Vis Comput (2014) 30:443–453
DOI 10.1007/s00371-013-0867-4

ORIGINAL ARTICLE

SalientShape: group saliency in image collections

Ming-Ming Cheng · Niloy J. Mitra · Xiaolei Huang · Shi-Min Hu

Published online: 31 August 2013

Abstract Efficiently identifying salient objects in large image collections is essential for many applications including image retrieval, surveillance, image annotation, and object recognition. We propose a simple, fast, and effective algorithm for locating and segmenting salient objects by analysing image collections. As a key novelty, we introduce *group saliency* to achieve superior unsupervised salient object segmentation by extracting salient objects (in collections of pre-filtered images) that maximize between-image similarities and within-image distinctness. To evaluate our method, we construct a large benchmark dataset consisting of 15 K images across multiple categories with 6000+ pixel-accurate ground truth annotations for salient object regions where applicable. In all our tests, group saliency consistently outperforms state-of-the-art single-image saliency algorithms, resulting in both higher precision and better recall. Our algorithm successfully handles image collections, of an order larger than any existing benchmark datasets, consisting of diverse and heterogeneous images from various internet sources.

Keywords Saliency detection · Group saliency · Object of interest segmentation · Image retrieval

M.-M. Cheng (✉) · S.-M. Hu
TNList, Tsinghua University, Beijing, China
e-mail: cmm.thu@gmail.com

N.J. Mitra
University College London, London, UK

X. Huang
Lehigh University, Bethlehem, USA

1 Introduction

The ubiquity of acquisition devices, e.g., cameras and smartphones, and the growing popularity of social media have resulted in an explosion of digital images accessible in the form of personal and internet photo-collections. Typically, such image collections are huge in size, have heterogeneous content, and are noisy due to diverse background and illumination conditions. Although such images form a well-established communication medium for sharing experiences or blogging events, we still lack efficient and effective methods to analyze and organize such images.

Determining characteristic or salient regions of images allows transitioning from low-level pixels to more meaningful high-level regions, and thus form an essential step for many computer graphics and computer vision applications, including interactive image editing [14, 16, 41, 51, 53], image retrieval [12, 27, 28, 30], and internet visual media processing [11, 33, 34, 40]. Recently, significant success has been reported in saliency-based image segmentation producing near ground-truth performance on simple images ([1, 13, 15, 41] and references therein). The next challenge is to reliably segment salient object regions in large heterogeneous image collections such as internet images, e.g., Flickr, Picasa. Since such collections contain rich information about our surroundings, their effective analysis will naturally provide improved understanding of image contents.

We introduce *SalientShape*, a group-saliency based framework for salient object detection and segmentation in image collections. We demonstrate that even when the shared content across image collections is small, e.g., 30 %, our framework produces superior results as compared to individually processing the images. Our proposed method is simple, scales well with increasing size of collections, has low memory-footprint, and can be effectively integrated with existing image handling systems.

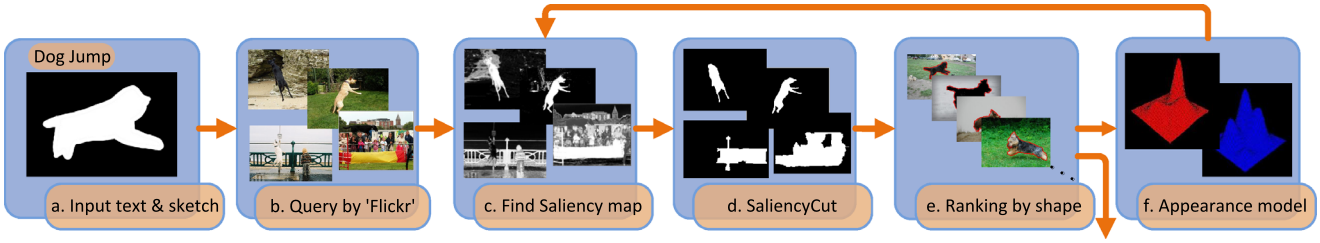


Fig. 1 System pipeline. Our system explicitly extracts salient object regions from a group of related images with heterogeneous quality offline (**a**–**d**, **f**) to enable efficient online (**e**) shape based query. To enable effective salient object segmentation for a large collection of images with heterogenous contents, our system only requires a simple input text keyword and a coarse sketch (**a**), for initial query of internet image candidates (**b**) and shape ranking (**e**). For a new query shape related to a processed keyword, the segmentation results (**d**) can be re-used for efficient query (only shape matching is required in this typical use case)

Our algorithm (see Fig. 1) runs in the following key stages: First, for a query object class, we retrieve candidate images by pre-filtering using keywords. Such retrieved images are usually noisy and contain outliers due to limitations in keyword-based image search, ambiguity of keywords, and heterogeneous tags. Next, we detect and segment salient object regions of each candidate image using SaliencyCut [15] and automatically remove candidates with fragmented scenes or with unreliable segmentation quality by jointly analyzing the salient regions. Then, in a key stage, we re-rank the remaining candidates based on the consistency between their saliency cuts and the user provided sketches. In this step, access to image collections proves critical since even in noisy ensembles we observe that segments corresponding to the inlier objects have *consistent* appearance and shape properties. To exploit this, we build global (group) foreground and background appearance models from the top-ranked candidate images for the query object class. Finally, we use the extracted appearance model for group saliency region detection and segmentation. We iterate the process to alternately improve saliency estimates and appearance models.

