EUROGRAPHICS 2015 / O. Sorkine-Hornung and M. Wimmer
(Guest Editors)

Volume 34 (2015), Number 2

Distilled Collections from Textual Image Queries

Hadar Averbuch-Elor[1] Yunhai Wang [2,3†] Yiming Qian [3] Minglun Gong [3] Johannes Kopf [4] Hao Zhang [5] Daniel Cohen-Or [1]

[1] Tel Aviv University [2] Shenzhen VisuCA Key Lab/SIAT [3] Memorial University [4] Microsoft Research [5] Simon Fraser University

Abstract

We present a distillation algorithm which operates on a large, unstructured, and noisy collection of internet images returned from an online object query. We introduce the notion of a distilled set, which is a clean, coherent, and structured subset of inlier *images. In addition, the object of interest is properly segmented out throughout the distilled set. Our approach is unsupervised, built on a novel clustering scheme, and solves the distillation and object segmentation problems simultaneously. In essence, instead of distilling the collection of images, we distill a collection of loosely cutout foreground "shapes", which may or may not contain the queried object. Our key observation, which motivated our clustering scheme, is that outlier shapes are expected to be random in nature, whereas, inlier shapes, which do tightly enclose the object of interest, tend to be well supported by similar shapes captured in similar views. We analyze the commonalities among candidate foreground segments, without aiming to analyze their semantics, but simply by clustering similar shapes and considering only the most significant clusters representing non-trivial shapes. We show that when tuned conservatively, our distillation algorithm is able to extract a near perfect subset of true inliers. Furthermore, we show that our technique scales well in the sense that the precision rate remains high, as the collection grows. We demonstrate the utility of our distillation results with a number of interesting graphics applications.*

1. Introduction

Billions of new images of all kinds of objects are uploaded to the internet every single day. Text-based object queries through search engines such as Google, Bing, and Flickr, allow combing through this sea of data and enable extracting large topical but otherwise unstructured wild image collections. The vast visual knowledge encoded in such collections has been previously tapped into to enable various applications, such as building 3D models of real places [SSS06], sketch-based photo composition [CCT*09, EHBA09], or data-driven analysis and synthesis [LWQ*08, ZGW*13].

A major challenge when working with unstructured internet image collections is, however, that the image-to-text association is noisy, and, hence, the retrieved collections typically contain many false positives: noise images that do not show the object of interest at all or only in insufficient quality due to resolution or distortion; see Figure 1(a). Moreover, many of the aforementioned applications require not just images that contain the object of interest but also that the object is extracted from its background.

The objective of our work is to generate a *distilled* image collection from raw internet search results based on a textual object query. The distilled collection is a clean, coherent, and consistent subset consisting only of inlier images, i.e., images that do contain the queried object. Ultimately, we would like the distilled set to be outlier-free. In addition, the object of interest is cleanly segmented out from the background in all images in the distilled collection. Figure 1 provides an example where the search was for "elephant" and Figure 1(c) shows the distilled collection.

Extracting only inliers from the noisy image collection and object segmentation are inter-dependent problems. An effective solution to either problem would facilitate a solution to the other. The challenge that both tasks face is that it would seem inevitable that some high-level semantic analysis is

† Corresponding author: Yunhai Wang (cloudseawang@gmail.com)

Figure 1: *(a) Unstructured image collections from text queries are noisy and typically contain many images that do not show the object of interest (such outliers are marked with red borders). (b) Single-image segmentation is unreliable and produces many erroneous shapes, which either do not contain the object, cut off parts of it, or include chunks of the background. (c) Our distillation algorithm extracts only a subset of kernel of inlier shapes and organizes them into clusters (marked with colored boundaries).*

necessary to identify the object of interest in every image and to measure the relevancy of each image to the object query. This would call for object-specific prior knowledge to be acquired and then learned to solve the problems.

In this paper, we take an unsupervised approach without relying on any object-specific knowledge. Our approach not only breaks the inter-dependence between the two problems but also solves both of them simultaneously. In essence, instead of distilling the collection of images, we distill a collection of loosely cutout foreground "shapes" which may or may not contain the queried object. Our key realization is that outlier shapes (e.g., wrong object or poor segmentations that include background or miss parts of the foreground) should be expected to be random in nature. On the other hand, inlier shapes, i.e., regions that do tightly enclose the object of interest (perhaps in varying views), tend to be well supported by similar shapes (corresponding to similar object views) from other images. Such supports are expected to be significant if the raw image collection is sufficiently large.

We use this idea to develop an image collection distillation algorithm that does not rely on understanding the semantics of images or their parts. Its core component is a novel unsupervised *constructive* shape clustering algorithm. Since the object of interest exhibits large variation in size and appearance over the image collection, we use closed contours as the main clustering feature in this algorithm, as they tend to be robust against such changes. We obtain object contour candidates using a standard single-image segmentation technique, which by itself is unreliable in the sense that it produces a large fraction of bad segmentations (Figure 1b). However, our algorithm is able to filter these outliers as they do not form tight clusters.

Our algorithm is tuned to aggressively prune a raw image collection and conservatively extract only true inliers, at the expense of leaving out some *false* negatives, i.e., images that do contain the object of interest but our algorithm has not gained sufficient confidence in.

We show that conventional object co-segmentation methods do not perform well in this context. Nearly all techniques, except the most recent ones, are designed for homogeneous datasets and assume the object of interest is present in every input image. They are, thus, bound to fail with outlier images. More recent techniques are specifically designed to handle noise [RJKL13], however, as any co-segmentation method, they try to correctly label every pixel in every image. This, however, is extremely challenging for images that have a less common appearance in the dataset, even if they are inliers. Hence, their method still produces a relatively large fraction of erroneous segmentations and is ultimately not reliable enough for graphics applications.

Our distilled image collections support a variety of 3D and 2D applications, such as image-based viewpoint selection, upright orientation detection, color design, sketch2photo, Captcha and, in general, data-driven applications. Prior to describing some of these applications, we elaborate on a novel application of generating an abstract 3D model from the distilled collection. Since the distilled images are obtained from different instances in a variety of articulations and the view directions are unknown, this type of modelling is an extremely difficult problem. We demonstrate how the distilled set alleviates the process, and facilitates the quick construction of a 3D abstract model.

