Int J Comput Vis (2014) 108:82–96
DOI 10.1007/s11263-014-0716-6

Part and Attribute Discovery from Relative Annotations

Subhransu Maji · Gregory Shakhnarovich

Received: 25 February 2013 / Accepted: 14 March 2014 / Published online: 26 April 2014

Abstract Part and attribute based representations are widely used to support high-level search and retrieval applications. However, learning computer vision models for automatically extracting these from images requires significant effort in the form of part and attribute labels and annotations. We propose an annotation framework based on comparisons between pairs of instances within a set, which aims to reduce the overhead in manually specifying the set of part and attribute labels. Our comparisons are based on intuitive properties such as *correspondences* and *differences*, which are applicable to a wide range of categories. Moreover, they require few category specific instructions and lead to simple annotation interfaces compared to traditional approaches. On a number of visual categories we show that our framework can use noisy annotations collected via "crowdsourcing" to discover semantic parts useful for detection and parsing, as well as attributes suitable for fine-grained recognition.

Keywords Relative annotations · Crowdsourcing · Semantic parts · Fine-grained attributes

1 Introduction

In order for an automatic system to answer queries such as 'birds with short beaks and blue wings' or 'planes with engines on their nose', it would require an underlying representation that is *aligned* to the parts and attributes of the category in question. In recent years several such part and attribute based models have demonstrated excellent performance on a number of visual recognition tasks such as detection (Bourdev and Malik 2009; Bourdev et al. 2010), pose estimation (Agarwal and Triggs 2006; Felzenszwalb and Huttenlocher 2005; Ferrari et al. 2008), detailed recognition (Bourdev et al. 2011; Farhadi et al. 2010; Kumar et al. 2008), interactive categorization (Branson et al. 2010; Kovashka et al. 2012), etc. Most of these models rely on supervision in the form of a pre-defined set of parts and attributes provided by experts. In stark contrast, little attention has been paid to automatically discovering the set of parts and attributes useful for these high-level recognition tasks.

For some categories the set of part and attribute labels are easy to obtain—parts may be based on the anatomical structure for animals, attributes of birds may be obtained from a field guide. For these categories traditional methods for collecting annotations involve showing a *single* instance at a time with detailed instructions (Fig. 1 *left*). For part annotation he/she may mark the bounding boxes of parts or locations of landmarks. Similarly, they may indicate the presence or absence of a *given* attribute in each image. However, for a vast majority of categories such structure is absent or field guides are nonexistent, rendering the task of determining the set of labels to annotate to be a challenge. Furthermore, attributes present in field guides may not be suitable for the non-expert 'crowd' available via crowdsourcing platforms. Annotators may find it difficult to answer questions such as 'where is the elbow of a horse' or 'what color is the supercilium of a bird'. Some parts may be hard to localize in images due to self occlusion, e.g. 'where is the tail of a cat'.

Our framework for part and attribute annotation addresses some of these drawbacks. The key idea, as seen in Fig. 1 (*right*), is that we annotate properties of an object *relative* to another. As seen in Fig. 2, we rely on intuitive properties based on *correspondences* and *differences* between pairs

Communicated by Serge Belongie and Kristen Grauman.

S. Maji (✉) · G. Shakhnarovich
Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave, Chicago, IL 60637, USA
e-mail: smaji@ttic.edu

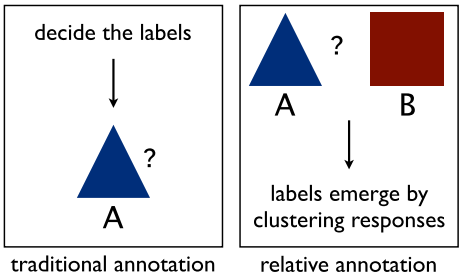


Fig. 1 Our relative annotation framework for label discovery. In contrast to commonly used annotation frameworks when labels are known, our approach consists of collecting *relative* annotations followed by *grouping* the instances to discover the labels implicitly defined by the groups

of instances. By analyzing these annotations across many such pairs, one can discover groups that correspond to parts and attributes respectively. Furthermore, as we demonstrate experimentally, these can be used to bootstrap a number of visual recognition tasks such as object detection via parts, or fine-grained attribute prediction.

In summary, we propose new annotations tools along with their associated clustering methods to discover parts and attributes of visual categories from annotations that can be collected via crowdsourcing with overhead. Such weakly structured annotations can be noisy, and much of our work aims to reduce this with a careful design of the user interface for collecting annotations and the method to analyze the collected data. Experimentally we show that semantically meaningful parts that are useful for recognition tasks such as detection and fine-grained parsing, as well as attributes useful for fine-grained discrimination, can be discovered for a number of visual categories such as buildings, airplanes, birds and texture patterns. This paper provides a unified view of our earlier work (Maji 2012; Maji and Shakhanarovich 2013, 2012) as well as some additional experiments on discovering and predicting fine-grained attributes of man-made textures.

1.1 Related Work

Relative or comparative information has been widely used for metric learning where user preferences of similarity are obtained over triplets of images (Frome et al. 2007; Tamuz et al. 2011). Our work is related to recent work in computer vision and human-computer interaction for recognition tasks and image annotation using humans 'in the loop'. These include games for annotating images such as ESP (Von Ahn and Dabbish 2004), PeekABoom (Von Ahn et al. 2006), as well as interactive methods for fine-grained recognition (Welinder et al. 2010). Below we describe some of the relevant work on semantic part and attribute discovery.

1.1.1 Semantic Part Annotations and Discovery

A large number of approaches for part discovery in computer vision are weakly supervised (Felzenszwalb et al. 2010; Felzenszwalb and Huttenlocher 2005; Singh et al. 2012; Weber et al. 2000), i.e., they rely on object level annotations only. However, the semantic alignment of the discovered parts are either nonexistent or unknown, which makes them less suitable for answering detailed questions such as 'is the person wearing a hat?', etc. In this work we focus on semantic parts learned or discovered using supervision in the form of part annotations.

Popular methods for part annotations typically involve drawing part bounding boxes or marking a predefined set of landmarks on instances. For bounding box annotations, annotators are typically asked to draw a tight bounding box around the part of interest. This is the staple mode of annotation for rigid parts and objects such as frontal faces and pedestrians in many datasets (Dalal and Triggs 2005; Everingham et al. 2010). More recently datasets such as Farhadi et al. (2010) also contain bounding boxes for parts of animals such as heads and legs, and parts of vehicles such as wheels.

When the extent of the part is less obvious, marking keypoints or landmarks can be more suitable. Here the annotators

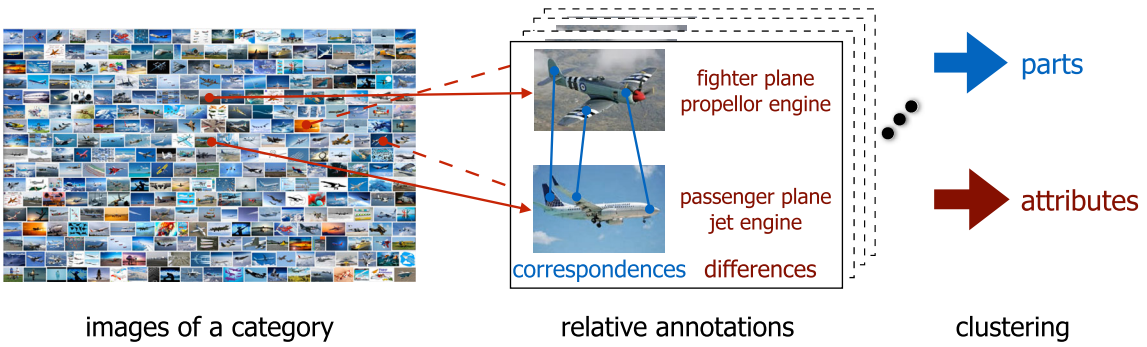


Fig. 2 Overview of the relative annotation framework for part and attribute discovery. Given a collection of images we pick random pairs and collect correspondences (via clicks) and differences (via text) between them on Amazon's mechanical turk. These annotations are then clustered across various pairs to obtain parts and attributes respectively

are asked to mark the location and/or presence of a predefined set of keypoints or landmarks in each instance of the object. These annotations can then be used to discover and learn part detectors that are aligned to these annotations. A notable example of this is the 'poselet' model (Bourdev et al. 2010; Bourdev and Malik 2009) that rely on a set of 10–20 keypoints per category, to learn a large library of discriminative patterns by finding repeatable and detectable configurations of these keypoints. Other examples include supervised deformable part-models (Yang and Ramanan 2011; Zhu and Ramanan 2012) and 'phraselets' (Desai and Ramanan 2012).