We compare the resultant segmentation results with state-of-the-art single image salient region segmentation methods [1, 15, 45] (see Fig. 7(b)), and to retrieval performance with SHoG [22] on 30 categories (see Table 1). Further, we introduce a benchmark dataset consisting of 15,000 images collected from Flickr along with 6000+ pixel-accurate ground truth salient object masks where applicable (to be made publicly available for academic use). To the best of our knowledge, our benchmark dataset with pixel accurate salient object region ground truth labeling is the largest of its kind (15× larger than [1]), while the images are more difficult and closer to real-world scenarios. In our extensive tests, group saliency consistently outperforms existing state-of-the-art alternatives, especially on images with cluttered backgrounds.

The improved performance is primarily due to the joint saliency estimation and (single-image and group/global) appearance models learning. Our system also benefits from meta-data,[1] visual saliency, and shape similarity to explicitly detect salient object regions and enable shape retrieval without influence from background clutter. In summary, we (i) introduce *group saliency* to extract object of interest regions from a group of correlated but heterogeneous images, and (ii) present a large benchmark dataset to objectively compare the superiority of group saliency over traditional single image saliency detection. Since our focus is on *consistent segmentation*, we show retrieval only as a potential application rather than being the focus of this work.

2 Related work

Salient region extraction Various methods have been proposed for extracting salient regions from single images: Ko and Nam [36] select salient regions using a support vector machine trained on image segment features, and then cluster these regions to extract salient objects. Han et al. [29] model color, texture, and edge features using a Markov random field framework and grow salient object regions from seed values in the saliency maps. Achanta et al. [1] average saliency values within image segments produced by mean-shift segmentation, and find salient objects by identifying image segments with average saliency above a threshold. Cheng et al. [13, 15] propose a saliency estimation method to automatically initialize an iterative version of GrabCut [46] to segment salient object regions. These methods aim at salient region extraction for *individual* images, while ignoring useful global information available from correlated image collections. Recently, co-saliency methods have been proposed to find common salient object(s) between pair of images [10, 39] or among multiple images [8, 52]. Such methods, however, require salient areas to contain parts of the foreground objects across most images. Further, the algorithms are difficult to scale to large number

[1] Meta-data is the current industry standard for image retrieval as popularized by search engines like Google image, Flickr, etc.

of images (largest demonstrated collection has 30 images). In contrast, we focus on detecting and segmenting correlated salient object regions from large (thousands or more) image collections with heterogeneous contents (e.g., internet images).

Internet image re-ranking Fergus et al. [25] use the top results returned from a web-based image search engine to train a classifier, and then use the classifier to filter the search results. Ben-Haim et al. [3] automatically segment images into regions and build color histogram features for each region, which are then clustered to obtain principal modes. The remaining images are then re-sorted based on the distance of their regions to the mean feature values of the principal clusters. Cui et al. [17] categorize a query image into one of several predefined categories, and employ a specific similarity measure in each category to combine image features for re-ranking based on the query image. Popescu et al. [44] re-rank images based on the visual coherence of queries using a diversification function to avoid near-duplicate images and ensure that different aspects of a query are presented to the user. None of these algorithms use visual attention and shape information of the desired object. In contrast, we use such information to capture potential appearances that a desired object class may have, enabling superior salient region extraction.

Sketch based image retrieval (SBIR) Early works by Hirata and Kato [31] perform retrieval by comparing shape similarity between user sketches and edge images in a database, expecting precise sketches from the users. Alberto and Pala [19] further employ elastic matching to a user-sketched template for robust retrieval, with the cost of expensive computation. Recently, Cao et al. developed a novel indexing technique [6] to support efficient contour-based matching for a retrieval system [7] that handles millions of images. However, the method does not provide translation, scale, or rotation invariance, and more importantly expects the desired object to appear at roughly similar positions, scale, and rotation as in the user-drawn query sketch.

In an important recent system, Eitz et al. [22] use local descriptors to achieve state-of-the-art retrieval performance. Their success is mainly attributed to translation invariance of local descriptors as well as using large local features (20 %–25 % image's diagonal length) to retain large-scale image characteristics. All such methods compare user sketches with image edges (or boundaries), suffering from influence of background edges when finding a desired object. Salient object region extraction [11, 12, 34] and multi-resolution region representation [32] have been used to handle background clutter. We also use explicit region information to support SBIR. However, instead of feature designing, matching, or indexing, we use visual attention and (learned) global appearance information to improve salient region extraction, which naturally supports shape retrieval with scale, rotation, and translation variations.

Segmentation transfer Our work is also related to recent advances in segmentation transfer. Kuettel and Ferrari [37] transfers segmentation mask from training windows that are visually similar to the target image windows. In an impressive concurrent effect, Kuettel et al. [38] successfully generate pixelwise segmentations for ImageNet [20], which contains 577 classes over 500 K images, by leveraging existing manual annotations in form of class label, bounding-boxes, and external pixel-wise ground truth segmentations in the PASCAL VOC10 dataset [23]. These methods also use class-wise appearance models, captured by the Gaussian Mixture Model, and model the segmentation problem in an extended MRF framework. However, in absence of appropriate methods to choose good segmentations before segmentation propagation, the accuracy degrades gracefully over the stages. Instead, we carefully choose good segmentations by measuring scene complexity, imprecise cut, region incompleteness, and shape consistence. This allows us to select reliable candidates leading to high quality global appearance, which accords with human understanding about the classes (see also Fig. 4 and supplemental materials).[2] Thus, instead of external pixel-accurate ground truth labeling, our method only requires a few (typically one is enough) sketches for each class to help learn useful global appearance information, thus significantly lowering required annotation efforts.