2. Related work

Image co-segmentation

Given multiple images with shared content, the goal of image co-segmentation is to simultaneously segment a specific object that appears in the entire collection. Early work in image co-segmentation focuses on extracting the same object from a pair of images [RMBK06, MSD09, HS09]. These techniques were recently extended to handle a large number of images and/or object classes (e.g., [JBP12, KX12]). The ClassCut [ADF10] technique, for example, aims at co-segmenting a set of images capturing object instances of an unknown class. Their method alternates between segmenting object instances and learning a class model. Our image segmentation method bears some similarity to the object co-segmentation technique [VRK11] which incorporates the notion of objectness into the co-segmentation framework to ensure the foreground segment is an object. However, their work is supervised and requires ground truth segmentation of pairs of images depicting similar objects to define the classifier.

Co-segmentation in noisy collections is addressed in the recent work of Rubinstein et al. [RJKL13]. They proposed an algorithm that automatically discovers and segments out a common object. Every pixel in the image is labelled as foreground or background, and outlier images can be identified as those images containing only background labelling. Chen et al. [CSG14] improved these numbers by using automatically learned visual priors. The average success rate is still not high enough to enable various applications we are interested in. We therefore designed our method to only label a *subset* of the collection, but at a significantly higher success rate.

Unsupervised object discovery

Unsupervised discovery of visual categories in a collection of images is a fundamental problem in computer vision and many solutions have been proposed [TLBB10]. It is usually approached by clustering the image collection into meaningful groups of shared visual properties. Contour-based methods which discover the common object shapes in an unlabelled multi-category collection of images (e.g., [LG09, PT10]) are most closely related to our work. However, we focus on discovering multiple shape variations of one object category. Furthermore, we experiment with a noisy internet image collection, in contrast to benchmark datasets that contain only images belonging to the specified categories.

2.1. Image Classification and Annotation

Image classification and annotation algorithms are important for scene understanding [LSFF09]. Many approaches to images classification design image features [OT01, VFJZ01, LFF07], and apply either discriminative [CW04, ZZ06] or generative models [CFF07, LFF07]. A recent breakthrough work [KSH12] outperforms previous state-of-the-art methods on large image collections and with a large number of classes using a deep convolutional neural network. Several new deep learning methods [SVZ13, JSD*14] were later proposed, which further improve the classification accuracy. Our distillation technique can be added as a post-processing step to image classification methods. Furthermore, our object segmentation provides a prior which can improve the classification accuracy [RLYFF12].

For image annotation, generative methods [BDF*03, JLM03, DBdFF02] attempt to learn the relationships between images and annotation terms with probabilistic models, while discriminative methods [YDH06, CCMV07] train classifiers for image labeling. In order to annotate large-scale image collection, Westo et al. [WBU10] propose to learn generative model with a method which learns to rank. Since annotations and classification can support each other, Wang et al. [WBL09] propose a probabilistic model for jointly modeling the image, its class label, and its annotations.

Supervised filtering methods

Our conservative distillation method bears some conceptual similarity to the aggressive filtering used in the Sketch2Photo system [CCT*09]. In their work, all shown results including the numeric evaluations are filtered with a user provided silhouette sketch. Furthermore, their method contains application-driven filtering stages, for example, pruning images with complex backgrounds. In this work, we are interested in an unsupervised technique that avoids such pre-filtering, while obtaining a lower false positive rate.

3. Overview

Our goal in this work is to automatically extract and segment a common object of interest from an unstructured image collection obtained using internet search engines.

This objective poses significant challenges: First, internet image collections are noisy. Typically, a significant fraction of images does not contain the object of interest at all or only depicts a portion of it. The remaining images, that show the full object, still vary considerably in pose, appearance, resolution, noise characteristics, etc. Moreover, the images depict different instances, and the object might be inherently deformable. This precludes using feature matching approaches, e.g., as common in Structure-from-Motion techniques, since they rely on non-deforming rigid scenes.

The distillation task is broken into two steps. First, we perform single-image segmentation to generate candidate shapes (Section 4). Then, the distilled inlier sets are formed by a constructive algorithm that considers the outer segment contours as the compared features and is thus robust against

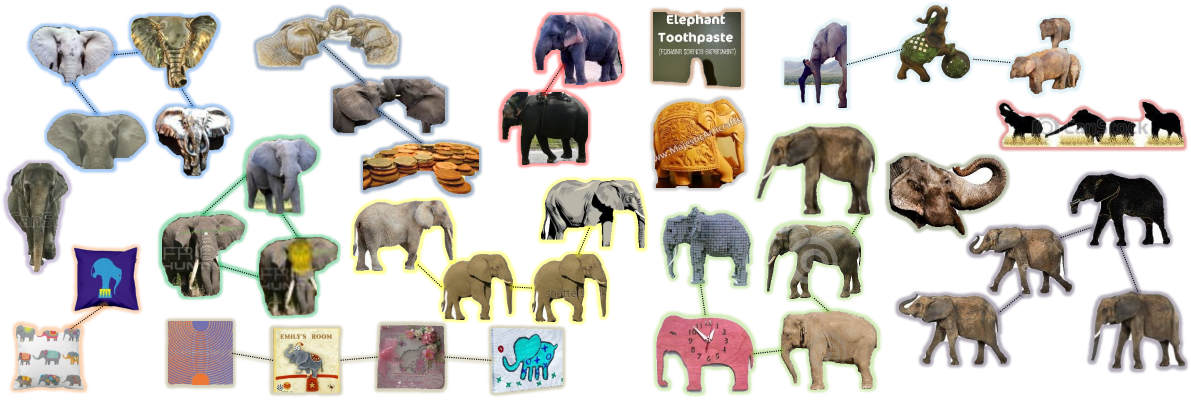

Figure 2: *Identifying hot spot candidates with a mutual kNN graph. Each shape forms a node in the kNN graph. Two nodes are connected if and only if they are both among each other's k-nearest neighbors in terms of their inter-contour distance. The hot spots are a subset of the connected components that contain at least three nodes.*

appearance changes (Section 5). The algorithm is constructive in the sense that it additively collects good segments rather than subtractively filtering outliers. The key idea in this operation is that tight clusters of *non-trivial* shapes are a strong indication of being inliers, because erroneous segments usually have non-repeated defects and thus do not exhibit strong commonality. Assembling a small clean distilled subset comes at the expense of removing a fraction of inliers, however, we show that the distilled sets preserve sufficient variety to enable many applications (Section 7).