The main drawback of these approaches is that they require the set of parts or landmarks be known ahead of time. Constructing such a set with the detailed instructions for annotation can be time consuming. Furthermore, to account for all variations in a structurally diverse category, such as buildings, the set has to be very large making the annotation task cumbersome. These pose significant challenges on both constructing the user interfaces for and reliably collecting annotations via crowdsourcing.

1.1.2 Semantic Attribute Annotation and Discovery

Much of recent work on attribute based learning and description has relied on a pre-defined set of attributes specified by experts, e.g. field guides. Automatic methods for attribute discovery can be broadly divided into two categories, those that rely on (1) images with captions, and (2) a specialized annotation task.

The work of Berg et al. (2010) lies in the former category where they use descriptions of products such as shoes, bags, jewelry, etc., collected from the web to mine phrases that appear frequently, which are analyzed to characterize and predict the visually discriminative attributes. The main drawback of such work is that such text is available for only few categories. Collecting descriptions via crowdsourcing is another option, but without quality control or detailed instructions, these captions may not be descriptive enough to mine fine-grained attributes.

An example of the latter is Duan et al. (2012), Parikh and Grauman (2011) where they discover task-specific attributes with humans 'in the loop' by considering projections of the data asking them to *name* the direction of variability. However, it assumes a feature space where describable directions can be easily found. Another related work is Patterson and Hays (2012) where they ask annotators to describe attributes (single words) that distinguish a set of images from another as a way of identifying discriminative attributes. This procedure was used to identify attributes for scene understanding. Although quite suitable for scenes, single words fail to describe localized attributes such as 'pointy beak' or 'engine on the nose', which might be more relevant for object categories.

2 Overview

Relative information about similarities and differences can be used to discover labels by *grouping* instances. Consider the following analogy; Suppose we want to label a set of points into k categories. If we know the categories we can simply label each instance as one of k. However, if we don't, we can collect similarities between pairs of instances, and use them to cluster the points into k groups. This enables simultaneous discovery of the categories and implicit labeling of the instances.

In this work we extend this analogy for discovering parts and attributes. Given a collection of images for we wish to discover parts and attributes, we randomly sample pairs for which we collect relative annotations. As seen in Fig. 2, our framework has two main ingredients, (a) the *user interface* to collect annotations and, (b) the *grouping method* to discover the clusters of instances. Both the part and attribute discovery framework follow the same overall idea, but the details vary, and are described below.

In Sect. 3 we describe our framework for semantic part discovery. We consider diverse visual categories such as buildings and chairs for which it is rather difficult to come up with a list of parts ahead of time—some of these parts are hard to name, others don't necessary correspond to a part (e.g. the middle point of the roof-line), and some others might have missed our attention. We propose a *semantic correspondence task* where annotators mark pairs of landmarks that belong to the same semantic part. Landmarks are then clustered using their appearance to discover semantic parts that can then be used for a variety of computer vision applications such as detection, semantic saliency prediction, and detailed parsing.

In Sect. 4 we describe our framework for fine-grained attribute discovery. Here we propose a *discriminative description task*, where annotators are asked to describe the differences between pairs of instances within a basic level category. The task forces the annotators to describe each instance in more detail than they would when each instance is shown in isolation. These descriptions are also highly structured which enables us to group words into clusters based on their co-occurrence statistics. We show how one can discover describable attributes for a number of categories such as airplanes, birds and man-made textures. Furthermore, the inferred attributes can be used to learn visual classifiers to predict attributes of unseen instances. We conclude and present directions of future work in Sect. 5.

A drawback of the approach is that the cost scales quadratically with the number of instances. However, one can simply compare each instance to a fixed number of others to reduce the cost. Our experiments suggest that even with a small number of such comparisons per instance (typically <10) the obtained annotations can be useful in a number of recognition tasks.

3 Semantic Part Discovery

The goal of this work is to discover the inherent structure of objects within a category via their parts. Such parts act as *diagnostic* elements for instances within the category; their presence and arrangement provides rich information regarding the presence and location of the object, its pose, size and fine-grained properties, e.g., architectural style of a building or type of a car. We consider two challenging man-made categories, churches and chairs, for which traditional methods for collecting part annotation fail for a number of reasons—presence or absence of parts, or the number of their appearances, varies across instances. The instances of these parts could differ drastically in their appearance, e.g., shape of windows for buildings, form of the armrests for chairs. Still, despite this structural flexibility and appearance variability, humans can reliably recognize corresponding points across instances, even when the observer does not have a name for the part and does not precisely know its function.

We leverage this ability through in our annotation paradigm that relies on people marking such correspondences, and propose a novel approach to construction of a library of parts driven by such annotations. Such annotations can enable discovery of parts that are aligned to human-semantics for categories that are otherwise hard to annotate using traditional methods of named keypoints, and part bounding boxes. We show the utility of the rich part library learned in this way for three tasks: object detection, category-specific saliency estimation, and fine-grained parsing.

3.1 User Interface

The annotator is presented with a pair of images of the category of interest, and asked to mark points in the two images that match. The interface (Fig. 3 top) allows the user to add correspondences by first clicking on the left image and then on the corresponding point in the right image. Each image has only a single instance of the object. To guide the process of annotation we show some examples of landmarks for the category of interest such as those in Fig. 3 (bottom). The interface lets the user adjust the landmark pairs or delete them. Once the user is done, he/she clicks a submit button to finish the annotation. Providing instructions for this task is significantly easier than providing the names and semantics of a pre-defined set of parts.

Dataset: We have collected annotations for 1,000 pairs among 288 images of churches, and 300 images of chairs, collected from Flickr. Annotations were collected on Amazon's mechanical turk (AMT). Landmark pairs, a few examples of which are shown in Fig. 4, include a variety of semantic matches: identical structural elements of buildings (windows, spires, corners and gables), and vaguely defined yet consistent matches, the likes of 'the mid-point of roof slope' and 'the mid-point of back rest' for chairs.

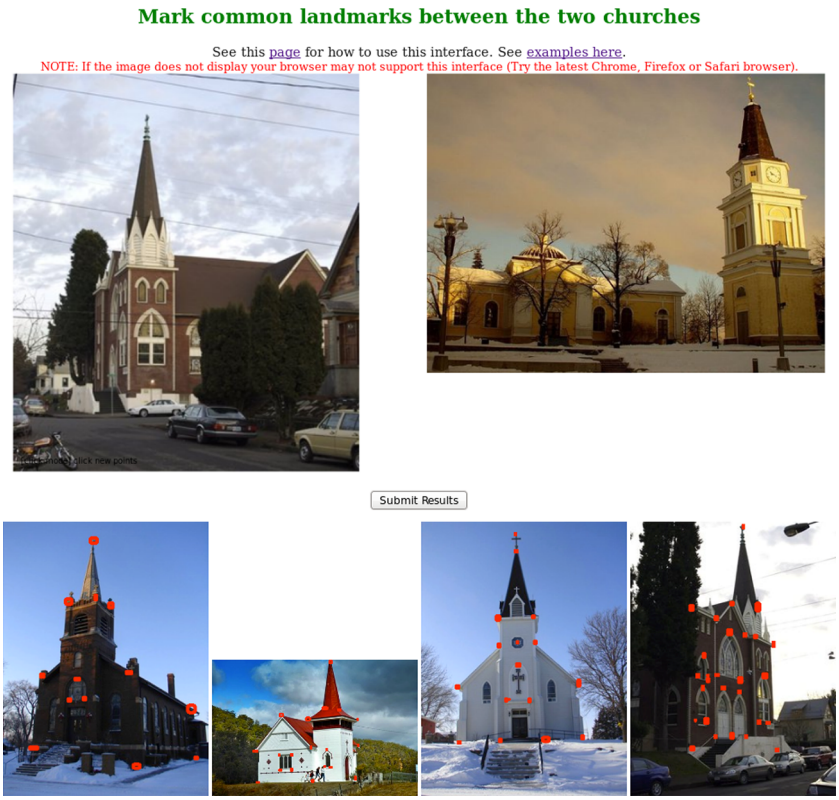


Fig. 3 User interface for labeling correspondences. The annotator is shown pairs of images and he/she mark correspondences by clicking on a point in the first image and its corresponding point in the second image. Examples of landmarks shown to the users to guide the annotation process

