ \qquad if\ \ r \notin \mathcal{F}_{il}, \forall i = 1 \cdots K_l \\ [0,\quad \min_j D(\mathcal{F}_{il}, \mathcal{B}_{jl}),\ \mathbf{f}_{\mathcal{F}_{il}}^T]^T \\ \qquad Otherwise \end{cases}$$

This formula indicates that a region inside a foreground will inherit the features of the foreground and the ones outside will compute feature based on foreground and background in this level, which makes the segments distinguishable in multiple levels. Our approach both discover multiple foregrounds and hierarchically describe their attentional importance, which produces more robust and effective representations for learning the model. This is especially useful in multi-attentional images. Finally, we adopt the SVR to train a model from this regional features.

For compositional change cues, we replace the single foreground distance feature in [27] with a concatenation of multiple foregrounds and multiple segmentation levels. Additionally, we enhance the compositional change measurements by introducing rules for multi-attentional model: (1) visual balance, measured by the distance between foreground region centroid to image center. (2) diagonal dominance, measured by the spanned range, average center position and average orientation of diagonal lines in the image. (3) background balance. measured by the portion of the background type number in the cropped image to the type number in the original image. This rule indicates the multi-type preference of an image, i.e. the cropped image should not just contain a single background type. We provide the details of background type type in supplementary material. Comparing to the rule learnt with GMM, these rules contain more fine-grained properties that give richer descriptions in avoiding cropping boxes falling in local optima. At last, we train a SVR model by using these features to measure the crop changes. With the regressed score, we compute the energy potential $E_{exl.}(\mathbf{B})$ and $E_{co.}(\mathbf{B})$ in the same way as [27].

Finally, our change-based likelihood distribution is defined as $P(c_{ch.}|\mathbf{B}, \theta_{ch.}) = \frac{1}{Z} exp(-(E_{exl.}(\mathbf{B}) + \mu E_{co.}(\mathbf{B}))$, where Z is a normalizing constant, $\theta_{ch.}$ is the regression parameter.

2.2. Efficient Inference by Cascade ESS

Given the learnt model, thanks to the cascade property of this generative model, we designed a very efficient inference scheme. Specifically, as shown in Eqn.(1), our model can be factorized as $\prod_{i=1}^{s} P(c_i|\mathbf{B}, \theta) P(\mathbf{B}|\theta)$, with different computation complexity ascending in a sequence as $P(\mathbf{B}), P(c_{in.}), P(c_{com.}), P(c_{exl.}), P(c_{co.})$. Here we omit the writing of $\mathbf{B}, \theta$ for simplification. We additionally factorize $P(c_{ch.})$ into $P(c_{exl.})$, and $P(c_{co.})$. Naively, one could directly perform cascade filtering by accumulating the distribution at each stage and filter bad results through a threshold over the distribution. However, it requires tedious cross validation for good thresholds.

We notice that due to the probability factorization in a discrete parameter space, i.e. $[t, l, h, w]$ for sliding window, the probability in the previous stage naturally serves as an upper bound for the latter stage. Formally, by taking two concatenated stages from this cascade, we have $P(B) \geq P(A|B)P(B)$. This property allows us using ESS [10] for optimal searching. Specifically, in ESS, a rectangle set is represented with an interval set $S = \{\{l, t, h, w\}, l \in L, t \in T, h \in H, w \in W\}$, by separating one of the intervals, e.g. let $T = T_1 \bigcup T_2$, we can split the rectangle set into two disjoint sets. This serves as a set branching process. For each set, by a fast computing upper bound UB and a generated lower bound LB, we could decide whether to keep branching and search in this set or not. Now let us take the first two stages cascade for example, i.e. $P(A|B)P(B)$, and the rest of the cascade can be performed so forth. We start from the whole rectangle set $S_0 = \{\{l, t, h, w\}, L = [1, \hat{w}], T = [1, \hat{h}], H = [1, \hat{h} - t], W = [l, \hat{w} - l]\}$, and branch by a reasonable split as in ESS [10], yielding $S_0 = \{S_{11}, S_{12}\}$. Then for upper bounds, we use $P(B)$ to quickly find the upper bound for all the split sets, i.e. $\{UB_{11}, UB_{12}\}$. For lower bound, we select top 50 rectangles with highest $P(B)$, then compute $P(A|B)P(B)$ and take the median as a reasonable overall lower bound LB, which gives us a pruning threshold. With this criterion, we can keep branching from the set having largest upper bound, and pruning the sets with upper bound UB smaller than LB. At the end, we will results in sets of crops with their $P(B)$ larger than the LB. After the first stage, we reunion the remaining crops in to the initial sets S_1 for the next stage which maximizes $P(C|B)P(A|B)P(B)$. In this time, we compute $P(A|B)P(B)$ for S_1 to find sets' upper bound and the similar procedure could be performed iteratively until the last stage. During this cascade, the resulting set will become smaller and smaller, and the upper bound will become tighter and tighter. Notice at the last stage, the LB should be updated when we get the full probability of the model, which ensure us to find the best results. Though this cascade ESS does not ensure global optimal, in the experiments, we obtain good results efficiently due to that good cropping should achieve high probability in all stages.