4. Generating Candidate Segments

We use standard internet search engines and automatic query expansion to construct the initial unstructured image collection [RJKL13]. Our distillation algorithm requires a set of candidate image segmentations as input. These can be generated either using existing object co-segmentation techniques or using a simple single-image segmentation algorithm. For simplicity, we use a single-image segmentation algorithm, which yields smaller distilled collections but comparable precision and recall values, and rely on the distillation algorithm to perform the co-analysis. In Section 6, we compare the results obtained by either bootstrapping with single-image segmentation or state-of-the-art co-segmentation techniques and demonstrate that all techniques benefit from the distillation algorithm.

We obtain one candidate segment from every image using the following procedure. First, we detect the bounding box of the main objects using the *objectness* detector [ADF12], keeping only the highest scoring box. Next, we use GrabCut [RKB04] to extract the detected object from the background. In our GrabCut implementation we use the response from a modern contour detector [DZ13] as the smoothness term. Note that all the ingredients mentioned above have publically available implementations, so this algorithm is easily reproducible.

5. Collection Distillation

At this point, we have a set of segmented candidate object shapes either obtained from co-segmentation or the algorithm described in the previous section. Some of the shapes are true contours of the query object but there are also many outliers and bad segmentations, as each of the previous stages (internet search, object detection, segmentation) can introduce outliers and errors.

5.1. Distance Measure

Our clustering algorithm requires a distance measure that is robust to slight articulation and deformations, so that objects captured from similar viewpoints induce small distances. We measure distances between outer contours. First, we normalize each contour by translating its center of mass to the origin and scale it so it has unit average distance to the origin.

Given two contours c_1 and c_2 we define their distance as

$$d(c_1,c_2) = d_G(c_1,c_2) + \lambda d_L(c_1,c_2) \qquad (1)$$

where the d_L and d_G terms incorporate local and global features, described below. $\lambda = 1.5$ is a balancing coefficient.

For the local term, d_L, we start by sampling and matching the contours using inner-distance shape context [LJ07]. This technique extends shape context [BMP02] by replacing Euclidean distances with inner-distances. It computes descriptors for each sampled point on the contours that are robust against articulation. We set d_L to the average spatial distance between a point on c_1 and its most similar point (in descriptor space) on c_2.

The global term, d_G, is defined as the sum of differences of two global attributes,

$$d_G(c_1,c_2) = \|\mathbf{w}_1 - \mathbf{w}_2\| + |a_1 - a_2|, \qquad (2)$$

where $\mathbf{w}_i$ are the principal directions of the contours (unit vectors computed using PCA) and a_i are the aspect ratios of their bounding boxes.

Hot Spots Identification

Now we are ready to describe the core of our distilling algorithm. When distilling a collection, we are not interested in *all* clusters but only the most significant ones, i.e., the *hot spots*, which likely contain contours capturing similar objects from similar viewpoints.

To extract the hot spots, we identify and select clusters that pass two tests: they are (1) *visually informative* (contain non-trivial shapes, see below), and (2) *tight* (small distance between all members). Just considering tightness is not sufficient, because clusters that contain *trivial* shapes (e.g., rectangular, circular, or, in general, low entropy shapes) are often tight, too; however, these shapes often result from segmentation errors. On the other hand, erroneous *non-trivial* shapes are unlikely to form tight clusters, because usually each has inconsistent defects. We describe the exact definition of these two tests in the following subsections.

We generate cluster candidates using the Mutual kNN technique [MHVL07]. This algorithm is commonly used when the goal is not necessarily to cluster all nodes but rather the most significant ones. The clusters are generated by extracting the connected components of the mutual kNN graph. This graph has a node for every contour, and an edge between two nodes if and only if both corresponding contours are part of each other's k-nearest neighbors according to the distance measure defined in Section 5.1. We set $k = 2$ to be conservative and only consider clusters that contain at least three members. This conservative choice of parameters may result in different clusters that contain similar shapes. Increasing the minimum cluster size threshold will yield more distinct clusters but will also result in a smaller distilled collection, as the smaller clusters will be thrown out. Varying k yields a similar trade-off. See Figure 2 for an illustration of the sparse kNN graph.

Visually Informative Clusters

We are interested in clusters composed of non-trivial shapes that are significant in their respective images. To quantify significance, we check if at least three cluster members pass two simple self-saliency tests [WSZ02]: the shape center of gravity is less than 25% away from the image center and the shape area is greater than 5% of the image area.

We are also interested in clusters that contain non-trivial shapes. To quantify this we check whether they are composed mostly of shapes that have low contour entropy or are nearly rectangular. If either condition is true, the cluster is discarded at this stage.

We first compute the contour entropy of each shape as described by Page et al. [PKS*03]: quantize the contour directions, compute the probability p_i of each curvature direction, and compute the entropy $-\sum_i p_i \log p_i$. Next, we define the entropy h of each cluster as the average entropy of its members, and compute the mean entropy μ_h and standard deviation σ_h of all clusters. Finally, we regard all clusters whose entropy is below $\mu_h - \sigma_h$ as containing trivial shapes and discard them.

We check whether a shape is near-rectangular by computing the ratio of its area and the area of its bounding box. We prune clusters whose average ratio exceeds 0.75. The reason for including this test in addition to the entropy check lies in the fact that many outlier segmentations are *near*-rectangular, but non-trivial entropy-wise.

Tight Clusters

We are only interested in tight clusters, i.e., those with small distances between their members (measured as the distance between the cluster's medoid and its most distant member). However, we observe that complex shapes are inherently more variable than simple ones. Thus, it is necessary to define an adaptive tightness threshold and prune low-entropy shapes more aggressively. We compute the mean tightness μ_t and standard deviation σ_t of all the visually informative clusters and only keep those whose tightness is below the threshold

$$T(h) = \begin{cases} (1-\alpha)\cdot(\mu_t - \sigma_t) + \alpha \cdot \mu_t, & h < \mu_h, \\ \mu_t, & else, \end{cases} \quad (3)$$

where h is the cluster entropy and $\alpha = \frac{h-(\mu_h-\sigma_h)}{\sigma_h}$ is the weight balancing coefficient. This equation describes a ramp that is clipped at the mean cluster tightness value (dashed line in Figure 3). For clusters that fail the test above, we remove the member farthest from the medoid and iterate the test until the threshold is satisfied or the cluster contains less than three members.