3 Unsupervised segmentation of individual candidate images

For any given keyword (e.g., dogs, jumping dogs, etc.), we first retrieve a set of candidate images using Flickr, typically around 3,000 (see Fig. 1). For each such image, we perform unsupervised segmentation to estimate a salient object, as described next. The key stage comes later (Sect. 4) when we exploit correlation in salient objects' appearance and shape among related images for a query object class, toward group saliency.

3.1 Saliency guided image segmentation

We briefly describe our previous SaliencyCut [15] work, which is used here for single image saliency estimation. Segmenting a color image $I := \{I_i\}$ consisting of pixels I_i in the RGB color space amounts to solving for corresponding *opacity* values $\alpha := \{\alpha_i\}$ with $\alpha_i \in \{0, 1\}$ at each pixel. We

[2] http://mmcheng.net/gsal/.

enable unsupervised segmentation by building a Gaussian Mixture Model (GMM) for foreground/background color distribution G, which we then use to directly extract a binary segmentation mask to avoid manual thresholding.

We use GrabCut formulation [46] to model single image saliency-based segmentation problem and use their suggested parameters. The segmentation problem can be solved by optimizing a *Gibbs* energy function E as

$$\min_{\alpha} E(\alpha, G, I) = \min_{\alpha}\big(U(\alpha, G, I) + V(\alpha, I)\big) \quad (1)$$

where $U(\alpha, G, I)$ evaluates the fitness of the opacity distribution α with respect to the data I under a given color model G and $V(\alpha, I)$ measures the smoothness of α. The fitness term $U(\alpha, G, I)$ of a pixel I_i is defined as the negative log probability of this pixel belonging to its corresponding color distribution G (α is a binary value. $\alpha = 1$ indicates a foreground GMM, while $\alpha = 0$ denotes a background GMM). The smoothness term $V(\alpha, I)$ is defined as the sum of neighboring pixel color similarity when they take different α values (see [46] for details about the measurement and its parameters estimation). In SaliencyCut [15], the continuous saliency values are thresholded to automatically initialize foreground and background color models in GrabCut [46]. To improve robustness to noisy automatic initialization, the segmentation process is iterated, with morphological operations to improve performance (see [15] for more details).

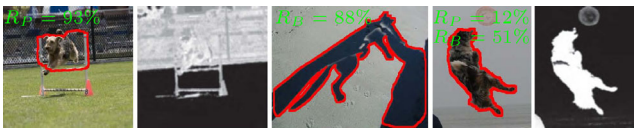
Implementation details Unlike [15], we use a new initialization procedure that avoids the un-intuitive threshold choosing process. Foreground and background are modeled with color GMMs. Instead of assigning pixel colors to a model using a threshold, we treat every pixel color as a weighted sample that contributes to both foreground and background color GMMs, e.g., for a pixel with saliency value 0.7, we use weight factors 0.7 and 0.3 when building the foreground and background color GMMs, respectively. Although this soft assignment incurs a small computational overhead (<10 %) in each SaliencyCut iteration, it introduces more accurate initialization, which reduces the number of iterations required. In the experiments, the overall segmentation quality and computational time is similar to [15] with a manually chosen threshold. During GMM estimation, we use the color quantization bin (see [15]) as a unit of samples instead of each pixel color for computational efficiency.

3.2 Measuring the reliability of SaliencyCut

Keyword based image retrieval often produces a large percentage of irrelevant images (see also Sect. 5.3 and supplementary materials). Luckily, in our retrieval application, users are interested in the precision rate of the top ranked images (e.g., top 50 images) rather than the recall rate of the entire searched results (typically a few 1000 s). Hence, we aggressively prune away likely outlier images, as described next.

original image segmentation [24] saliency [15]

Fig. 2 Complex and cluttered scenes usually lead to segmentations composed of many small regions, or fragmented saliency maps



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

Fig. 3 Different SaliencyCuts are marked in *red* in (**a**), (**c**), and (**d**). While (**a**) is imprecise and (**c**) produces an incomplete region-of-interest, (**d**) yields a good cut. Undesirable cuts are detected based on relevant foreground probability maps (**b**, **e**). Segmentation quality ranking scores (imprecise R_P and incomplete R_B) are overlaid on top

Scene complexity Saliency maps are often poor for complex/cluttered scenes. We use the number of regions produced by segmentation [24] as an indicator of scene complexity (see Fig. 2). Intuitively, images with a small number of segments are simpler. We sort the images based on increasing number of regions and retain only the top T_R images for subsequent stages. We use $T_R = 70$ % in our tests leading to around a thousand images being discarded.

Note that in our problem setting huge sets of candidates are available, e.g., internet images, and the users simply want to easily find some high quality desired targets rather than explore the whole collection. Hence, as a design choice, we decided to favor higher precision over higher recall. We empirically threshold values and use them for all our tests. This resulted in reliable global statistics for group segmentation.

Segmentation quality Even for relatively simple scenes, SaliencyCut can have imprecise or incomplete boundaries, which we detect as follows: (i) For each image, we use its SaliencyCut region and its remaining parts to train GMMs for foreground and background regions, respectively. We then estimate the foreground probability of relevant image pixels according to these two GMMs. Specifically, we take the sum of foreground probabilities for pixels inside a narrow band (of 30 pixel width) surrounding the SaliencyCut as a measure for imprecise cut (e.g., Fig. 3a): the higher this sum, the lower the predicted quality of the cut. (ii) We take the total number of SaliencyCut region pixels within a narrow band (of 20 pixel width) along the image border as a

measure for incompleteness of the object-of-interest region (see Fig. 3c): the higher this number, the more likely the cut object region is incomplete. We sort the images according to increasing order of the above two measures and retain the top T_P and T_B of images for subsequent stages. We use $T_P = 80$ % and $T_B = 80$ % in our experiments. The retained images are next analyzed for image collection consistency.