In practice, we stop branching when all the rectangle set intervals are small (< 50 pixel in our experiments for 1000×1500 images). This is like a non-maximum suppression to enforce diversity. Due to that our probability distributions are generally not fuzzy and non-uniform, e.g.

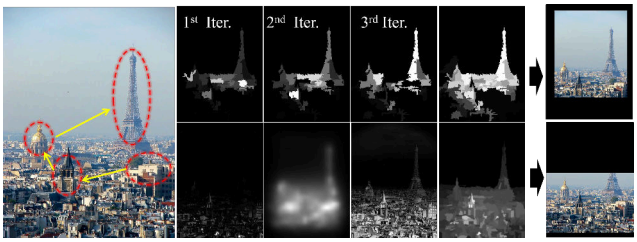


Figure 6. We show one human attentional path in the left image. In traditional methods, (in the bottom line, from left to right, FT[1], GBVS[5], HC[3], RC[3]), one of the main subjects is suppressed by the texture dense and highlighted parts from bottom of the image. In the top line, we show the three maps generated from each iteration and our final saliency map which includes all the main subjects.

Fig. 4, we can achieve very fast performance by pruning a large amount of cropping boxes at the early stages. Additionally, when the type of distribution is parametrized with a known form as the GMM or is easily to estimate, we divide the distribution $P(A|B)$ by its maximum value, which provides a better bound for pruning without loss of generality.

3. Attentional models

Saliency estimation and foreground detection are the keys to make cropping effective. In this section, we introduce our robust saliency estimation and multi-level segmentation strategy.

Saliency estimation for composition. As in Fig. 6, recent methods [1, 5, 3] are focusing on distinguishing the most salient regions. However, attentional model for composition is more complex than detecting the most saliency region, and less attentional parts play important roles, such as balancing the main subject etc. Inspired by [21], we know human attention forms a saccadic path from the most attractive subject to the parts less important.However, different from the model of [21], we are more interested in fast region-wise saliency for objects, which leads us to an iterative saliency estimation strategy that mimics such a process.

Specifically, we take our superpixel segmentation, $\{r_i\}_{i=1}^{i=N_R}$, based on which our algorithm updates the saliency map iteratively. In each iteration, we compute the uniqueness of each region by combining multiple regional contrast cues and generate a saliency map $\mathbf{S}$. If there is a foreground region, we record it and remove this region in the next iteration when computing the regional uniqueness. The saliency of a region in our case is computed by aggregating several attentional cues inspired from psychology analysis of human attention including: (1) a local and global region contrast cues, which are complementary as indicated in [2]. (2) a soft center bias, which is observed in human tendency [22]. (3) the background cues in connecting the boundary. The final formulation for calculating the saliency for a region is,

$$\mathbf{S}(r_k) = -w_b(r_k)w_c(r_k)\sum_{r_i \neq r_k} w_s(r_i, r_k)w_a(r_i)D_c(r_i, r_k),$$

where $w_b(r_i) = exp\{-\text{Bd}(r_i)/\delta_b^2\}$ measures the region connectivity to the boundary and $\text{Bd}(r_i)$ is defined as a feature in Sec. 2.1 for learning the background, we set $\delta_b = 0.2$. $w_c(r_i) = exp\{-c(dx_i^2/w^2 + dy_i^2/h^2)\}$ is the center bias for each region and dx_i and dy_i is the Euclidean distance from the region center to the image center. We set $c = 2$. $w_s(r_i, r_k)$ is the spatial similarity between r_i and r_k which is defined as,

$$w_s(r_i, r_k) = \begin{cases} exp\{-d_s(r_k, r_i)/\delta^2\} & if \;\; adj(r_k, r_i) = 0 \\ 1 & otherwise \end{cases}$$

where $d_s(r_k, r_i)$ is the Euclidean distance between the center position of regions k, i, and $adj(r_k, r_i)$ is an indicator function judge whether the two regions are adjacent or not. This enhances the contrast between the region and its neighbourhoods. We set $\delta = 3$. $w_a(r_i)$ is the area portion of r_i to the image, which emphasizes the contrast to larger regions. $D_c(r_i, r_k)$ measures the appearance distance between the two regions, which is adopted from [8]. Finally, our output map $\mathbf{S}$ is normalized into the range of $[0, 1]$.