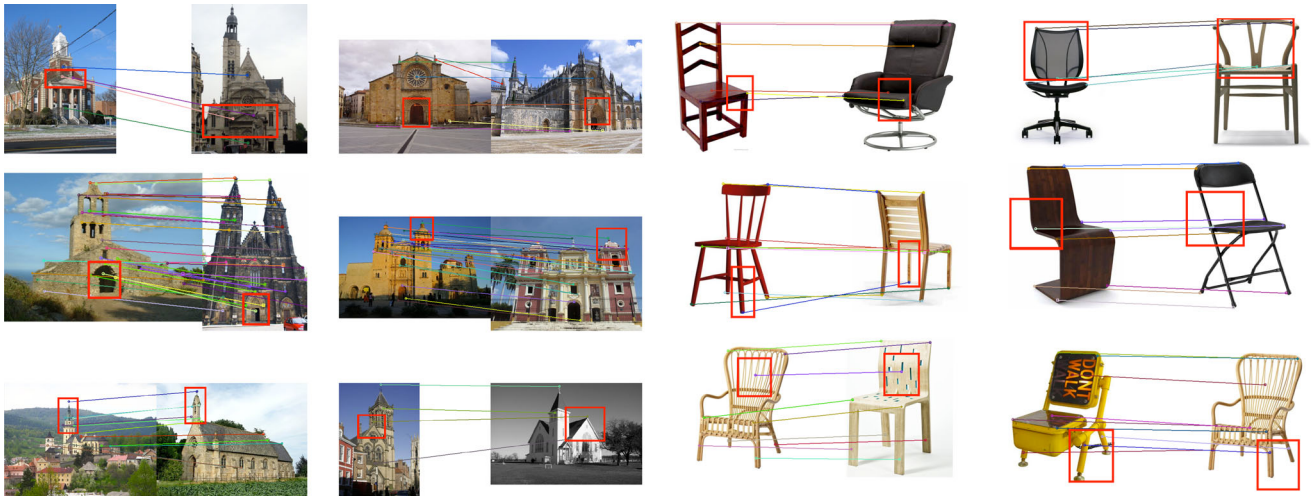


Fig. 4 Example correspondence annotations. Some annotations collected using our interface for *church* (*left*) and *chair* (*right*) categories. Given a source window in one of the images in a pair, these correspondences allow us to *automatically* find the target window in the other (shown as *red boxes*) (Color figure online)

Fig. 5 Consistency of annotations. Landmarks marked by various users shown with different color in each image. The locations of landmarks provided by different annotators tend to agree at salient locations on the image

Are the annotations consistent? Our hypothesis is that annotators can mark correspondences without knowing the names of the parts. In Fig. 5 we visualize where different annotators click on an image as it is displayed with other images. As one can see there is a high level of consistency across annotators—points near the tip of the towers, corners of windows, etc., are consistently clicked by different annotators. Moreover, the number of clicks at a given location provides a rough estimate of the frequency of a part within the category.

Cost of annotation For the church category, users spend 48 seconds and marked 3 landmarks on average per image (i.e., 48/6 = 8s per landmark). For chairs, users spend 34 s and marked 2 landmarks on average, (i.e., 34/4 = 8.5s per landmark). Note that using our interface we get annotations for two images at the same time, hence double the number of landmarks. As a comparison collecting keypoint annotations using the interface of Maji (2011) takes 44 seconds and users mark 6 keypoints on average for the chair category, (i.e., 7.3s per landmark). Thus, our interface is similar in terms of the time spend by the annotators.

3.2 Part Discovery by Clustering Appearance

The partial correspondence information between pairs can be propagated over the set of images using the underlying semantic graph of correspondences $G = (V, E)$. The vertex set V corresponds to images, while the edges E correspond to the collected pairwise correspondences. Correspondences can be propagated in the semantic graph from one image to another as long as there is a path connecting the two. In this manner one can obtain a large number of potential landmarks across images that are in *correspondence* with a given landmark in an image as shown in Fig. 6.

We can then use an *appearance* model to group these landmarks and obtain semantic parts. We use HOG features (Dalal and Triggs 2005) to model part appearance. Given a sampled 'seed', we initialize the model by training the HOG filter $\mathbf{w}^{(0)}$ to separate the seed patch from a set of background patches; this step resembles the exemplar-SVM of Malisiewicz et al. (2011). Next, we propagate the correspondence from the seed window using breadth-first search in the semantic graph as shown in Fig. 6. This provides a set of hypothesized locations

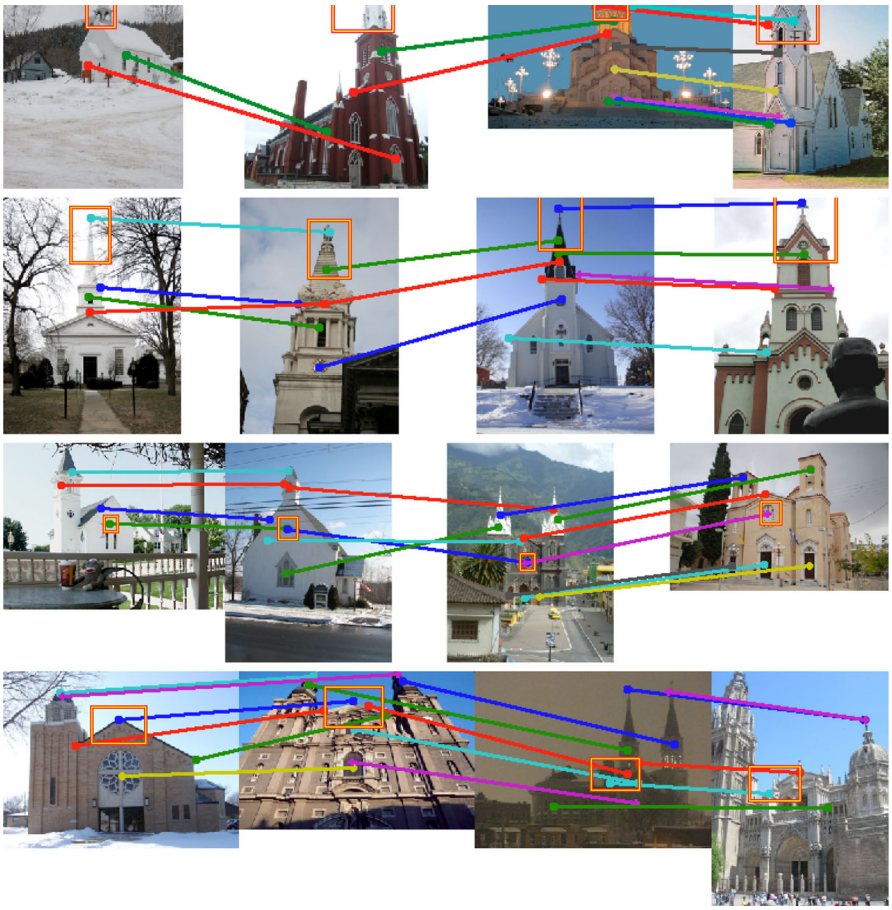


Fig. 6 Depth-first correspondence propagation in the semantic graph. Images on the *left column* in each row are not directly connected to the right ones, but the correspondences can be propagated in a pairwise manner to them

for the part in other images. We denote them $\mathbf{x}(I_i, L_i^{(0)}, s_i^{(0)})$, for $i = 1, \ldots, k$, where $\mathbf{x}(I, L, s)$ is the patch extracted from image I at location L and with scale s. We would like to use these additional likely examples of the part to retrain the model.

Since the correspondence is sparse, the estimated location and scale of these initial hypothesized matches is likely to be noisy. Furthermore, some of these matches may belong to a different visual sub-type of the part, e.g., a different kind of window or door. Therefore we treat the unknown location and scale of the matches as latent variables, and train the model using the following iterative algorithm. In iteration t, we find for each hypothesized match the location and scale *near* the initial estimate obtained using the semantic graph that maximizes the response of $\mathbf{w}^{(t-1)}$:

$$\left(L_i^{(t)}, s_i^{(t)}\right) = \underset{L,s \in \mathcal{N}(L_i^{(0)}, s_i^{(0)})}{\operatorname{argmax}} \langle \mathbf{w}^{(t-1)}, \mathbf{x}(I_i, L, s) \rangle \qquad (1)$$

where, $\mathcal{N}(L, s)$ denotes all the locations and scales for which the corresponding rectangles have an overlap (defined as the intersection over union of areas) greater than $\tau = 0.5$ with the rectangle at (L, s). Then, we retrain $\mathbf{w}^{(t)}$ using the updated list of matches $\mathbf{x}(I_i, L_i^{(t)}, s_i^{(t)})$ as positive examples, and continue to next iteration until convergence. To make the process robust under visual diversity, we only retain $\mathbf{w}^{(t)}$ using the k matches with the highest score under $\mathbf{w}^{(t-1)}$. In practice the process converges in a few iterations.