4 Group saliency

Retrieved images in a collection (e.g., Flickr) are largely correlated, but can have differences due to pose, appearance, etc. (see Fig. 8 and also supplementary). We use rough sketches as an indicator of the poses that the user is interested in (e.g., when user searches for a specific style of 'dog jump'), while we use consistency across the retrieved images to extract what are plausible appearance models for the salient object (e.g., color of the 'dog'). Such saliency, which we refer to as *group saliency*, favors both *similarities* between images and *distinctness* within each image. Specifically, we first use an efficient cascade model to rank single image unsupervised segmentation results according to their shape consistency with user input sketch. Top ranked results are then used to train a global appearance statistical model for refining group saliency and segmentation results.

4.1 Cascade filtering for sketch based retrieval

Having reliable SaliencyCut boundaries from initial images, we benefit from existing shape matching algorithms [5, 49] to retrieve desired images based on consistency with user-input sketches. We use the following simple measures, which are easy to calculate on regions with clean background:

- circularity: $Perimeter^2/Area$ [49],
- solidity: $RegionArea/ConvexHullArea$ [26],
- Fourier Descriptor [50], and
- Shape Context [2].

We proceed in a cascaded fashion. For each measure, we sort the shapes in decreasing order based on their similarity to user sketch, and retain a top percentage of the candidates. The measures are arranged in increasing complexity, allowing efficient and early rejection of candidates with large dissimilarity. In our experiments, we keep $T_C = 80$ %, $T_S = 80$ %, and $T_F = 70$ % images according to circularity, solidity, and Fourier descriptor, respectively; the corresponding dimensions for these descriptors are 1, 1, and 15. We compare these descriptors using simple Euclidean distance with corresponding features of the user sketch. We finally use the Shape Context (with default parameters and matching methods as suggested in the original paper [2]), which is complex but effective to properly order the remaining candidates. While one can employ more complex shape descriptors (e.g., [22]), we find the above selection sufficiently diverse to prune out most shape outliers (see also supplementary material). Note that at this stage we are left with only $T_R \cdot T_P \cdot T_B \cdot T_C \cdot T_S \cdot T_F \approx 20.0$ % of the image shapes, which are used for appearance consistency, as described next.

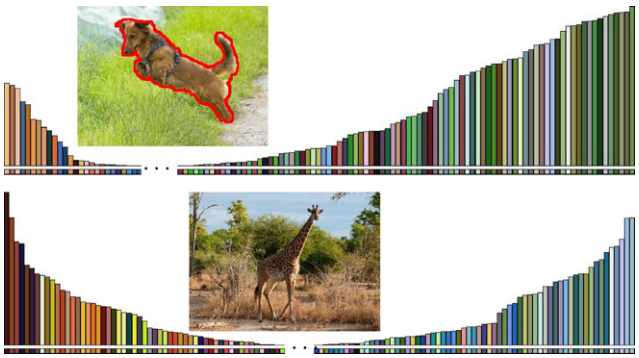


Fig. 4 Appearance histogram of sample colors in object of interest region in 'dog jump' images according to the learned global prior. Left shows typical foreground colors, while right shows typical background colors. Probability values $\{p\}$ are ordered and histogram height is $|p - 0.5|$. We ignore color samples with probability around 0.5 as chance. Inset shows a typical input image example

4.2 Statistical global appearance models

After cascade filtering, the top ranked images typically have high precision. We use the top 50 images as a high quality training set to learn a global appearance prior to guide subsequent group saliency detection and segmentation. We choose GMM models to capture such a prior $\bar{G}$ for two reasons: (i) GMM models generalize better on small amounts of training data than histogram models [35]; (ii) GMM priors can be easily integrated in our unsupervised image segmentation framework (see Sect. 3). For example, Fig. 4 shows foreground and background GMM models for the 'dog jump' example indicating that dogs are typically yellow or dark in color and are like to play on green grass/fields; for the 'giraffe' we find typical background colors consist of blue/green indicating sky/trees, as typically associated with context information for giraffes. Although other attributes like texture and visual vocabulary can be considered, we currently use only color. We empirically chose 8 Gaussians each to model major appearance of foreground/background per category and found that this number is not sensitive.

Note that since shape features are typically orthogonal to appearance attributes, the samples we retrieved based on shape largely preserve their appearance diversity and can be used to learn representative appearance models. Typically only a fraction of such images (15 %–57 % in our tests) contain desired objects. These objects may have different

colors, textures, and even a single object may comprise of several regions with very different appearance (e.g., butterfly). We found that considering the largest appearance cluster [3] or top-ranked internet images [25] as an initial set to be unsatisfactory. In an interesting effort, Chang et al. [8] use repeatedness among images as a global prior of multiple images and assume that most images contain at least parts of the foreground object, an assumption that is often violated in our setting. Further, since each image is compared with all others, the method cannot be used for large collections (e.g., they considered image sets of maximum size 30, while we handle a few 1000 s).

4.3 Estimating group saliency

Finally, we use the learned global appearance statistics to improve the saliency detection and SaliencyCut of each image. Since the estimated global color prior $\bar{G}$ is encoded as GMMs, we simply add a global prior constraint to our single image unsupervised segmentation energy function of Eq. (1). The new energy function takes the form:

$$E(\alpha, G, \bar{G}, I) = \lambda U(\alpha, \bar{G}, I) + (1-\lambda)U(\alpha, G, I) + V(\alpha, I) \quad (2)$$

where, the additional term $U(\alpha, \bar{G}, I)$ evaluates the fitness of the opacity distribution α to the global color prior $\bar{G}$, while weight λ (0.3 in our test) balances between global color prior and per image color distribution. Here, the global color prior $\bar{G}$ reflects similarity between the targets, while the per-image color distribution G is trained according to the individual image saliency map and captures distinctness within an image.

