A Taxonomy of Part and Attribute Discovery Techniques

Subhransu Maji

Abstract This chapter surveys recent techniques for discovering a set of *Parts and Attributes* (PnAs) in order to enable fine-grained visual discrimination between its instances. *Part and Attribute* (PnA) based representations are popular in computer vision as they allow modeling of appearance in a compositional manner, and provide a basis for communication between a human and a machine for various interactive applications. Based on two main properties of these techniques a unified taxonomy of PnA discovery methods is presented. The first distinction between the techniques is whether the PnAs are semantically aligned, i.e. if they are human interpretable or not. In order to achieve the semantic alignment these techniques rely on additional supervision in the form of annotations. Techniques within this category can be further categorized based on if the annotations are language-based, such as *nameable* labels, or if they are language-free, such as *relative similarity comparisons*. After a brief introduction motivating the need for PnA based representations, the bulk of the chapter will be dedicated to techniques for PnA discovery categorized into *non-semantic*, *semantic language-based*, and *semantic language-free* methods. Throughout the chapter we will illustrate the trade-offs among various approaches though examples from the existing literature.

1 Introduction

This chapter surveys a number of part-based and attribute-based models proposed in the last decade in the context of visual recognition, learning, and description for human-computer interaction. Part-based representations have been very successful for various recognition tasks ranging from detecting objects in cluttered scenes [9, 34], segmenting objects [16, 107], recognizing scene categories [45, 72, 77, 92], to recognizing fine-grained attributes of objects [10, 111, 98]. Parts provide robust-

Subhransu Maji
University of Massachusetts, Amherst e-mail: smaji@cs.umass.edu

ness to occlusion – the head of a person can be detected even when the legs are occluded. Parts can also be composed in different ways enabling generalization to novel viewpoints, poses, and articulations of objects. Two popular methods, namely the *Deformable Part-based Model* (DPM) of Felzenszwalb et al. [34] and the *poselets* of Bourdev et al. [9, 11], exploit this property to build robust object detectors.

The compositional nature of part-based models is also the basis for *Convolutional Neural Networks* (CNNs). While traditional part-based models can be seen as shallow networks where the representations are hand-designed, CNNs learn all the model parameters from raw-pixels to image labels in an end-to-end manner using a deeper architecture. When trained on large labeled datasets, deep CNNs have led to breakthrough results on a number of recognition tasks [44, 48, 87], and are currently the dominant approach for nearly all visual recognition problems.

Beyond recognition, a set of parts provides a means for a human to indicate the pose and articulation of an object. This is useful for recognition with humans "in the loop" where a person can annotate a part of the object to guide recognition. For instance, Branson et al. [13] interactively categorize birds by asking users to click on discriminative parts leading to significant improvement over the computer vision only baseline. In such cases it is desirable that the parts represent semantically-aligned concepts since it involves communication with a human.

Along with parts, *visual attributes* provide a means to model the appearance of objects. The word "attribute" is extremely generic as it can refer to any property that might be associated with an object. Attributes can describe an entire object or a part, e.g., a tall person or a long nose. Attributes can refer to low-level properties such as color and texture, or high-level properties such as age and gender of a person. Attributes can be shared across categories, e.g., both a dog and a cat can be "furry", allowing the description of previously unseen categories. Semantically aligned attributes provide a basis for learning interpretable visual classifiers [33], create classifiers for unseen categories [52], debugging recognition systems through attribute-based explanations [3, 76], and providing human feedback during learning and inference [14, 46, 51, 78].

Thus, PnAs provide a rich compositional way of describing and recognizing categories. Techniques for PnA discovery are necessary as the desired set of parts and attributes often depend on the underlying task. While it may not be necessary to model the gender, hair-style, or the eye-color of a person for detecting them, it may be useful for identifying a particular individual. One motivating reason for the unified treatment of PnAs in this chapter is that their roles are interchangeable for recognition and description. For instance, in order to distinguish between a red-beaked and a yellow-beaked bird, one could have two parts, "red beak" and "yellow beak" and no attributes, or a single part "beak" with two attributes, red and yellow. Therefore, from a representation point of view it is more fruitful to think of the joint space induced by various part-attribute interactions instead of each one of them independently. In other words we can think of attributes being localized, i.e. associated with a part, or not.