6. Evaluation

We performed quantitative and qualitative evaluations to analyze the performance of our distillation method. We conducted a quantitative evaluation on ground truth images and compared against three recent state-of-the-art co-segmentation methods [RJKL13, FI13, WHG13]. We performed the evaluation on three publicly available datasets provided by Rubinstein et al. [RJKL13] (Figure 4). A random subset of their images has manually annotated foreground segmentations, which are considered "ground truth". As in previous work, we report the precision P (percentage of correctly labelled pixels) and Jaccard similarity J (intersection over union of result and ground truth segmentations).

In addition to the internet datasets provided by [RJKL13], we generated twelve datasets containing diverse categories that span a wide range of objects (man-made to natural,

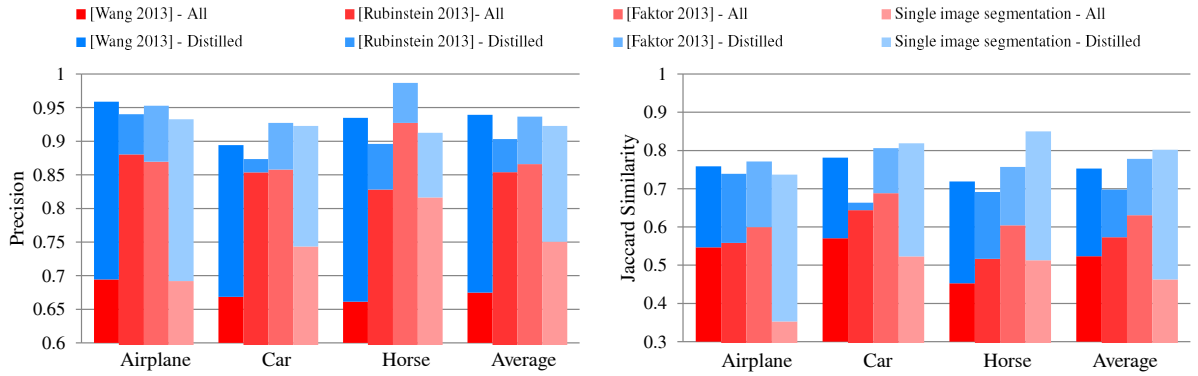

Figure 4: *Comparison against Rubinstein et al. [2013] , Wang et al. [2013] and Faktor et al. [2013] on the available datasets by reporting the precision P (left) and Jaccard similarity J (right). For each method, we report numbers for all images as well as for the distilled subsets by running our distilling method on each method's segmentation results. The distilled set size varies according to the segmentation method. Each plot shows a breakdown of the numbers for each method on the dataset followed by the average per-method numbers.*

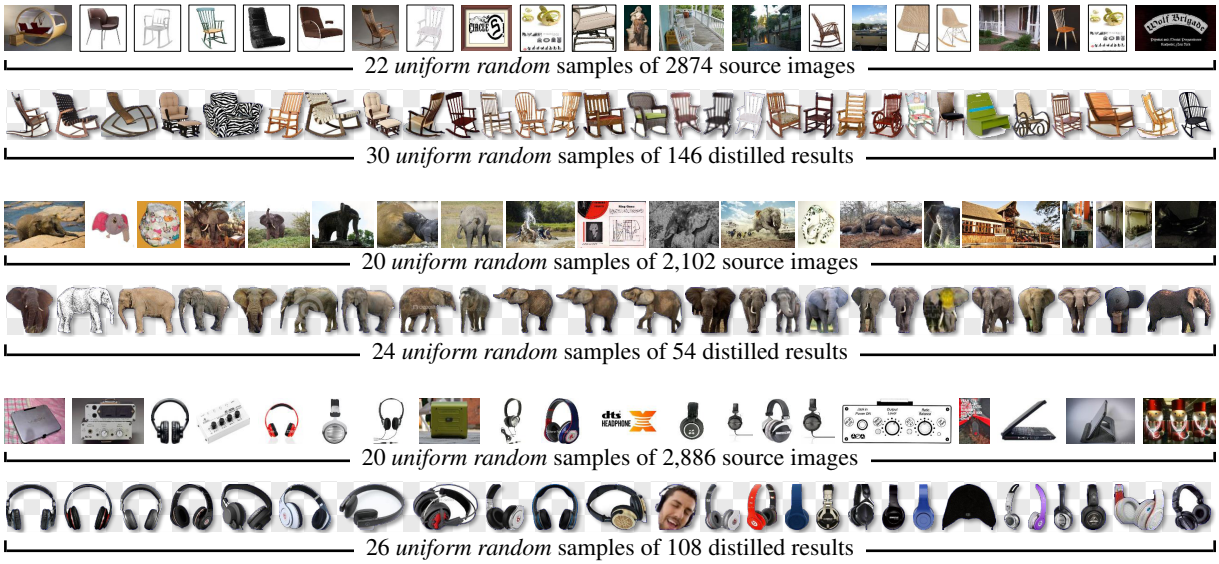

Figure 5: *Results for the queries "Rocking Chair", "Elephant", and "Headphone". For each set we show random samples from the input images and distilled results.*

small to large, many variants to almost all similar, etc.). Furthermore, we collected human labels for all the images belonging to the visually informative clusters in four of our sets to perform a quantitative evaluation on these additional sets. The labels were all manually inspected and refined. A qualitative evaluation on these datasets is provided as well.

We demonstrate how our technique yields a more reliable subset for three state-of-the-art co-segmentation methods. We further show that the distilled collection size depends on the initial segmentations, and that the method scales with the input collection size. Lastly, we examine the behaviour of our method as a function of the tightness threshold, and we show that a more conservative setting will yield a cleaner but smaller distilled collection. In what follows, we elaborate on each of these experiments.