Similar to Eq. (1), we optimize Eq. (2) to get group saliency segmentation results. We then encode group saliency maps as probability maps of pixels belonging to the object-of-interest region obtained by group saliency segmentation. Note that although the change compared to Eq. (1) is small, the improvement in estimated saliency is significant with only marginal computational overhead.

Figure 5 demonstrates typical improvements in saliency cut and segmentation using global color priors. In the 'dog jump' image, the green parts of the image are estimated to be more likely to be background rather than foreground according to the learned global color prior. Similarly, in the plane example, missing object regions are correctly recovered with the help of global statistics (see supplementary material for more examples).

5 Experiments

We implemented our framework in C++ and evaluated it using a Quad Core i7 920 CPUs with 6 G RAM. We use the

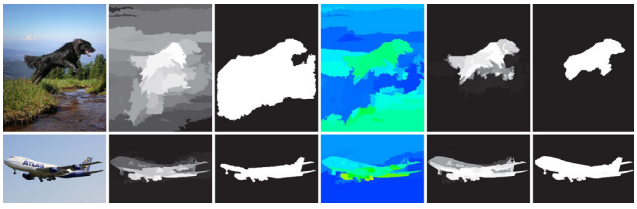


Fig. 5 Examples of using statistics to refine unsupervised segmentation: (**a**) source image, (**b**) single image saliency map, (**c**) SaliencyCut (using Eq. (1)), (**d**) global color prior, (**e**) group saliency map, and (**f**) group saliency segmentation (using Eq. (2)). Note the improvement from (**c**) to (**f**) (see supplementary material)

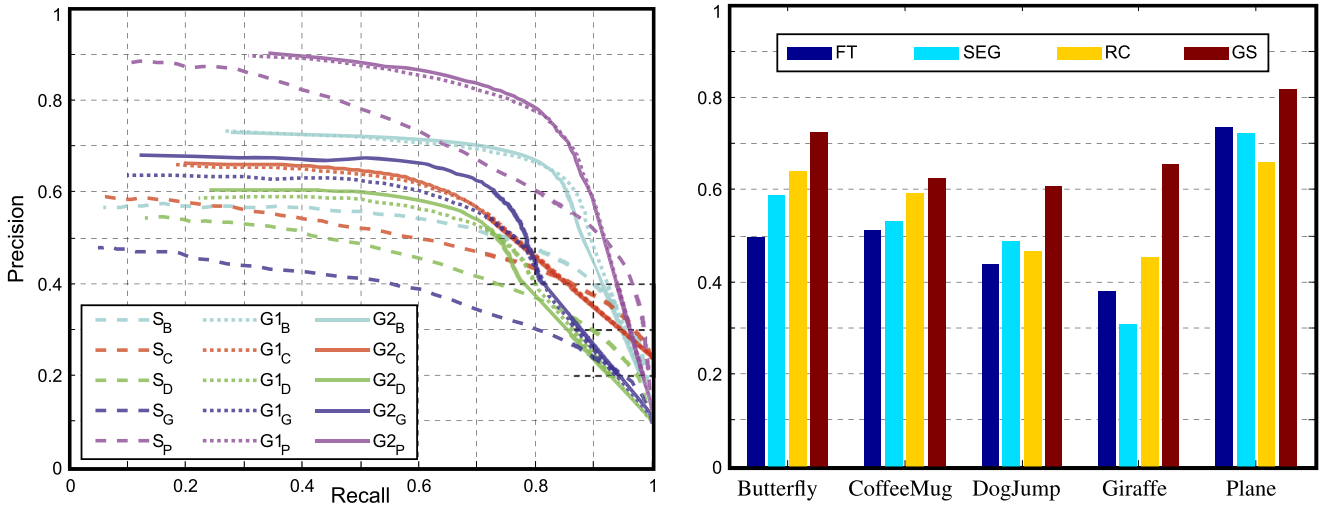
Fig. 6 Example images from the benchmark dataset that correspond to the keyword 'dog jump', with pixel-accurate ground-truth labeling for the corresponding object of interest regions (if such a region exists; 4 out of the 7 here)

group saliency based retrieval results to re-train new appearance statistics and iteratively improve saliency segmentation (see Fig. 1). Experimentally, we found two rounds of iterations to be sufficient.

We evaluated the proposed method for three different applications using a benchmark dataset: (i) fixed thresholding of group saliency maps, (ii) object of interest segmentation, and (iii) sketch based image retrieval. For the first two applications, we consider the average segmentation performance only over those images that do contain the target object (according to annotated ground truth).

5.1 Benchmark dataset for saliency segmentation

We collected a labeled dataset of categorized images initially extracted by querying with keywords from Flickr. We downloaded about 3,000 images for each of the 5 keywords: 'butterfly', 'coffee mug', 'dog jump', 'giraffe', and 'plane', and *manually annotated* saliency maps to mark the object regions (see Fig. 6 and supplementary material for examples; full dataset to be publicly made available). To normalize these images, we uniformly scale them so that their maximal dimension is 400 pixels long. Some images in the dataset do not contain any salient object matching the keyword; we leave such images unlabeled. Further, since partially occluded objects are less reliable for shape matching, we only label object regions that are mostly un-occluded. In the end, we got 6000+ images with pixel accurate ground