We use linear discriminant analysis (LDA) method of Hariharan et al. (2012) to learn $\mathbf{w}$. The method replaces the entire negative set by a Gaussian distribution estimated from a large number of image which speeds up the learning procedure as it avoids the hard-negative mining step commonly during training. However, we still have to perform the latent updates described in Eq. 1 during training. We dub this method 'latent LDA'.

This procedure is illustrated in Fig. 7. The left shows the initial hypothesized matches found using semantic graph (ordered by depth at which they were found). The middle shows the refined matches after the training converges, with location and scale at which the response to the filter $\mathbf{w}$ is maximized. The ordering now reflects the response in (1). In our experiments we restrict the maximum depth of our breadth-first search to two.

For our experiments we divided the set of 288 annotated images as described into a training set of 216 images, and a test set of 72 images. We call this dataset *church-corr*. During training we only use the semantic graph edges entirely contained in the training set (*church-corr-train*), resulting in

Fig. 7 Illustration of the part discovery process. *Left* Top matches found using BFS in the annotation graph. *Center* Top matches sorted by similarity to the source window using a HOG model learned from the source window and negative images. *Right* Visualization of the coefficients of the learned HOG model: *top* shows positive, *bottom* negative weights. The learning procedure refines both the order and the location (shown in *blue*), given the initial estimate (shown in *red*) (Color figure online)

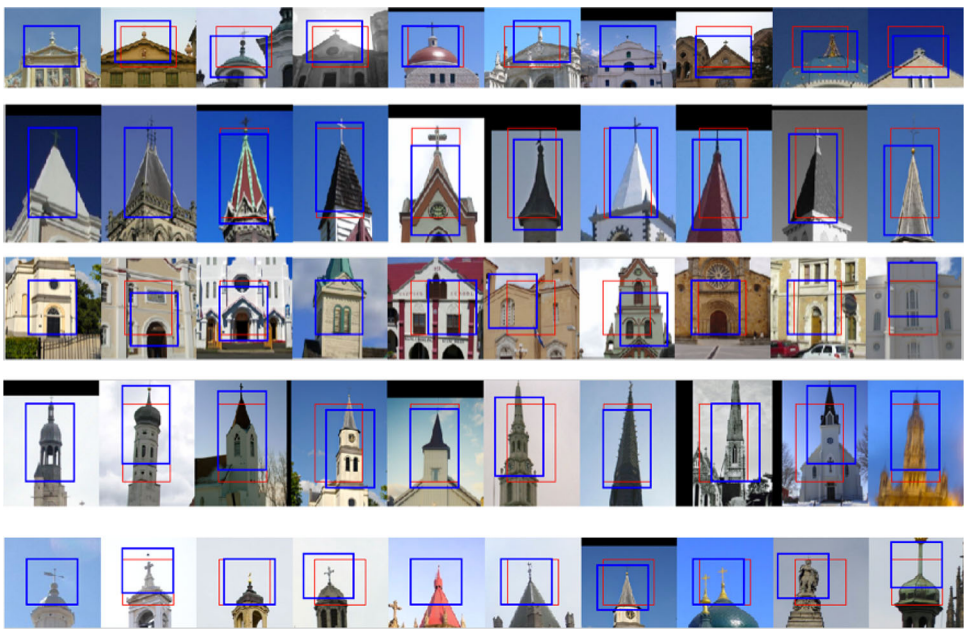


Fig. 8 Examples of discovered parts for churches. On each part the 'latent' location discovered is show in *blue*, while the initial graph-based location is shown in *red*. In each row examples are ordered according to their similarity to the first one, i.e., the score defined by Eq. 1 (Color figure online)

617 pairs, each labelled with an average of five landmarks. The test set (*church-corr-test*) is used to evaluate the utility of parts for predicting the location of the human-clicked landmarks. Since the *church-corr* dataset contains church buildings that occupy most of the image, we collected an additional set of 127 images where the church building occupies a small portion of the image to test the utility of parts for localizing them. The chance performance of detection in these images is small. For these images we also obtained bounding box annotations and the set is further divided into a training set of 64 images and a test set of 63 images. The training set is used to learn bounds regression models for our various methods. We call this dataset *church-loc*.

3.3 Utility of Learned Parts

Figure 8 shows some discovered parts that were highly discriminative (measured as their detection performance) for churches. These parts can then be used to perform a range of recognition tasks such as detection, semantic saliency prediction and detailed parsing using *words* assigned to the parts (when possible).

Object detection We use a simple Hough voting based detector (Leibe et al. 2004) for combining detections from multiple parts. Votes from multiple part detections are combined in a greedy manner. For each image, part detections are sorted

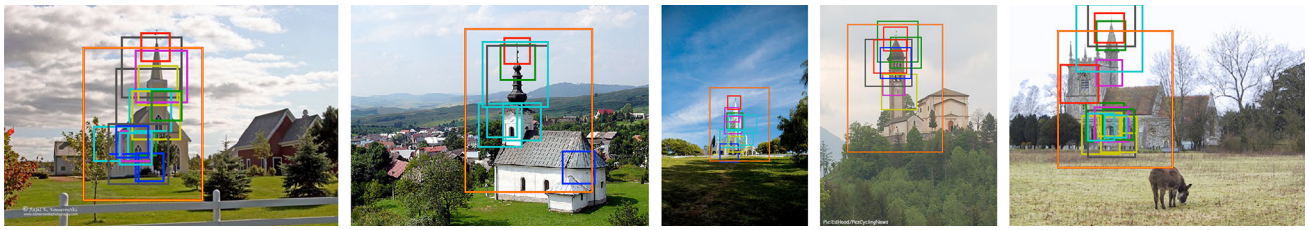


Fig. 9 Example detections of church buildings along with the locations of parts shown in *different colors* (Color figure online)

by their detection score (after normalizing to [0,1] using the sigmoid function) and considered one by one to find clusters of parts that belong together (based on the overlap of their predicted bounding boxes being greater than $\tau = 0.5$). We stop after $n = 500$ part detections are considered. Each cluster represents a detection, from which we predict the overall bounding box as the weighted average of the predictions of each member and score as the sum of their detection scores. This is similar to the detection strategy using poselets (Bourdev et al. 2010).

Figure 9 shows example detections of the church buildings using the discovered parts. Combining the predictions of the top 30 parts our detector achieves an average precision (AP) of 39.90 % compared to a DPM model (Felzenszwalb et al. 2010; Girshick et al. 2012) that achieves an AP = 34.74 %. Our method also outperforms parts obtained using unsupervised 'discriminative patches' (Singh et al. 2012) that obtains AP = 38.34 %, but is orders of magnitude faster during training. Ignoring the correspondence information ('exemplar LDA'), i.e., training parts by restricting the depth of our graph search to zero leads to an AP = 19.95 %, while ignoring the locations of clicks, i.e., random patches leads to an AP = 16.67 %. This shows that the correspondence annotations lead to significantly better parts for detection.

Semantic saliency prediction A landmark saliency map is a function $s(x, y) \rightarrow [0, 1]$, $\sum_{x,y} s(x, y) = 1$, which is a likelihood that a location of the image is a landmark. We can evaluate the likelihood of a given set of ground truth landmark locations under the saliency map as a measure of its predictive quality. Assume a set of n images are all scaled to contain the same number of pixels m. Let $S_k, k = 1, \ldots, n$, denote the set of landmarks in the k^{th} image. The Mean Average Likelihood (MAL) is defined as:

$$\mathsf{MAL} = \frac{1}{n} \sum_{k=1}^{n} \left(\sum_{(x,y) \in S_k} \frac{ms(x, y)}{|S_k|} \right) \tag{2}$$

According to this definition, the *uniform* saliency map has $\mathsf{MAL} = 1$ since $s(x, y) = 1/m, \forall x, y$.

Our saliency detector uses the top 30 parts sorted according to their part detection accuracy on the training set. Given an image, the highest scoring detections above the threshold, up to a maximum of 5 detections, are found for each part. Each detection contributes a saliency proportional to the detection score to the center of the detection window. The contributions are accumulated across all detections to obtain the initial saliency map. This is then smoothed with a Gaussian with $\sigma = 0.01d$, where d is the length of the image diagonal, and normalized to sum to one, to obtain the final saliency map. We set the number of pixels $m = 10^6$.