The algorithm stops when no salient region can be detected, or it reaches the maximum iteration number $N_{Iter} = 4$. To judge whether a salient region is included or not, we train a kernel SVM by using the SPS saliency feature between images with and without attentional regions similar with [24].

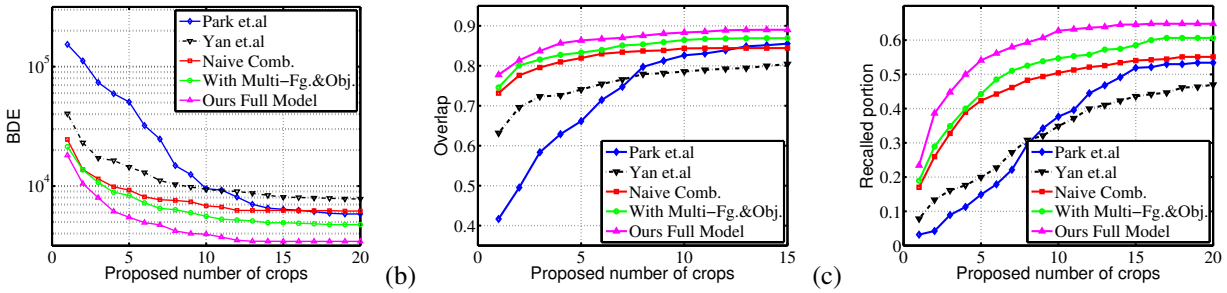
Intuitively, the content found later is less important and it might already weighted in previous iterations. Thus, we take a mono-decreasing weight, i.e. $\mathbf{S}_{Iter} = \sum_{t=1}^{T} w_t \mathbf{S}_t$, where $\mathbf{w} = [w_1, \cdots, w_T]^T$ and we set the weight as a linear interception in the range of $[0.6, 1]$. Last, our final model also combines the responses from human detector [4] and face detector [26]. We set the combination weight as $\{1, 1\}$.

Foreground segmentation. With the saliency estimation at hand, we generate our multiple foreground segmentation by proposing multiple thresholds respect to foreground area since a good foreground for evaluation should contain a certain content but not cover the whole image.

we formulate such limitation by a minimum foreground area a_{min} and a maximum foreground area a_{max}. We find a minimum threshold Th_1 which generates foreground with its area smaller than a_{max}, and corresponding a maximum threshold Th_{N_T} for a_{min}. The rest of the thresholds are linearly interpolated within $[Th_1, Th_{N_T}]$, resulting in a set of thresholds $\{Th_1, \cdots, Th_{N_T}\}$. We set $a_{min} = 0.01$ and $a_{max} = 0.6$.

4. Experiments

We evaluate our approach on the database from Yan et.al [27] and our self-built database as images from [27] mostly contain a single object. The details of data collection



(a) (b) (c)

Figure 7. The quantitative comparison between our method and other state-of-the-art methods: Park et.al [17] and Yan et.al [27]. (a) BDE. (b) Overlap. (c) Recall given 0.75 overlap in Eqn. 2.

is included in Sec.1 in our supplementary material. We will release the dataset with the publication of this paper. We compare our results with two state-of-the-art learning strategies: Yan et.al [27] with supervised change-based model and Park et.al [17] with unsupervised learnt compositional model. We implement the other two algorithms with our best efforts due to lack of author's code, including parameter tuning and retraining over our collected images.

Quantitative results We first compare the results based on the setting and criteria from [27] to proof our implementation. In Tab 1, our implementation of [27] performs almost as good as that in the paper. We can see our algorithm gives a higher accuracy over all the three testing sets as we handle better on multiple objects.

In addition, as users may interested in selecting from multiple cropping results. We propose three standard evaluation strategies for measuring the performance of multiple proposed cropping results, i.e. minimum bounding box displacement error (BDE), maximum overlap and ground truth recall rate with respect to top m proposed results. For a fair comparison, we let all the compared algorithms propose top m results for evaluation. Formally, given the generated results $\{\mathbf{B}_i\}_{i=1}^{m}$ and NG human cropped results $\{\mathbf{G}_j\}_{j=1}^{NG}$ the criteria is defined as:

$$BDE = \min_{i,j} \|\mathbf{B}_i - \mathbf{G}_j\|;\ Overlap = \max_{i,j} \text{Olp}(\mathbf{B}_i, \mathbf{G}_j);$$
$$Recall = \frac{1}{\text{NG}} \sum_j \mathbb{1}(\exists i,\ s.t.\ \text{Olp}(\mathbf{B}_i, \mathbf{G}_j) > 0.75), \quad (2)$$