(a) Fixed thresholding saliency maps

(b) Object of interest region segmentation

Fig. 7 Evaluation results on our benchmark dataset. (**a**) Precision-recall curves for naive thresholding of saliency maps. S, G1, G2 represent single image saliency, group saliency after the 1st and 2nd iterations, respectively. Subscripts B, C, D, G, P represent groups of 'butterfly', 'coffee mug', 'dog jump', 'giraffe', 'plane', respectively. (**b**) Comparison of F_β for image groups using single image saliency segmentation methods (FT [1], SEG [45], RC [15]) vs. group saliency (GS) segmentation. The RC and GS in (**b**) corresponds to results of Eq. (1) and Eq. (2) respectively

truth segmentation (see Fig. 6 for sample images and supplemental material for more statistics). Note that Eitz et al. [22] introduce a benchmark dataset for evaluating SBIR systems by annotating how well a given sketch and image pair match. Our benchmark contains pixel-accurate segmentation of the targets, when present, thus allowing evaluation of corresponding segmentation algorithms. Our dataset is 15× larger than previously largest public available benchmark [1] with pixel accuracy salient region annotation. In contrast to the benchmark in [1], where salient regions are unambiguous, clearly separated from the background, and often positioned near the image centers, images in our proposed dataset are more challenging and represent typical cluttered real-world scenes.

5.2 Fixed thresholding of group saliency maps

We threshold the saliency map with $T \in [0, 255]$ and compare the segmentation results with ground truth labeling (see also [1]). The precision and recall curves in Fig. 7(a) show that our group saliency algorithm stabilizes after 2 iterations and significantly outperforms state-of-the-art single image saliency detection method [15].

5.3 Object of interest segmentation

We also evaluate how accurately our algorithm extracts target objects from heterogeneous internet images. For images containing a target, we compare their pixel-level labeling with our group-saliency segmentation according to precision, recall, and F-Measure, which is defined as

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}. \quad (3)$$

Note that for salient region segmentation, precision is more important than recall [1, 15], since recall can trivially be 100 % by taking all image regions as targets. For internet retrieval, precision is more important as a false detection is undesirable over missing some good candidates among thousands of possibilities. Hence, we use $\beta^2 = 0.3$ to weight precision more than recall for fair comparison with state-of-the-art methods [1, 15, 45]. Figure 7(b) illustrates the improvement due to group saliency segmentation.

Note that most of the nature-scene images contain multiple objects and be associated with multiple text tags in by the search engine. We use the group saliency based segmentation to extract saliency object regions in images of the same group. This allows us to extract objects in an image even when the text tags differ.

5.4 Sketch based image retrieval

As an application, we compare our retrieval algorithm with state-of-the-art SBIR proposed by Eitz et al. [22] (using author implementations). Our method explicitly extracts object of interest regions from images, thus enabling us to exploit the power of existing shape matching techniques. For heterogeneous internet images, the combination of group-saliency segmentation and shape matching effectively selects good images containing target objects, leading to im-

Table 1 True positive ratios (TPR) among top 50 and 100 retrievals using Flickr, our method, and SHoG [22], for 30 different categories

TPR (%)	Among top 50			Among top 100		
	Flickr	Our	SHoG	Flickr	Our	SHoG
Bottle	58	98	84	60	94	82
Butterfly	28	54	36	28	50	40
Cake	56	84	78	55	81	79
Cattle	48	94	14	37	81	12
Coffee mug	58	94	82	51	90	78
Cow	52	94	54	54	92	55
Crow	36	96	76	38	86	61
Dog jump	56	92	74	55	85	73
Eagle	18	94	82	23	95	80
Elephant	40	96	76	38	87	68
Flag	56	96	54	52	91	57
Fox	32	96	38	37	80	42
Giraffe	30	64	18	25	51	18
iPhone	48	96	68	43	90	54
Jeep	54	94	38	59	93	43
Mailbox	52	86	68	52	87	60
Orange	24	70	70	19	62	52
Parrot	58	90	60	56	79	57
Pear	64	84	82	63	82	81
Plane	44	94	90	48	93	91
Sailboat	52	86	56	67	81	54
Seagull	56	92	88	59	90	86
Sheep	54	78	44	50	81	40
Snail	62	94	60	63	87	61
Strawberry	50	70	76	53	69	67
Swallow	24	76	74	31	66	68
Tank	48	88	56	44	84	45
Tortoise	48	62	46	48	62	45
Wolf	24	50	30	25	55	29
Zebra	30	66	28	25	56	31
Average	**46.0**	**84.2**	**60.0**	**45.3**	**79.3**	**57.0**

proved results (see Table 1, Fig. 8, and supplementary material). We leave exploring benefits of hybrid systems using additional attributes including appearance [18], local features [22], or additional lines [7] to future research.

We pre-process each image in the database by performing single image unsupervised segmentation (about 100 images per minute). Further, we pre-process each category using a representative sketch to initialize a good appearance learning and unsupervised segmentation of the object. In Fig. 8(e), we show retrieval results for 'plane' with two *different* input sketches, for which results for the second sketches just have to compare salient shapes generated with the help of the first sketch. Our results contain explicit region information allowing input sketches to retrieve results with more relevant pose.

Note that currently our system expect users to supply both keywords and sketches. Existing shape matching techniques, which are able to effectively select high quality matchings from shapes with clean background, even with very rough sketches (e.g., state-of-the-art method [48]), could achieve 93.3 % accuracy in the very challenging MPEG7 shape dataset. Once the user inputs a rough sketch to help distinguish between desired object and irrelevant region shapes, it would help us to get useful global appearance information. The low correlation between shape feature and appearance statistics allows us to reuse the input sketch, learned appearance, and segmented regions. Recent advances in human object sketch classifications [21] can potentially be used in conjunction with our system toward a keyword-free retrieval interface, which can be attractive for gesture-based devices. At runtime, we only compare a new user sketch with object shapes using the cascade filtering process (see Sect. 4) taking less than 1 second to handle an initial retrieved set of 3,000 images. For larger databases, efficient retrieval algorithms using shape context [42] may be useful. Our method only segments the most salient object region from each image and perform retrieval based on that object region. Since there are a huge number of internet images, we are mainly focused on quality of the top ranked results rather than the recall of every image.