Our approach can be seen as 'category-specific interest points', and we compare this approach to a baseline that uses standard unsupervised scale-space interest point detectors based on Differences of Gaussians (DoG) and the Itti and Koch saliency model (Itti and Koch 2001). Table 1 shows the MAL scores for various approaches on the *church-corr-test* subset of our dataset. According to our saliency maps, the landmarks are 6.4× more likely than the DoG saliency, and 4.2× more likely than the Itti and Koch saliency. The 'latent LDA' parts outperform both the 'exemplar LDA' parts and 'discriminative patches' (Singh et al. 2012) based saliency. Figure 10 shows example saliency maps for a few images for a variety of methods. As one might expect, our part-based saliency tends to be sharply localized near doors, windows, and towers.

Fine-grained parsing Beyond the standard classification and detection tasks, the rich library of correspondence-driven parts allows us to reason about fine-grained structure of visual categories. For instance, we can attach semantic meaning to a set of parts at almost no cost by simply showing a human a few high-scoring detections. If the parts appear to correspond

Table 1 Performance on 'semantic saliency' prediction task

Method	MAL
Difference of Gaussian	1.23
Itti and Koch Itti and Koch (2001)	1.86
Discriminative patches Singh et al. (2012)	6.14
Exemplar LDA (Landmark seeds)	5.79
Latent LDA on the graph	**7.84**

Bold value indicates best performs
Mean Average Likelihood (MAL) of landmarks according to various saliency maps

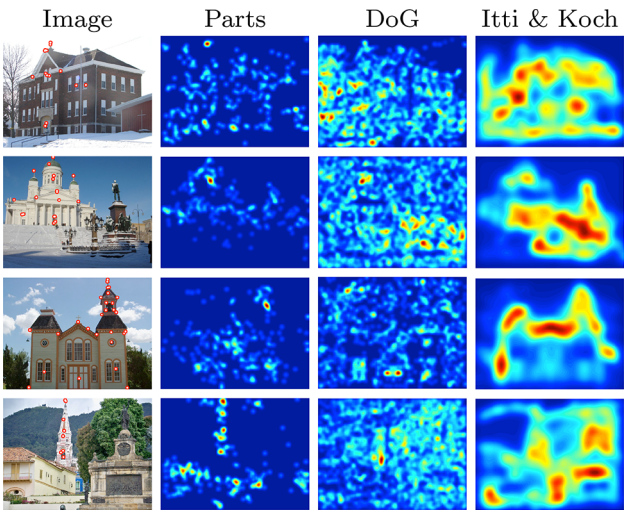


Fig. 10 Predicted saliency maps. From *left* to *right*—images shown with the landmarks; saliency maps from our parts, Difference of Gaussian (DoG) interest point operator, and the Itti and Koch model

to a coherent visual concept with a name, say, 'window' or 'tower', the name for the concept is recorded. Figure 11 (*top row*) shows such labels assigned to various such parts. These semantic labels can be visualized on new images by pooling the part detections across models that correspond to the same label. Figure 11 (*bottom row*) shows example images from the SUN dataset (Xiao et al. 2010), where we have visualized each image with labels positioned at the center of the detection window. Such parsing may be used for search and retrieval of images based on attributes such as 'churches with windows on towers', 'churches with two towers', etc.

4 Fine-Grained Attribute Discovery

Beyond parts, attributes provide an effective language-based interface for humans to query particular instances of a category. Some successful applications include searching faces

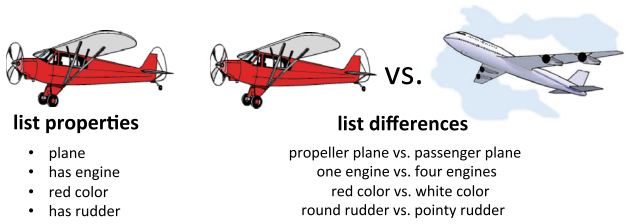


Fig. 12 Discriminative description. Fine-grained attributes are better revealed in the discriminative description task (*right*), than in the traditional description task (*left*)

with desired attributes (Kumar et al. 2008), shopping websites that support structured search, etc. From the computer vision perspective, such attributes can provide insights into which representations are useful for recognition. Indeed, in recent years, vision systems have benefited both in terms of recognition rates and their ability to generalize to new categories by using attributes as an intermediate representation (Bourdev et al. 2011; Farhadi et al. 2010).

This work addresses the issue of identifying the set describable attributes in order to enable fine-grained discrimination within a basic level category. For an attribute to be useful it should achieve the twin goals of *communication* and *discrimination*, i.e., it should be easy to describe the attribute, and it should be useful for discriminating one instance from another. We use this intuition to design our annotation task—we show annotators pairs of images and ask them to describe the differences between them. Thus, the descriptions that we elicit from this process is likely to be more specialized than what we may get by collecting descriptions of instances one at a time. An example shown in Fig. 12 explains the intuition.

The framework also allows us to discover attributes that are relevant to the set of images in hand. For e.g., if all the planes in our dataset were propeller planes, we would discover attributes that distinguish propeller planes from one another.

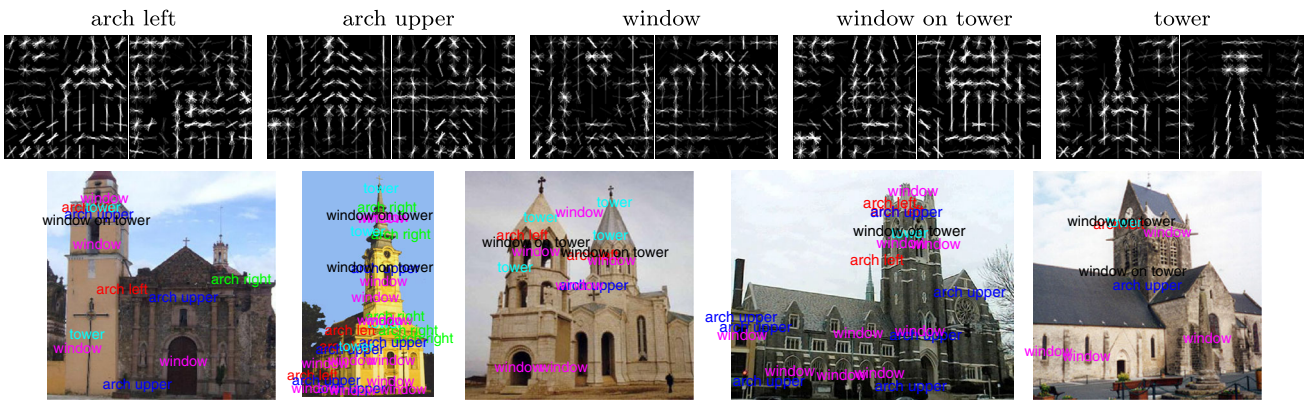


Fig. 11 Fine-grained parsing of images. On the *top row* are labels assigned to parts by humans and on the *bottom row* are localized labels obtained by pooling the corresponding part detections on images (Color figure online)

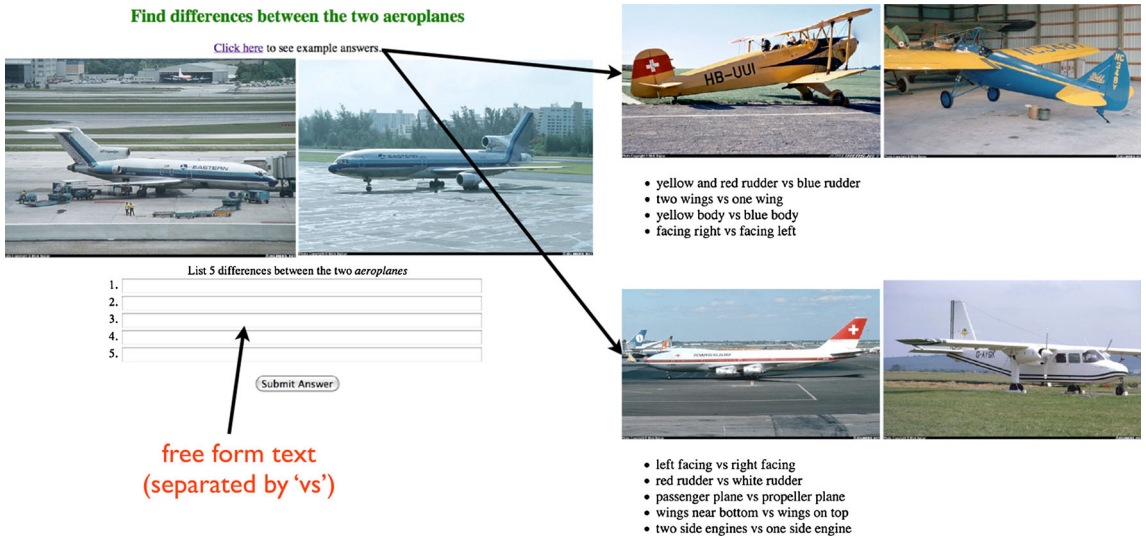


Fig. 13 User interface for collect descriptions on Amazon's mechanical turk (*left*). We also provide some example annotations (*right*) to guide the annotators