The next section provides an overview of the rest of the chapter, and describes a unified taxonomy of recent PnA discovery methods.

1.1 Overview

Although there are many ways to categorize the vast number of methods for PnA discovery in the literature, the particular one described in this chapter was chosen because it is especially useful for fine-grained domains which is our main focus. Often these domains have a rich structure described through language, visual illustrations, and other modalities, which can be used to guide representation learning. Translating all this information to useful visual properties is one of the main challenges of these methods. The proposed taxonomy categorizes various PnA methods based on:

- the degree to which the models explicitly try to achieve *semantic alignment* or *interpretability* of the underlying PnAs,
- the nature of the source of semantics, i.e. if they are language-based or not.

When semantic alignment is not the primary goal, the PnAs can be thought of as a intermediate representation of the appearance of objects. Example methods for part discovery in this setting include DPMs [34], and CNNs [48, 56]. Here the learned parts factorize the appearance variation within the category and are learned without additional supervision apart from the category labels at the object or image level. Hence, semantic alignment is not guaranteed and parts that arise tend to represent visually salient patterns. Similarly non-semantic attributes can be thought of as the coordinates in a transformed space of images optimized for the recognition task. Such methods are described in Section 2.1 and Section 2.2.

Language is a natural source of semantics. Although the vocabulary of parts and attributes that arise in language are a result of multiple phenomena, they provide a rich source of interpretable visual PnAs. For instance, parts of animals can be based on the names of anatomical parts. Various existing datasets that contain part annotations follow this strategy. This include the *Caltech-UCSD Birds* (CUB) dataset [100], *OID:Airplanes* dataset [98], and part annotations of animals in PASCAL VOC dataset [9, 20]. Similarly, attributes can be based on common color, texture, and shape terms used in language, or can be highly specialized language-based properties of the category. For example, the CUB dataset annotates parts of birds with color attributes, while the Berkeley "attributes of people" dataset [10] contains attributes describing gender, clothing, age, etc. We review techniques for collecting language-based attribute and part annotations in Sections 3.1 and Section 3.4 respectively.

Task-specific language-based PnAs can also be *discovered* by analyzing descriptions of objects (Section 3.2). For example, Berg et al. [6] analyze captioned images on the web to discover attributes. Nameable attributes may also be discovered *interactively* by asking annotators to *name* the principal directions of variations within the data [79], by selecting a subset of attributes that frequently discriminate instances [80], or by analyzing descriptions of differences between instances [63]. We review such techniques in Section 3.3.

Beyond language, semantic alignment of PnAs may also be achieved by collecting language-free annotations (Section 4). An example of this is through similarity

comparisons of the form *"is A more similar to B than C"*. The coordinates of the embedded space that reflects these similarity comparisons can be viewed as an semantic attribute [101] (Section 4.1). Another example is when an annotator clicks on landmarks between pairs of instances. Such data can be collected without having to name the parts providing a way to annotate parts for categories that do not have a well defined set of *nameable* parts [65]. The resulting pairwise correspondence data can be used for learning semantic part appearance models (Section 4.2).

Figure 1 shows the taxonomy pictorially. Existing approaches are divided into three main categories: *non-semantic PnAs* (Section 2), *semantic language-based PnAs* (Section 3), and *semantic language-free PnAs* (Section 4). Within each category we further organize approaches into various sections to illustrate the scenarios when they are applicable and the computational *v.s.* annotation-cost trade-offs they offer. We describe some open questions and conclude in Section 5.

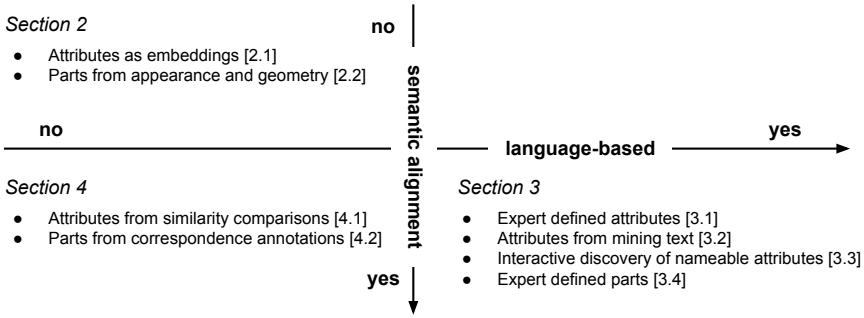


Fig. 1 A taxonomy of PnA discovery techniques discussed in this chapter based on the degree of semantic alignment (y-axis) and if they are language-based (x-axis). Various sections and subsections in this chapter are listed within each quadrant.