where $\text{Olp}(\mathbf{B}_i, \mathbf{G}_j) = \mathbf{B}_i \cap \mathbf{G}_j / \mathbf{B}_i \cup \mathbf{G}_j$. For evaluation, we randomly and evenly separate the images into 5 folders and report the average results by 5 fold validation. The comparison results are shown in Fig. 7, from which we can see that in all criteria, our method provides the best results especially at the top few ranks of output cropping boxes. For the red curve, we combine the priori, unsupervised learnt rules and compositional energy from Yan et.al [27] with saliency map [3] they adopted, which already improves results significantly. For the green curve, we plug in multiple foreground segmentations and features. We can see it performs better due to our more accurate foreground capturing strategy. For the pink curve, by adding our iterative saliency map, the results are further improved due to more in-detailed compositional information. This proves that all the cues we induced are complementary in photo cropping tasks.

More importantly, thanks to the cascade ESS procedure, our approach is 20X faster than that of Yan et.al [27]. This is because much fewer candidates are passed to the final evaluation layer which is computationally expensive. Specifically, for an image of 1000×1500, based on our unoptimized C++ implementation (at 3.4GHz PC), we can get our results around 1.5(s), while [27] takes around 25(s).

Testing set	Photographer 1	Photographer 2	Photographer 3
Park et.al [17]	0.6034 (0.1062)	0.5823 (0.1128)	0.6085 (0.1102)
Yan [27]	0.7275 (0.0692)	0.7199 (0.0732)	0.7245 (0.0743)
Ours	**0.7823** (**0.0623**)	**0.7697** (**0.0617**)	**0.7725** (**0.0701**)

Table 1. Performance comparison on database from [27]. First number is average overlap ratio. Second number (in parentheses) is average boundary displacement error. Best values are shown in boldface.

Qualitative results Fig. 8 shows the examples of our approach and the cropped results from the other two techniques. For a better illustration, we pick the best solution with smallest BDE from the returned top 5 results for all compared algorithms. From these examples, we can see that unsupervised learning such as [17] would results in crops that sometimes dramatically change original image and fall in undesired local optima due to the existence of multiple and false attentional subjects. Built on supervised data, Yan et.al [27] can only train the model from images with human labels, which yields less robust compositional models for selecting candidate regions, e.g. a main subject is often located in the center (e.g. third row at (b)). Fig.6 in their paper [27] also reveals this issue. In addition, due to single foreground assumption, the algorithm fails in handle multiple objects cases and sometimes prunes the main subjects (e.g. third row at (e)). Our approach handles these issues well by carefully integrating several complementary cues and learning strategies, which robustly generates multiple style cropping candidates for users, we show more comparison examples in the supplementary material.

References

[1] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

(a) (b) (c) (d) (e)

Figure 8. Qualitative comparisons. The first row is original images, the second row is the results from Park et.al [17], the third row is the results from Yan et.al [27]. We show our results in the last row.

[2] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012.

[3] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[4] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, pages 1–11, 2010.

[5] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[6] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[8] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng. Automatic salient object segmentation based on context and shape prior. In *BMVC*, pages 1–12, 2011.

[9] B. Krages. *Photography the Art of Composition*. All-worth Press, 2005.

[10] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12):2129–2142, 2009.

[11] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Comput. Graph. Forum*, 29(2):469–478, 2010.

[12] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *CVPR*, pages 2097–2104, 2011.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] J. Luo. Subject content-based intelligent cropping of digital photos. In *ICME*, pages 2218–2221, 2007.

[15] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian. Learning to photograph: A compositional perspective. *IEEE Transactions on Multimedia*, 15(5):1138–1151, 2013.

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

[17] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo re-arrangement. In *ICIP*, pages 2741–2744, 2012.

[18] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *ICCV*, pages 1052–1059, 2011.

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

[20] P. Siva, C. Russell, T. Xiang, and L. de Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, pages 3238–3245, 2013.

[21] X. Sun, H. Yao, and R. Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *CVPR*, pages 1552–1559, 2012.

[22] B. W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):1–17, Nov. 2007.

[23] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[24] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li. Salient object detection for searched web images via global saliency. In *CVPR*, pages 3194–3201, 2012.

[25] P. Wang, G. Zeng, R. Gan, J. Wang, and H. Zha. Structure-sensitive superpixels via geodesic distance. *International Journal of Computer Vision*, pages 1–21, 2013.

[26] R. Xiao, H. Zhu, H. Sun, and X. Tang. Dynamic cascades for face detection. In *ICCV*, pages 1–8, 2007.

[27] J. Yan and S. Lee. Learning the change for automatic image cropping. In *CVPR*, 2013.

[28] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, 96(3):353–383, 2012.

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

[30] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W.-Y. Ma. Auto cropping for digital photographs. In *ICME*, pages 438–441, 2005.