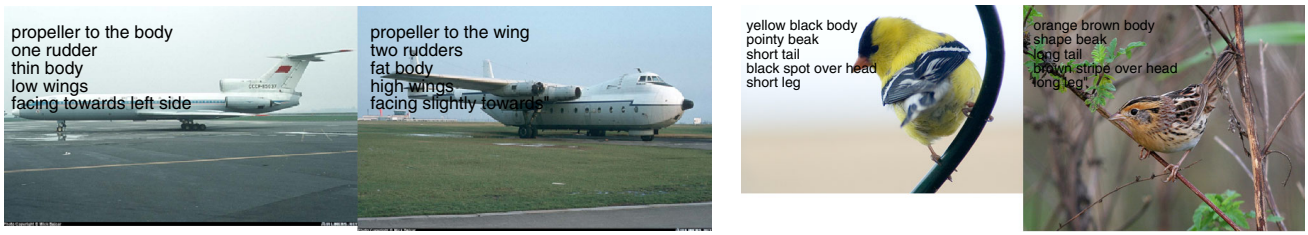


Fig. 14 Example annotations collected using our interface for airplanes (*left*) and birds (*right*)

4.1 User Interface

Our annotation task consists of showing annotators pairs of images and asking them to list 5 visual properties that are *different* between them in *free-form* English. Each sentence is required to have the word 'vs.', which separates the left and the right property as seen in Fig. 13. In addition we show some example annotations to guide the annotation process. It is important to not constrain the descriptions to avoid annotation biases. Once again, as in the correspondence task, there is significantly less effort required to a collect annotation using our interface compared to providing instructions for traditional attribute labeling. Figure 14 shows some annotations collected on Amazon's mechanical turk (AMT) using our interface overlaid on the images.

For a given set of images one can sample pairs at random. The random sampling strategy is good because it biases the discovery process towards those that split the dataset evenly. If a binary attribute is present in a fraction p of the dataset, then the likelihood that it will be revealed in a pairwise comparison is upper bounded by $2p(1-p)$. We need on average 50 pairs of images to find an attribute that appears on 1 % of the dataset. Thus, the pairwise comparison technique is extremely effective in mining discriminative attributes.

4.2 Attribute Discovery by Clustering word Utterances

The discriminative description task provides us with pairs of sentences that can be analyzed to discover a lexicon of parts and attributes. The key observation is that in form of text we collect, each sentence pair typically describes only one part and its modifier. As an example, one may describe a difference between a pair of airplane images as containing 'red rudder vs. blue rudder'. From this sentence pair, one may infer that the noun that is being described is 'rudder', and that it is being modified by 'red' and 'blue'. Moreover, the words 'red' and 'blue' must belong to the same semantic category, which in this case is 'color'. The last one is an additional and a powerful constraint we obtain because we consider pairs of images.

These relationships can be captured by the generative model as shown in Fig. 15. At the top level, topics encode parts and modifiers that are shared across the corpus. Noun topics capture parts, whereas modifier topics capture semantic properties such as 'color' or 'cardinality'. A single noun

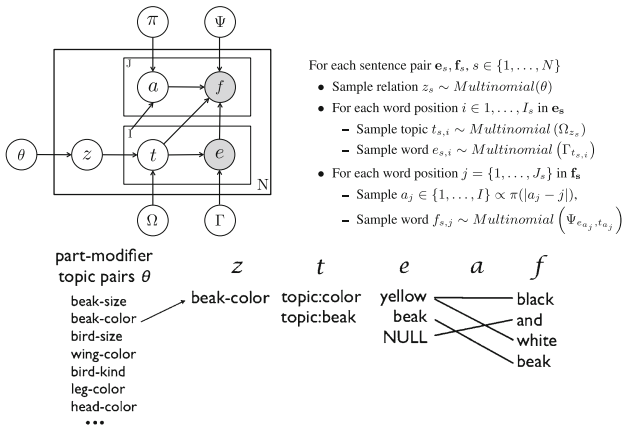


Fig. 15 *Top* The generative model of the of sentence pairs $\{\mathbf{e}_s, \mathbf{f}_s\}$. Each sentence pair in our annotation comes from one *noun* and one *modifier* topic. These topics are shared across all sentences and are estimated from the data using variational EM algorithm. *Bottom* We show the generative process for the sentence pair '*yellow beak* versus *black* and *white beak*'

topic may be modified by several modifier topics (e.g. 'red beak' and 'pointy beak'), and a single modifier topic may modify several noun topics (e.g. 'red beak' and 'red wing'). The set of attributes, i.e., relations between parts and modifiers can thus be expressed as a bipartite graph between the nouns and modifiers topics.

We refer the readers to Maji (2012) for details on parameter estimation and initialization. We note that the generative model proposed here is similar to the IBM word alignment model (Brown et al. 1990), popular in machine translation to initialize translation tables across a pair of languages. Compared to the IBM models, we have also introduced topics to capture semantically coherent noun and modifier topics. This is similar to the Latent Dirichlet Allocation(LDA) model (Blei et al. 2003) and its variants such as Correspondence-LDA (Blei and Jordan 2003) which are used to identify topics in an unsupervised manner from documents. The main difference is that we constrain the topic proportions in each sentence to be *bipartite* corresponding to part-modifier relations.

4.3 Attribute Saliency

As an instance is compared to different ones, different attributes become relevant for discrimination. By computing the frequency with which each attribute was used we can get an estimate of its salient attributes. Figure 16 shows an example where we list the sentences used to describe the shown bird (a 'florida scrub jay') in the order of their frequency. The most frequent sentences used are 'blue wings', 'black wings', 'long tail', etc., which are highly discriminative attributes of this species as can be seen in the wikipedia [1] description shown in the right.

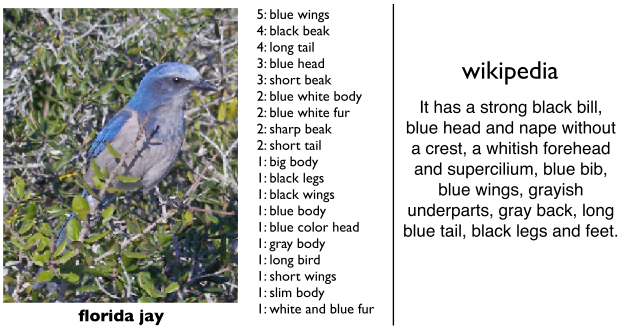


Fig. 16 Attribute saliency. Annotations for the image accumulated over different pairings. The annotations are shown as the raw text input by the annotators along with the frequency with which each phrase was used to describe a difference. The frequency of the property is an indicator of how discriminative it is