Boosting co-segmentation scores

We performed our distillation technique on the segmentations obtained using both our standard single-image foreground extraction method and the segmentations obtained using the three mentioned co-segmentation techniques. As Figure 4 demonstrates, our distillation process consistently and dramatically improves the scores. These results further imply that the distillation process is not dependent on a specific segmentation technique, but rather can be added as a

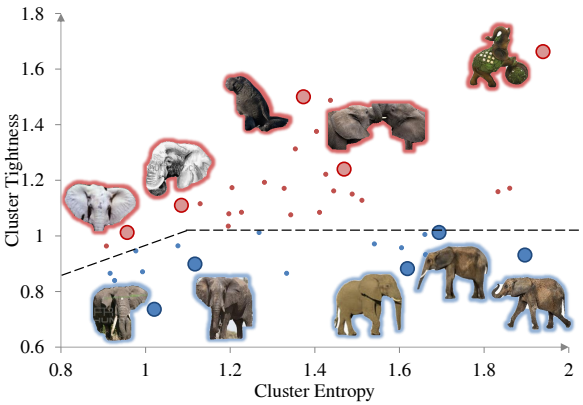

Figure 3: *Cluster entropy vs. tightness. The 41 dots represent all the visually informative clusters belonging to our Elephant set. For illustration purposes a few dots are enlarged and the cluster medoids are displayed alongside. The dashed line shows the adaptive tightness threshold. The tight clusters below the threshold (shown in blue) form our distilled collection, while the remaining clusters (red) are pruned.*

	Airplane	Car	Horse	Average
[Wang 2013]	3.69	5.78	9.7	6.39
[Rubinstein 2013]	11.59	6.49	8.68	8.92
[Faktor 2013]	7.49	12.35	11.7	10.51
Single image	3.47	1.19	1.66	2.11

Table 1: *Recall scores (in percentage) of the distilled collections on the available datasets. These scores complement the P and J scores that are reported in Figure 4 by the blue bars.*

post-process step to any co-segmentation pipeline, to extract a subset of highly confident inlier images. In Table 1, we report our recall numbers for different datasets and segmentation methods. As our method only extracts a subset of the inliers, the recall rates are generally low.

The number of distilled images depends on the segmentation technique. Starting from one of the co-segmentation techniques yields 4-8 times more distilled images. For example, starting from [RJKL13] yields 6.2 times more distilled images on average (the breakdown for the different datasets is 4.2, 8.1, 8.3 times more images). Therefore, if the extra complexity is acceptable, it is indeed desirable to use one of these techniques for initialization rather than the naive single-image segmentation approach.

Scaling to large distilled collections

We conducted experiments to analyze the scaling behaviour of our algorithm. Starting with the internet datasets provided by Rubinstein et al. [RJKL13], we successively reduced their size by removing random images. After each reduction step we ran our algorithm and examined the accuracy and number of distilled images as a function of the input size. We found that the number of distilled images is roughly proportional to the number of input images, while the two accuracy scores remain nearly constant (see Figure 6). This analysis suggests that a large distilled set could be obtained if one starts with a larger input collection of similar characteristics.

Additional datasets

Twelve datasets that span a wide range of objects were generated similarly to [RJKL13] to further evaluate our technique. In Figure 5 we show uniform random samples from the input collection and our distilled set. In addition, we provide thumbnails for the *full* distilled sets for all categories in the supplementary material. Please refer to these results for assessing the high quality of our results. Recall that we can increase the size and richness of our distilled sets by 4-8 times by starting from co-segmentation rather than single-image segmentation.

Size vs. quality of the distilled collections

To demonstrate the trade-off between the distilled set size and the obtained P and J scores, we collected ground-truth annotations for all the images belonging to the visually informative clusters in our Elephant, Full-Body, Hippo, and Rubber-Duck sets. We examined the aggressiveness of our algorithm by adding different constants, in the range $[-2\sigma_t, 2\sigma_t]$, to Equation 3. See Figure 7 for an illustration of the trade-off between the size and the scores. As the figure illustrates, both P and J scores drop mildly as the set size increases. We can tune the conservativeness of our algorithm in this manner.

7. Applications

We demonstrate our results in the context of several applications. The commonality between these applications is that they all rely on outlier-free input, and, so, only our distillation algorithm enables running them in an unsupervised manner. We first present a novel method for reconstructing an abstract 3D shape from the distilled contours (Section 7.1) and then show a variety of smaller 2D and 3D applications (Section 7.2).

7.1. Abstract 3D Modeling

The problem of reconstructing 3D shapes from images is gaining attention. To attempt this challenging problem, Vincente et al. [VCAB13] rely on selected images that all contain the object of interest and manually created segmentation masks and corresponding landmarks. In what follows, we present a fully-unsupervised approach that utilizes our distilled collections, where the user can optionally refine the result. Yet, all the results shown in the paper were generated automatically.

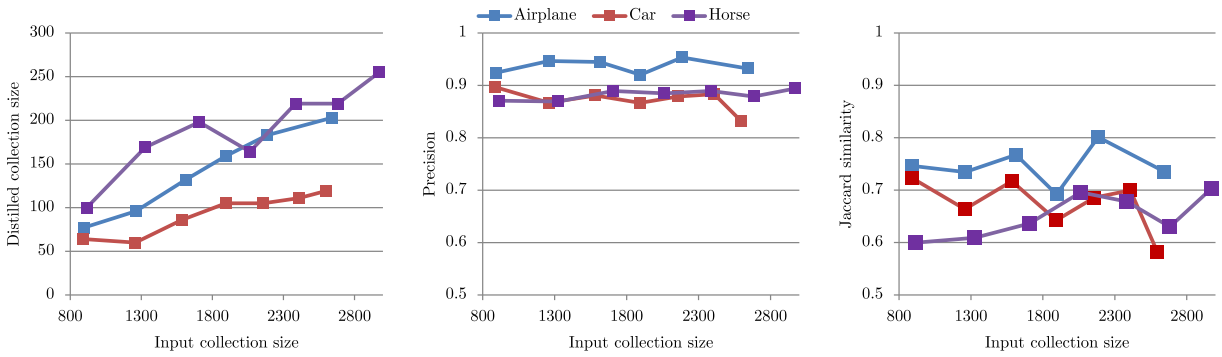

Figure 6: *Scaling evaluation. The figure demonstrates that the number of distilled images is proportional to the number of input images (left), while the P and J scores remain mostly constant (right). This analysis suggests that a large distilled set could be obtained from larger input collections of similar characteristics.*