6 Conclusion

We introduced a method to exploit correlations across internet images within same categories to achieve superior salient object segmentation and image retrieval. Starting from a simple user sketch, we estimate high quality image labeling to build appearance models for target image regions and their backgrounds. These appearance models are in turn used to improve saliency detection and image segmentation. We introduced a benchmark consisting of 6000+ pixel-accurate labeled dataset initially obtained by querying keywords from Flickr and use it to demonstrate that our proposed method produces high quality saliency maps and segmentation, with potential application to SBIR. Our approach makes use of the powerful user sketch information to select good segmentation candidates for getting global appearance information (see Sect. 4.2). This selection process avoids error accumulative problems which typically exist in iterative segmentation transfer methods [38], resulting in the consistent result improvements observed in our experiments.

In the future, we plan to learn additional texture and shape statistics [9] to further improve the segmentation. We also plan to investigate efficient shape indexing algorithms [4, 43] and GPU speed up [47] for increased efficiency.

(a) butterfly (b) coffee mug (c) dog jump (d) giraffe (e) plane

Fig. 8 SBIR comparison. In each group from left to right, *first column* shows images downloaded from Flickr using the corresponding keyword; *second column* shows our retrieval results obtained by comparing user-input sketch with group saliency segmentation results; *third column* shows corresponding sketch based retrieval results using SHoG [22]. Two input sketches with their retrieval results are shown in (**e**)

Acknowledgements We would like to thank the anonymous reviewers for their constructive comments. This research was supported by the 973 Program (2011CB302205), the 863 Program (2009AA01Z327), the Key Project of S&T (2011ZX01042-001-002), and NSFC (U0735001). Ming-Ming Cheng was funded by Google Ph.D. fellowship, IBM Ph.D. fellowship, and New Ph.D. Researcher Award (Ministry of Edu., CN).