4.4 Crowdsourced Discovery of Attributes

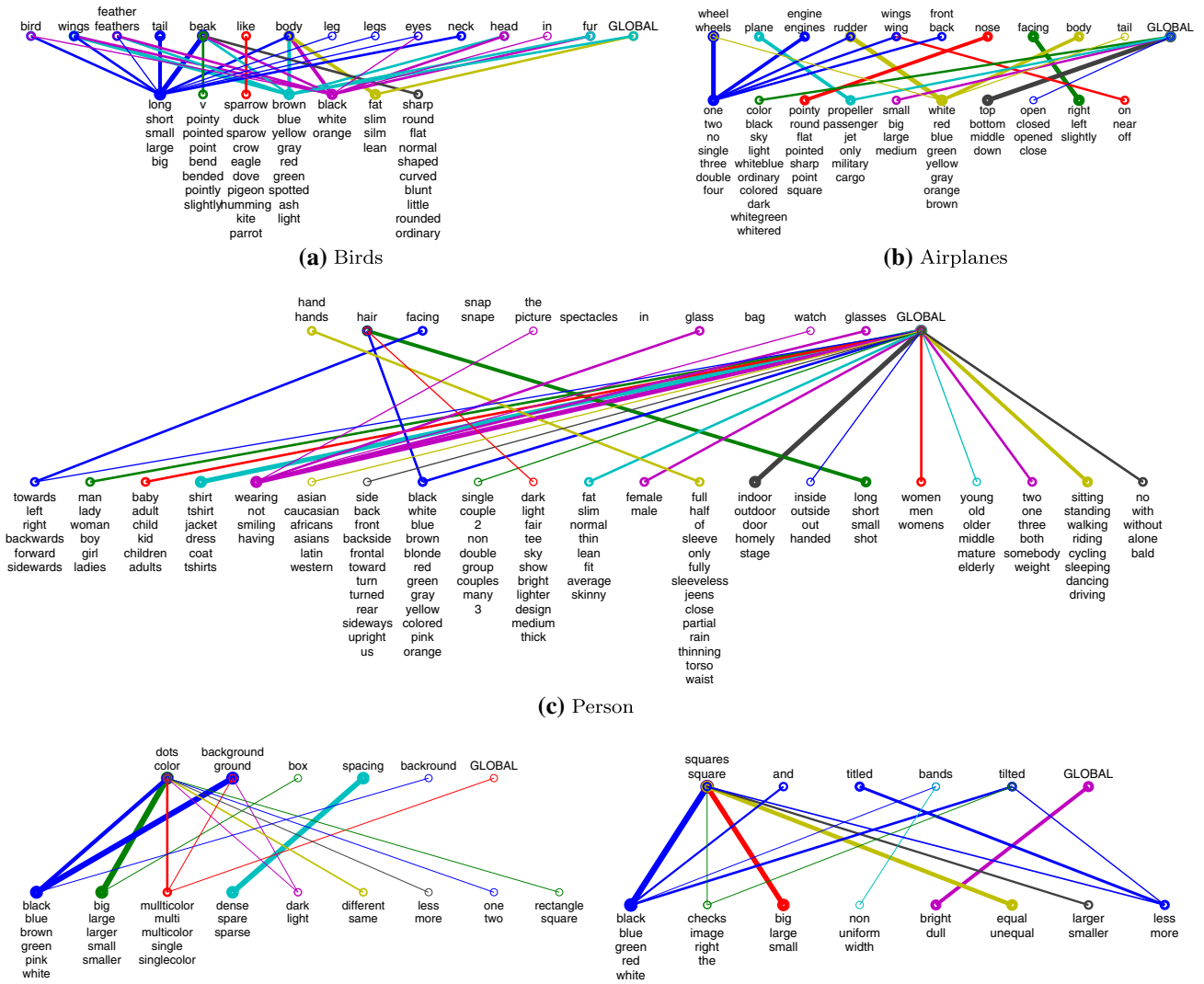
Here we analyze images for a number of basic level categories and use our framework for discovering fine-grained attributes. Figure 17 shows the attributes discovered for birds, airplanes, people and man-made textures. In each figure, the nouns (or parts) are are shown on the top row, modifiers on the bottom row, and the bipartite relation between parts and modifiers which indicate attributes are shown using edges connecting them. The thickness of the edge indicates the frequency of the attribute. The discovered attribute labels can then be used for exhaustive labeling for using traditional annotation methods, or can be used for recognition (Sect. 4.5). Below we describe the datasets and the discovered attributes in more detail.

4.4.1 Caltech-UCSD Birds

The dataset (Welinder et al. 2010) consists of 200 species of birds and was introduced for fine-grained visual category recognition. We gathered 200 images, one random image from each category, and collected annotations for 1,600 pairs sampled uniformly at random.

Figure 17a, shows the learned topics and attributes for birds category. The discovered parts and modifiers refer to parts of the bird such as the *body, beak, wings, tail, head, etc.*, and semantic categories such as *size, color, shape, etc.*, respectively. The most frequent attribute that discriminates birds from one another is the $\{beak\ size\} \leftrightarrow \{small, large\}$, followed by $\{tail\ size\} \leftrightarrow \{long, short, small, ..\}$. Other distinguishing features are colors of various parts such as

[1] http://en.wikipedia.org/wiki/Florida_Scrub_Jay.



(a) Birds **(b)** Airplanes **(c)** Person **(d)** Polka-dotted textures **(e)** Chequered textures

Fig. 17 Discovered parts (*top row*), modifiers (*bottom row*), and attributes (*edges*) for birds, airplanes, person, polka-dotted texture and chequered texture categories (Fig. *a–e*). Each modifier topic is shown with the most frequent words that appear in the clusters. The figures suggests that these modifiers correspond to semantically meaningful adjectives such as 'color', 'shape', 'size', 'cardinality', etc. In each image the thickness of the edge connecting the part and modifier is proportional to the frequency of with which the attribute was used to discriminate pairs of instances in the dataset shown to the annotator. This provides a sorted list of attributes which can be used for fine-grained discrimination

the body, tail and head, and beak shape which can be pointy or round, etc. An interesting attribute that is discovered is $\{like\} \leftrightarrow \{sparrow, duck, crow, eagle, \ldots\}$. The annotators choose to describe birds based on their similarity to a commonly seen ones as the actual species of birds were unknown to the user. Similarity to prototypical birds is a discriminative visual attribute and is often present in field guides.

We compared the attributes discovered by our algorithm to the ones the creators of the Caltech-UCSD birds dataset choose (Welinder et al. 2010). Out of the 12 parts of birds, which are *forehead, crown, bill, eye, throat, nape, breast, back, wing, belly, leg* and *tail*, 6 of them were discovered. Parts such as crown and nape which are sub-parts of the head region were missed likely because they were unfamiliar to non-experts.

4.4.2 Airplanes

We collected 200 images of airplanes from airliners.net, a website of airplane photographs maintained by airplane enthusiasts. We sampled 1,000 pairs uniformly at random to collect annotations.

Figure 17b, shows the discovered attributes. Here, the most frequent attribute is the color of the rudder. Our dataset has many airliners and they often have different rudder colors reflecting the airliner company (e.g. Lufthansa vs. Emirates). The number of wheels is the

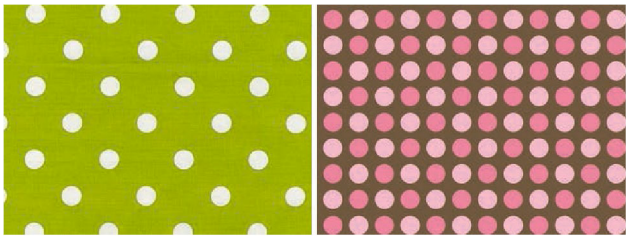


Fig. 18 Example annotation collected on AMT asking users to describe differences between texture pairs

second most distinguishing feature followed by the facing direction. Other discovered attributes are the shape of the nose $\in \{pointy, round, flat, \ldots\}$, kind of the plane $\in \{propeller, passenger, jet, \ldots\}$, overall size $\in \{small, big, large, medium\}$, and the location of the wing relative to the body. Cardinality affects parts such as wheels, engines and rudders, while color modifies the rudder and body. All these are salient properties that distinguish one airplane from another in our dataset.

4.4.3 PASCAL VOC Person

A dataset consisting of attributes of people from the PASCAL visual object challenge (VOC) dataset was introduced by Bourdev et al. (2011). We collected 400 random images from the `trainval` subset of the dataset and we sampled 1,600 pairs uniformly at random and obtained annotations.

Figure 17c shows the discovered attributes for this dataset. We find attributes such as *gender, hair style, hair length, dress type, wearing glasses, hats, etc* which are also identified in Bourdev et al. (2011). In addition, we discover attributes such as the action being performed—*sitting, standing, dancing*, etc.

4.4.4 Man-Made Textures

We collected annotations for 100 'polka-dotted' and 'chequered' texture images collected from the web [these images are also a part of Cimpoi et al. (2014)]. Our automatic analysis yields attributes that describe properties such as the color and size of the dots, their density, color of the background, etc., as shown in Fig. 17d. The same for 'chequered' textures shows that these textures vary according to the size and color of the squares, the color of the background and the tilt as seen in Fig. 17e. Figure 20 shows some examples of these textures.

4.5 Predicting Fine-Grained Texture Attributes

In the earlier section we used the text annotations to discover attributes suitable for fine-grained discrimination. However, these attributes may also be used to train classifiers to predict these attributes from low-level image features. As an example, we show how to predict the dot-size for polka-dotted texture, or the size of the chequered textures.