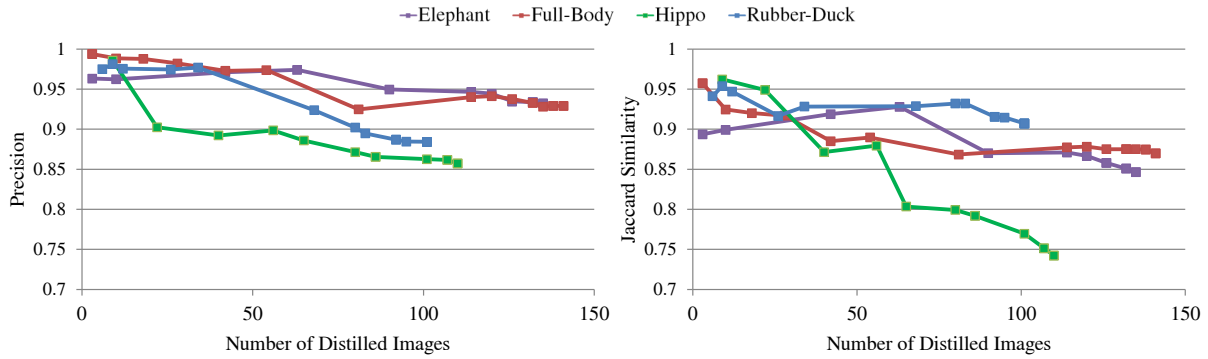

Figure 7: *Evaluating the trade-off between the set size and the P and J scores. Demonstrated above are the precision and Jaccard similarity scores obtained on four of our distilled collections, as different offsets are added to Equation 3. The data points correspond to successive increments of* $\frac{\sigma_i}{5}$*. As the figure demonstrates, a more conservative setting allows for a nearly perfect, but smaller, distilled collection.*

Shape-from-Silhouettes is a well-known approach for reconstructing a 3D model from multiple 2D contours. However, it is extremely difficult to generate 3D models from noisy internet collections or even from our distilled collections, because (1) the 2D shapes are the projections of different instances, (2) these instances might be articulated differently, and (3) the projection parameters used for each contour are unknown.

To alleviate these challenges, we assume that there are at least two orthogonal views among the distilled collection and then generate the 3D model from these views [RDI10]. However, identifying the two orthogonal views remains a non-trivial task [HO05]. We simplify this problem by assuming that our objects have at least one bilateral symmetry axis (i.e., reflection symmetry). This assumption is reasonable for nearly all natural and man-made objects [TB98]. Given a symmetric view and an orthogonal view we build a 3D shape by computing the visual hull surface [Lau94]. The smoothed surface yields an abstract 3D representation of the common object in the distilled images (Figure 8).

Since each distilled cluster contains similar shapes we only consider each cluster's medoid as a representative in the following computations. We first find the symmetric view by computing the contour distances between each shape and its mirror shapes along both vertical and horizontal axes and choosing the most symmetric one. We then consider all other shapes and select the farthest one as the orthogonal view. Since the direct contour distance measure in Equation 1 is unreliable for measuring large view changes we instead consider the furthest shape in *diffusion distance*. For more details on the diffusion distance, please refer to [CL06].

Since we cannot guarantee that the above method always finds good orthogonal views, a user could optionally assist and refine the process by selecting other views. See Figure 8 for some representative abstract 3D models along with the corresponding diffusion maps and selected orthogonal views.

Figure 8: *Reconstructing abstract 3D shapes from a distilled collection. We select two orthogonal views (middle) using the diffusion maps (left). The visual hull of the two yields an abstract 3D representation of the queried object.*

7.2. Other applications

In the following we present several smaller applications that are enabled by having distilled sets.

Viewpoint selection:

The problem of selecting good viewpoints for presenting a rendered 3D model has recently become an active area of research in computer graphics [FSG09, SLF*11], however, these methods are fairly complex. Our distilled segments can offer a simple alternative. We consider the silhouettes of the 3D object rendered from all directions and compare against our distilled countours. After filtering similar views we can sort the projections by their contour distance to determine the most representative views of the 3D model. Figure 9 illustrates the top three views for a given headphone model and the corresponding closest distilled images. Note that these three views reveal almost all part structures of the headphone.

Upright orientation detection:

This application is similar in spirit to the previous one. The majority of the top views are assumed to be captured with an upright orientation. This assumption is reasonable since images are usually associated with an upright orientation. Nevertheless, some objects, like airplanes in the sky, are not necessarily captured with an upright orientation.

Figure 9: *The selected three viewpoints for a 3D headphone model. The bottom row shows the selected three views while the above row shows the corresponding matched image for each view.*

Color design:

Finding harmoneous color themes is a difficult but crucial task in graphic design. There has been recent work that tries to extract color themes (usually consisting of 5-color palettes) from whole images [OAH11, LH13]. Using whole images does not lead to satisfactory results, however, if the goal is to design a color theme for an *object*, because background colors creep into the color theme, such as green grass color in Figure 10a. Using standard segmentation to extract colors only from the "foreground" in each image does not fully alleviate this problem because due to different errors invalid colors might still make it into the theme (e.g. the pink in Figure 10b). Extracting a color theme from our distilled set yields better results because they are clean, coherent, and avoid poor segmentations.

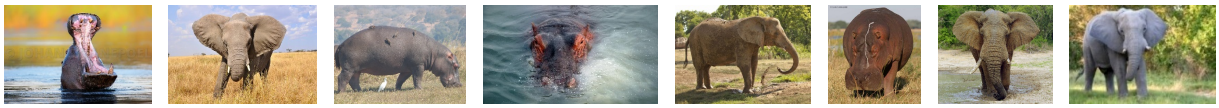

Figure 11: *Image-based Captcha. Subsets of our distilled Hippo and Elephant sets form a classification puzzle that is easy for humans and difficult for bots to solve, turning it into an effective automatic Captcha device.*

(a) Top Google whole images (b) Top Google segmented (c) Distilled

Figure 10: *Extracting color themes for the query "elephant": (a) extracting from whole images causes background colors (grass, sky) to dominate the color theme; (b) using foreground segmentation helps but there are still noise colors (pink) from outlier images; (c) our distilled sets lead to a clean color theme that does not contain any irrelevant colors.*

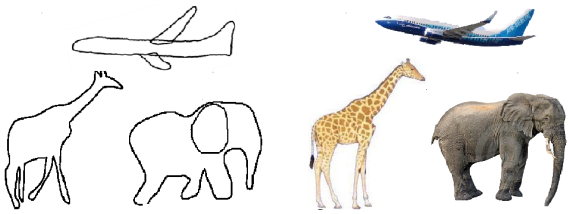

Figure 12: *An example of Sketch2Photo. Left: The input sketch, Right: the retrieved image candidates from our distilled collections.*

Captcha:

Our distilled collections can also serve as training data for various supervised applications, such as Captcha [VABHL03], image classification [CHV99], and object detection [VJ01]. Figure 11 demonstrates an example for an image-based Captcha, where a set of images form a puzzle and the task is to classify the set into two classes. Such task is extremely hard for an algorithm but easy for a human, turning it into an effective Turing test (i.e., a means to tell humans and robots apart). To generate such puzzles automatically the input set of images should be outlier-free.