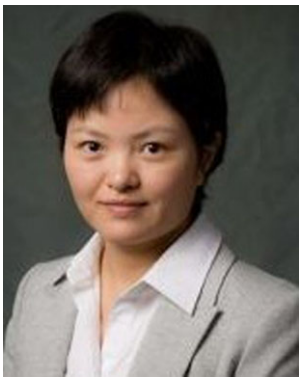
References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: IEEE CVPR, pp. 1597–1604 (2009)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
3. Ben-Haim, N., Babenko, B., Belongie, S.: Improving web-based image search via content based clustering. In: IEEE CVPRW, p. 106 (2006)
4. Biswas, S., Aggarwal, G., Chellappa, R.: An efficient and robust algorithm for shape indexing and retrieval. IEEE Trans. Multimed. **12**, 371–385 (2010)
5. Bouet, M., Khenchaf, A., Briand, H.: Shape representation for image retrieval. In: ACM MM, pp. 1–4 (1999)
6. Cao, Y., Wang, C., Zhang, L., Zhang, L.: Edgel index for large-scale sketch-based image search. In: IEEE CVPR, pp. 761–768 (2011)
7. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., Zhang, L.: Mindfinder: interactive sketch-based image search on millions of images. In: ACM MM, pp. 1605–1608 (2010)
8. Chang, K., Liu, T., Lai, S.: From co-saliency to co-segmentation: an efficient and fully unsupervised energy minimization model. In: IEEE CVPR, pp. 2129–2136 (2011)
9. Charpiat, G., Faugeras, O., Keriven, R.: Shape statistics for image segmentation with prior. In: IEEE CVPR, pp. 1–6 (2007)
10. Chen, H.: Preattentive co-saliency detection. In: IEEE ICIP, pp. 1117–1120 (2010)
11. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet image montage. ACM Trans. Graph. **28**(5), 124:1–124:10 (2009)
12. Chen, T., Tan, P., Ma, L.Q., Cheng, M.M., Shamir, A., Hu, S.M.: Poseshop: human image database construction and personalized content synthesis. IEEE Trans. Vis. Comput. Graph. **19**(5), 824–837 (2013)
13. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Salient object detection and segmentation. Tech. rep., Tsinghua University (2011). http://mmcheng.net/SalObj/. Submission NO. TPAMI-2011-10-0753
14. Cheng, M.M., Zhang, F.L., Mitra, N.J., Huang, X., Hu, S.M.: Repfinder: finding approximately repeated scene elements for image editing. ACM Trans. Graph. **29**(4), 83:1–83:8 (2010)
15. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: IEEE CVPR, pp. 409–416 (2011)
16. Chia, Y.S., Zhuo, S., Gupta, R.K., Tai, Y.W., Cho, S.Y., Tan, P., Lin, S.: Semantic colorization with internet images. ACM Trans. Graph. **30**(6) (2011). doi:10.1145/2024156.2024190
17. Cui, J., Wen, F., Tang, X.: Real time Google and live image search re-ranking. In: ACM MM, pp. 729–732 (2008)
18. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 1–60 (2008)
19. Del Bimbo, A., Pala, P.: Visual image retrieval by elastic matching of user sketches. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 121–132 (1997)
20. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE CVPR, pp. 248–255 (2009)
21. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. Graph. **31**(4) (2012). doi:10.1145/2185520.2185540
22. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: benchmark and bag-of-features descriptors. IEEE Trans. Vis. Comput. Graph. **17**, 1624–1636 (2011)
23. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The Pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
24. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)
25. Fergus, R., Perona, P., Zisserman, A.: A visual category filter for Google images. In: ECCV, pp. 242–256 (2004)
26. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., et al.: Query by image and video content: the QBIC system. Computer **28**(9), 23–32 (1995)
27. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. IEEE Trans. Image Process. **21**(9), 4290–4303 (2012)
28. Gao, Y., Wang, M., Zha, Z., Tian, Q., Dai, Q., Zhang, N.: Less is more: efficient 3d object retrieval with query view selection. IEEE Trans. Multimed. **11**(5), 1007–1018 (2011)
29. Han, J., Ngan, K., Li, M., Zhang, H.: Unsupervised extraction of visual attention objects in color images. IEEE Trans. Circuits Syst. Video Technol. **16**(1), 141–145 (2006)
30. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: IEEE CVPR, pp. 3005–3012 (2012)
31. Hirata, K., Kato, T.: Query by visual example-content based image retrieval. In: Advances in Database Technology-EDBT, pp. 56–71 (1992)
32. Hu, R., Wang, T., Collomosse, J.: A bag-of-regions approach to sketch-based image retrieval. In: IEEE ICIP, pp. 3661–3664 (2011)
33. Hu, S.M., Chen, T., Xu, K., Cheng, M.M., Martin, R.R.: Internet visual media processing: a survey with graphics and vision applications. Vis. Comput., 1–13 (2013). doi:10.1007/s00371-013-0792-6
34. Huang, H., Zhang, L., Zhang, H.: Arcimboldo-like collage using Internet images. ACM Trans. Graph. **30**(6), 155 (2011)
35. Jones, M., Rehg, J.: Statistical color models with application to skin detection. Int. J. Comput. Vis. **46**(1), 81–96 (2002)
36. Ko, B., Nam, J.: Object-of-interest image segmentation based on human attention and semantic region clustering. J. Opt. Soc. Am. **23**(10), 2462–2470 (2006)
37. Kuettel, D., Ferrari, V.: Figure-ground segmentation by transferring window masks. In: IEEE CVPR, pp. 558–565 (2012)
38. Kuettel, D., Guillaumin, M., Ferrari, V.: Segmentation propagation in ImageNet. In: ECCV, pp. 459–473. Springer, Berlin (2012)
39. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE Trans. Image Process. **20**(12), 3365–3375 (2011)
40. Liu, H., Zhang, L., Huang, H.: Web-image driven best views of 3d shapes. Vis. Comput., pp. 1–9 (2012). doi:10.1007/s00371-011-0638-z
41. Margolin, R., Zelnik-Manor, L., Tal, A.: Saliency for image manipulation. Vis. Comput., pp. 381–392 (2013). doi:10.1007/s00371-012-0740-x
42. Mori, G., Belongie, S., Malik, J.: Shape contexts enable efficient retrieval of similar shapes. In: IEEE CVPR, pp. 723–730 (2001)
43. Peter, A., Rangarajan, A., Ho, J.: Shape L'ane rouge: sliding wavelets for indexing and retrieval. In: IEEE CVPR, pp. 1–8 (2008)
44. Popescu, A., Moëllic, P., Kanellos, I., Landais, R.: Lightweight web image reranking. In: ACM MM, pp. 657–660 (2009)

45. Rahtu, E., Kannala, J., Salo, M., Heikkila, J.: Segmenting salient objects from images and videos. In: ECCV, pp. 366–379 (2010)
46. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut"—interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3), 309–314 (2004)
47. Schmid, J., Guitián, J.A.I., Gobbetti, E., Magnenat-Thalmann, N.: A GPU framework for parallel segmentation of volumetric images using discrete deformable models. Vis. Comput. **27**(2), 85–95 (2011)
48. Yang, X., Koknar-Tezel, S., Latecki, L.: Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In: IEEE CVPR, pp. 357–364 (2009)
49. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognit. **37**, 1–19 (2004)
50. Zhang, D.S., Lu, G.J.: Shape-based image retrieval using generic Fourier descriptor. Signal Process. Image Commun. **17**(10), 825–848 (2002)
51. Zhang, G.X., Cheng, M.M., Hu, S.M., Martin, R.R.: A shape-preserving approach to image resizing. Comput. Graph. Forum **28**(7), 1897–1906 (2009)
52. Zhang, L., Huang, H.: Hierarchical narrative collage for digital photo album. Comput. Graph. Forum **31**(7), 2173–2181 (2012)
53. Zheng, Y., Chen, X., Cheng, M.M., Zhou, K., Hu, S.M., Mitra, N.J.: Interactive images: Cuboid-based scene understanding for smart manipulation. ACM Trans. Graph. **31**(4) (2012). doi:10.1145/2185520.2185595

Ming-Ming Cheng received his Ph.D. degree from Tsinghua University in 2012. He is currently a research fellow in Oxford Brookes University, working with Professor Philip Torr.

Niloy J. Mitra received his Ph.D. degree from Stanford University in 2006. He is currently a reader (associate professor) at the University College London (UCL). He is on the editorial boards of Computer & Graphics, and Visual Computer.



Xiaolei Huang received her Ph.D. degrees from Rutgers University in 2006. She is currently an assistant professor in Lehigh University.

Shi-Min Hu received the Ph.D. degree from Zhejiang University in 1996. He is currently a chair professor of computer science in the CS Dept., Tsinghua University. He is on the editorial board of Computer Aided Design, The Visual Computer, and Computer & Graphics.