Figure 18 shows an example of the 'raw' text annotation collected for a pair of polka-dotted texture images. To obtain 'sanitized' annotations suitable for supervised learning we require a bit of additional supervision. In particular we need to group synonyms within a topic into a single group. This can be done by providing simple text processing rules, or may be automated using a dictionary. For example for the size attribute we can group words such as 'small' and 'smaller' into one group, and words such as 'large', 'larger' and 'big' into another. Similarly, for color we can simply use the list of all color words as separate categories. Using this we can automatically assign annotations to images. Figure 19 shows some 'sanitized' annotations. Note that the size attribute is relative and for each pair of images we can only infer the relative size from the text descriptions and the figures show the counts of how many times the dot-size was smaller or larger than in the other image in the pairwise com-

Fig. 19 Discovered and inferred attributes of polka-dotted and chequered texture instances from descriptions of these textures

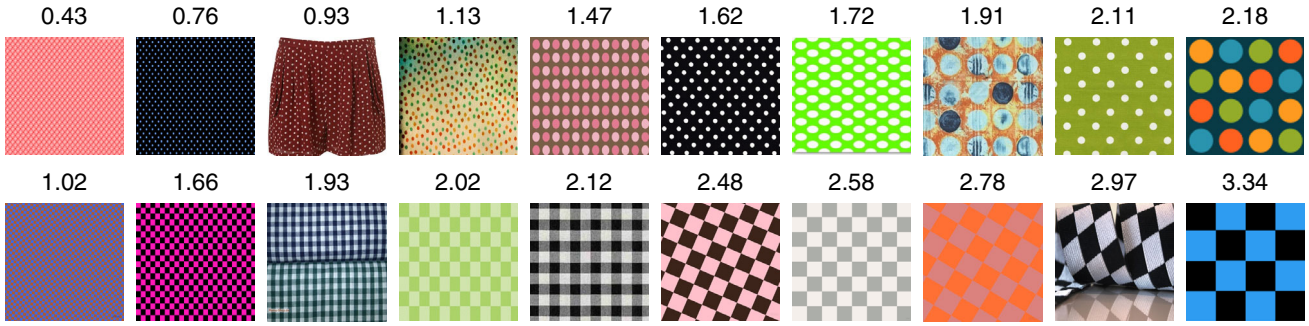


Fig. 20 Fine-grained texture attribute prediction. Automatic prediction of 'dot-size' (*top row*) and 'chequred-size' (*bottom row*) by classifiers trained using labels mined from the text descriptions and low-level coarseness features (Tamura et al. 1978). The images are sorted by the prediction score (shown on top of each image)

parisons. The names of the attributes are assigned manually which correspond to the edges in the bipartite-topic graph (Fig. 17).

We learn to the predict 'dot-size' attribute from low-level features and inferred annotations. For features we use histograms of coarseness values (Tamura et al. 1978) accumulated across all pixels in the image and learn a ranker that respects the ordering of 'dot-size' attribute in the annotations using a learning to rank framework (Joachims 2002). Figure 20 shows every 10th image of our 'polka-dotted' set sorted according to the predicted dot-size. To quantitatively evaluate this approach, we manually annotated the size of the dot in each of the 100 images, and found that our classifier correctly ordered 3,943 of 4,950 (79.66 %) pairs. As a comparison, using the true annotations with the same features correctly ordered 3,971 of 4,950 correct (80.22 %) pairs. Figure 20 shows the images in 'chequred' set sorted by the predicted size of the squares.

5 Conclusion

Studying the correspondences and differences between instances is a powerful means to uncover the structure of visual categories. We leverage such reasoning to design annotation tasks that are particularly effective in discovering parts and attributes. It is quite remarkable that one can obtain such detailed parts (Fig. 8) and attributes (Fig. 17) from crowdsourced data without the need of specialized instructions, careful curation or quality control. The key we believe was a combination of carefully designed interfaces to collect redundant annotations and robust learning methods that enabled us to discover the underlying structure within the data.

Our attribute discovery framework still depends on language in the sense that only namable parts emerge from the annotation process. To avoid this we can unify the part and attribute discovery framework in a single interface were we mark correspondences and list the differences between the clicked pair. This is implicitly being done by the annotators when they describe a localized attribute, e.g., 'red beak vs. black beak'. Furthermore, we can require these differences be in terms of basic properties such as 'color', 'shape', 'cardinality', and 'texture' which might lead to a framework for representing parts and attributes of categories in a language independent manner.

One can use the discovered attributes to group instances into clusters and obtain even fine-grained attributes by collecting differences between pairs of instances within a cluster to obtain a taxonomy of attributes. Somewhat paradoxically as things become more related one can describe more differences between them since more parts can be put in correspondence.

Although in this work we focussed on part and attribute discovery, one can leverage such similarity and difference comparisons more directly in a model of recognition such as 'memex' (Bush 1945), which has been recently popularized in computer vision by Malisiewicz and Efros (2009).

Acknowledgments Part of the work was done by SM during a workshop (http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/tduosn/) at the CLSP, Johns Hopkins University.

References

Agarwal, A., & Triggs, B. (2006). Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 44–58.

Berg, T., Berg, A., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *SIGIR* (pp. 127–134).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bourdev, L., Maji, S., Brox, T., & Malik, J. (2010). Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision*.

Bourdev, L., Maji, S., & Malik, J. (2011). Describing people: A poselet-based approach to attribute classication. In *International Conference on Computer Vision*.

Bourdev, L., & Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*.

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., & Belongie, S. (2010). Visual recognition with humans in the loop. In K. Daniilidis, P. Maragos & N. Paragios (Eds.), *Computer vision-ECCV 2010* (pp. 438–451). Berlin: Springer.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., et al. (1990). A statistical approach to machine translation. *Computational Linguistics, 16*, 79–85.

Bush, V. (1945). The atlantic monthly. *As we may think*.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In N. Dalal & B. Triggs (Eds.), *Computer Vision and Pattern Recognition* (pp. 886–893).

Desai, C., & Ramanan, D. (2012). Detecting actions, poses, and objects with relational phraselets. In *Computer vision-ECCV 2012* (pp. 158–172). Berlin: Springer.

Duan, K., Parikh, D., Crandall, D., & Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3474–3481).

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Farhadi, A., Endres, I., & Hoiem, D. (2010). Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition*.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, *61*, 55–79.

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition*.

Frome, A., Singer, Y., & Malik, J. (2007). Image retrieval and classification using local distance functions. In *Advances in neural information processing systems 19: Proceedings of the 2006 conference* (Vol. 19, p. 417). MIT Press.

Girshick, R. B., Felzenszwalb, P. F., & McAllester, D. (2012) Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/rbg/latent-release5/.

Hariharan, B., Malik, J., & Ramanan, D. (2012). Discriminative decorrelation for clustering and classification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato & C. Schmid (Eds.), *Computer vision-ECCV 2012* (pp. 459–472). Berlin: Springer.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133–142). ACM.

Kovashka, A., Parikh, D., & Grauman, K. (2012). Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2973–2980). IEEE.

Kumar, N., Belhumeur, P., & Nayar, S. (2008). Facetracer: A search engine for large collections of images with faces. In *European conference on computer vision*.

Leibe, B., Leonardis, A., & Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *ECCV workshop on statistical learning in computer vision* (pp. 17–32).

Maji, S. (2011). Large scale image annotations on amazon mechanical turk. Tech. Rep. UCB/EECS-2011-79, EECS Department, University of California, Berkeley (2011). http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-79.html

Maji, S. (2012). Discovering a lexicon of parts and attributes. In *Second International Workshop on Parts and Attributes, ECCV*.

Maji, S., & Shakhanarovich, G. (2013). Part discovery from partial correspondence. In *Computer vision and pattern recognition*.

Maji, S., & Shakhnarovich, G. (2012). Part annotations via pairwise correspondence. In *Human computation workshops at the AAAI*.

Malisiewicz, T., & Efros, A. (2009). Beyond categories: The visual memex model for reasoning about object relationships. In *Advances in neural information processing systems* (pp. 1222–1230).

Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Ensemble of exemplar-svms for object detection and beyond. In *International conference on computer vision*.

Parikh, D., & Grauman, K. (2011). Interactive discovery of task-specic nameable attributes. In *Workshop on fine-grained visual categorization, CVPR*.

Patterson, G., & Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2751–2758). IEEE.

Singh, S., Gupta, A., & Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato & C. Schmid (Eds.), *Computer vision-ECCV 2012* (pp. 73–86). Berlin: Springer.

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, *8*(6), 460–473.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. (2011). Adaptively learning the crowd kernel. In *International conference on machine learning (ICML)*. Bellevue, WA.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 319–326). ACM.

Von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 55–64). ACM.

Weber, M., Welling, M., & Perona, P. (2000). Towards automatic discovery of object categories. In *Computer vision and pattern recognition*.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). Caltech-UCSD birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3485–3492). IEEE.

Yang, Y., & Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *2011 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1385–1392). IEEE.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In: *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2879–2886). IEEE.