Sketch2Photo:

A distilled collection can be used for sketch-based image retrieval in a sketch2photo application [CCT*09]. Given a simple freehand sketch, we obtain the distilled shape with the smallest contour distance to the drawn sketch. Figure 12 provides a few examples, where sketches of an elephant, a giraffe and an airplane are matched to reasonable images. Using a distilled set dramatically improves the chance that the retrieved images are cleanly segmented inlier images.

8. Discussion, Limitation, and Future Work

We have introduced the novel notion of a distilled image set, which may originate from a source set that is highly unstructured, noisy, and outlier-ridden. We have developed a distillation algorithm that is applicable to a large, raw collection of internet images returned from an object query. Our approach is unsupervised, built on a novel clustering scheme, and returns a structured set of inliers — the distilled set. This implicitly suggests that one can potentially learn the essence of an object from vast amount of raw image query results without any object-specific knowledge or semantic analysis. We have shown, through several examples, the applicative potential of distilled image collections.

Our approach relies on the idea that outlier shapes are random in nature, and therefore, outliers do not tightly couple and persist into our distilled set. While mostly true, this realization is not impermeable in the sense that sometimes certain outlier shapes do appear multiple times in such a dense collection. For example, in our Motorcycle set, there were quite a few images capturing a motorcycle vest, a semantically-related object but certainly not a motorcycle. Consequently, our distilled set contains a cluster of images capturing different motorcycle vests. We have tried to show that with a more conservative setting, we can somewhat overcome this limitation. However, it comes at the expense of a smaller distilled collection.

Having said that, we should stress that there is yet more latent information in the clusters of a distiled set, i.e., knowledge about view directions, which we have not utilized. After forming the clusters, we may consider the inter-relations among the clusters, assuming that they are related to view direction, to re-enforce inlier clusters. Considering such relations among clusters, we may learn that a motorcycle vest is a relevant object, but not a motorcycle.

An important and intriguing question that we encountered is "what is an interesting shape". In Section 3, we have developed a criterion to define when a contour is interesting. Although this criterion usually removes the outlier shapes, it can also remove a significant portion of inliers, especially if the object has a trivial shape. If the queried object is a match box, for example, most shapes will be removed due to their low entropy or nearly rectangular shapes. We believe that this is an interesting research problem in its own right. There has been much research on "saliency", especially in the context of image analysis. Here we ask a related question, while posing it on shapes or geometry in general.

We plan to further extend the value of our distilled collection. We would like to enhance our 3D modeling technique and ultimately allow for a *Shape-from-Text* operator, a method to convert a noun into a 3D shape. Our vision is to do this without any supervision and starting from just a textual image query. In our 3D modeling application, we learned that in many cases, finding two orthogonal views provides a strong basis towards recovering an abstract 3D shape. As a by-product of this work, we developed means to identify such orthogonal views assuming the object is bilaterally symmetric. Analyzing the view directions of an arbitrary shape is a problem we would like to investigate more. In general, we believe that now, with the ever increasing number of internet images, unsupervised learning tasks of a large collection of images belonging to some family will gain much more attention.

Acknowledgements

This work supported by the Israel Science Foundation, the NSFC Grant (Number 61202222) and the NSERC Grant (Number 293127).

References

[ADF10] ALEXE B., DESELAERS T., FERRARI V.: Classcut for unsupervised class segmentation. In *Proc. Euro. Conf. on Comp. Vis.* (2010), Springer, pp. 380–393. 3

[ADF12] ALEXE B., DESELAERS T., FERRARI V.: Measuring the objectness of image windows. *IEEE Trans. Pat. Ana. & Mach. Int. 34*, 11 (2012), 2189–2202. 4

[BDF*03] BARNARD K., DUYGULU P., FORSYTH D., DE FREITAS N., BLEI D. M., JORDAN M. I.: Matching words and pictures. *The Journal of Machine Learning Research 3* (2003), 1107–1135. 3

[BMP02] BELONGIE S., MALIK J., PUZICHA J.: Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 4 (2002), 509–522. 4

[CCMV07] CARNEIRO G., CHAN A. B., MORENO P. J., VASCONCELOS N.: Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*, 3 (2007), 394–410. 3

[CCT*09] CHEN T., CHENG M.-M., TAN P., SHAMIR A., HU S.-M.: Sketch2photo: internet image montage. *ACM Trans. Graph. (SIGGRAPH Asia) 28*, 5 (2009), 124. 1, 3, 10

[CFF07] CAO L., FEI-FEI L.: Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. Int. Conf. on Comp. Vis.* (2007), IEEE, pp. 1–8. 3

[CHV99] CHAPELLE O., HAFFNER P., VAPNIK V. N.: Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on 10*, 5 (1999), 1055–1064. 10

[CL06] COIFMAN R. R., LAFON S.: Diffusion maps. *Applied and computational harmonic analysis 21*, 1 (2006), 5–30. 8

[CSG14] CHEN X., SHRIVASTAVA A., GUPTA A.: Enriching visual knowledge bases via object discovery and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (March 2014). 3

[CW04] CHEN Y., WANG J. Z.: Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research 5* (2004), 913–939. 3

[DBdFF02] DUYGULU P., BARNARD K., DE FREITAS J. F., FORSYTH D. A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer VisionąłECCV 2002* (2002), Springer, pp. 97–112.

[DZ13] DOLLÁR P., ZITNICK C. L.: Structured forests for fast edge detection. In *Proc. Int. Conf. on Comp. Vis.* (2013), IEEE. 4

[EHBA09] EITZ M., HILDEBRAND K., BOUBEKEUR T., ALEXA M.: Photosketch: a sketch based image query and compositing system. In *ACM SIGGRAPH - Talk Program* (2009). 1

[FI13] FAKTOR A., IRANI M.: Co-segmentation by composition. In *Proc. Int. Conf. on Comp. Vis.* (2013), IEEE, pp. 1297–1304. 5

[FSG09] FEIXAS M., SBERT M., GONZÁLEZ F.: A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Transactions on Applied Perception (TAP) 6*, 1 (2009), 1. 9

[HO05] HALL P. M., OWEN M.: Simple canonical views. In *BMVC* (2005). 8

[HS09] HOCHBAUM D. S., SINGH V.: An efficient algorithm for co-segmentation. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2009), IEEE, pp. 269–276. 3

[JBP12] JOULIN A., BACH F., PONCE J.: Multi-class cosegmentation. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2012), IEEE, pp. 542–549. 3

[JLM03] JEON J., LAVRENKO V., MANMATHA R.: Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM, pp. 119–126. 3

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia* (2014), ACM, pp. 675–678. 3

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105. 3

[KX12] KIM G., XING E. P.: On multiple foreground cosegmentation. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2012), IEEE, pp. 837–844. 3

[Lau94] LAURENTINI A.: The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pat. Ana. & Mach. Int. 16*, 2 (1994), 150–162. 8

[LFF07] LI L.-J., FEI-FEI L.: What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (2007), IEEE, pp. 1–8. 3

[LG09] LEE Y. J., GRAUMAN K.: Shape discovery from unlabeled image collections. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2009), IEEE, pp. 2254–2261. 3

[LH13] LIN S., HANRAHAN P.: Modeling how people extract color themes from images. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems* (2013), ACM, pp. 3101–3110. 9

[LJ07] LING H., JACOBS D. W.: Shape classification using the inner-distance. *IEEE Trans. Pat. Ana. & Mach. Int. 29*, 2 (2007), 286–299. 4

[LSFF09] LI L.-J., SOCHER R., FEI-FEI L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 2036–2043. 3

[LWQ*08] LIU X., WAN L., QU Y., WONG T.-T., LIN S., LEUNG C.-S., HENG P.-A.: Intrinsic colorization. *ACM Trans. Graph. (SIGGRAPH Asia) 27*, 5 (2008), 152. 1

[MHVL07] MAIER M., HEIN M., VON LUXBURG U.: Cluster identification in nearest-neighbor graphs. In *Algorithmic Learning Theory* (2007), Springer, pp. 196–210. 5

[MSD09] MUKHERJEE L., SINGH V., DYER C. R.: Half-integrality based algorithms for cosegmentation of images. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2009), IEEE, pp. 2028–2035. 3

[OAH11] O'DONOVAN P., AGARWALA A., HERTZMANN A.: Color compatibility from large datasets. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 63. 9

[OT01] OLIVA A., TORRALBA A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision 42*, 3 (2001), 145–175. 3

[PKS*03] PAGE D. L., KOSCHAN A., SUKUMAR S. R., ROUI-ABIDI B., ABIDI M. A.: Shape analysis algorithm based on information theory. In *Proc. IEEE Conf. on Image Processing* (2003), vol. 1, pp. 229–232. 5

[PT10] PAYET N., TODOROVIC S.: From a set of shapes to object discovery. In *Proc. Euro. Conf. on Comp. Vis.* (2010), Springer, pp. 57–70. 3

[RDI10] RIVERS A., DURAND F., IGARASHI T.: 3d modeling with silhouettes. *ACM Trans. Graph. (SIGGRAPH Asia) 29*, 4 (2010). 8

[RJKL13] RUBINSTEIN M., JOULIN A., KOPF J., LIU C.: Unsupervised joint object discovery and segmentation in internet images. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2013), IEEE, pp. 1939–1946. 2, 3, 4, 5, 7

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (SIGGRAPH) 23*, 3 (2004), 309–314. 4

[RLYFF12] RUSSAKOVSKY O., LIN Y., YU K., FEI-FEI L.: Object-centric spatial pooling for image classification. In *Computer Vision–ECCV 2012* (2012), Springer, pp. 1–15. 3

[RMBK06] ROTHER C., MINKA T., BLAKE A., KOLMOGOROV V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2006), vol. 1, pp. 993–1000. 3

[SLF*11] SECORD A., LU J., FINKELSTEIN A., SINGH M., NEALEN A.: Perceptual models of viewpoint preference. *ACM Transactions on Graphics (TOG) 30*, 5 (2011), 109. 9

[SSS06] SNAVELY N., SEITZ S. M., SZELISKI R.: Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph. (SIGGRAPH) 25*, 3 (2006), 835–846. 1

[SVZ13] SIMONYAN K., VEDALDI A., ZISSERMAN A.: Deep fisher networks for large-scale image classification. In *Advances in neural information processing systems* (2013), pp. 163–171. 3

[TB98] TROJE N. F., BÜLTHOFF H. H.: How is bilateral symmetry of human faces used for recognition of novel views? *Vision Research 38*, 1 (1998), 79–89. 8

[TLBB10] TUYTELAARS T., LAMPERT C. H., BLASCHKO M. B., BUNTINE W.: Unsupervised object discovery: A comparison. *Int. J. Comp. Vis. 88*, 2 (2010), 284–302. 3

[VABHL03] VON AHN L., BLUM M., HOPPER N. J., LANGFORD J.: Captcha: Using hard ai problems for security. In *Advances in Cryptology?aEUROCRYPT 2003* (2003), Springer, pp. 294–311. 10

[VCAB13] VICENTE S., CARREIRA J., AGAPITO L., BATISTA J.: Reconstructing pascal voc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 41–48. 7

[VFJZ01] VAILAYA A., FIGUEIREDO M. A., JAIN A. K., ZHANG H.-J.: Image classification for content-based indexing. *Image Processing, IEEE Transactions on 10*, 1 (2001), 117–130. 3

[VJ01] VIOLA P., JONES M.: Robust real-time object detection. *International Journal of Computer Vision 4* (2001), 34–47. 10

[VRK11] VICENTE S., ROTHER C., KOLMOGOROV V.: Object cosegmentation. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.* (2011), IEEE, pp. 2217–2224. 3

[WBL09] WANG C., BLEI D., LI F.-F.: Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 1903–1910. 3

[WBU10] WESTON J., BENGIO S., USUNIER N.: Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning 81*, 1 (2010), 21–35. 3

[WHG13] WANG F., HUANG Q., GUIBAS L. J.: Image cosegmentation via consistent functional maps. In *Proc. Int. Conf. on Comp. Vis.* (2013), IEEE, pp. 849–856. 5

[WSZ02] WANG W., SONG Y., ZHANG A.: Semantics-based image retrieval by region saliency. In *Image and Video Retrieval* (2002), Springer, pp. 29–37. 5

[YDH06] YANG C., DONG M., HUA J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, IEEE, pp. 2057–2063. 3

[ZGW*13] ZHANG C., GAO J., WANG O., GEORGEL P., YANG R., DAVIS J., FRAHM J., POLLEFEYS M.: Personal photo enhancement using internet photo collections. *IEEE Trans. Vis. & Comp. Graphics* (2013). 1

[ZZ06] ZHOU Z.-H., ZHANG M.-L.: Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems* (2006), pp. 1609–1616. 3