2 Non-semantic PnAs

A common theme underlying techniques for non-semantic PnA discovery is that the parts and attributes arise out of a framework where the goal is a *factorized* representation of the appearance space. Pictorially, one can think of PnAs as an intermediate representation between the images and high-level semantics. The factorization results in better computational efficiency, statistical efficiency, and robustness of the overall model.

2.1 Attributes as embeddings

A typical strategy of learning attributes in this setting is to constrain the intermediate representation to be low-dimensional or sparse. Techniques for dimensionality reduction, such as *k-means* [59], *Principal Component Analysis* (PCA) [42], *Locality Sensitive Hashing* [37], *auto-encoders* [4], and *spectral clustering* [68], can be applied to obtain compact embeddings.

An early application of such approach for recognition is the eigenfaces of Turk and Pentland [97]. PCA is applied to a large number of *aligned* frontal faces to learn a low-dimensional space corresponding to the first few PCA basis. These capture the major axes of variations, some of which are aligned to factors such as lighting, or facial expression. The low dimensional embedding was used for face recognition in their setting. One can use an image representation such as Fisher Vector [81, 82] instead of pixel values before dimensionality reduction for additional invariance. These techniques have no explicit control over the semantic alignment of the representation, and are not guaranteed to lead to interpretable attributes.

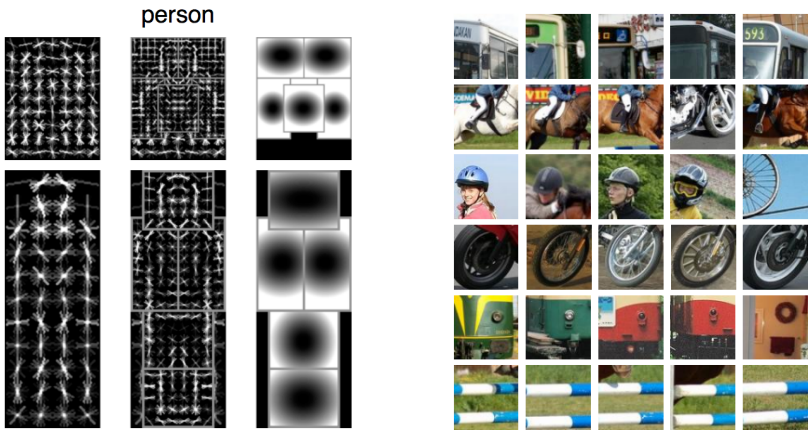
In a *task-specific setting* the intermediate representation can be optimized for the final performance. An example of this is a two-layer neural network for image classification that takes raw pixels as input and produces class probabilities via an intermediate layer which can be seen as attributes.

There are many realizations of this strategy in the literature that vary in the specifics of the architecture and the nature of the task. For example, the "picodes" approach of Bergamo et al. [7] learns a compact binary descriptor (e.g., 16 bytes) that has a good object recognition performance. Attributes are parametrized as $a(\mathbf{x}) = \mathbf{1}[\mathbf{w}^T\mathbf{x} > 0]$, for some weight vector $\mathbf{w}$ for an input representation $\mathbf{x}$. Rastegari et al. [86] use a similar parameterization but use a notion of "predictability" measured as attributes that achieve high separation between classes as the objective. Yu et al. [109] learn attributes by formulating it as a matrix factorization problem.

Experiments reported in the above work show that the task-driven attributes achieve better performance compared to unsupervised methods for attribute discovery on datasets such as Caltech-256 [40] and ImageNet [28]. Moreover, they provide a compact representation of images for efficient retrieval and other applications.

2.2 Part discovery based on appearance and geometry

In addition to appearance, part-based models can take into account the geometric relationships between the parts during learning. In the unsupervised, or task-free setting, parts may be obtained by clustering local patches using any unsupervised method such as k-means, spectral clustering, etc. This is the one of the key steps in the bag-of-visual-words representation of images [24] and their variants such as the Fisher Vector [81, 82] and Vector of Locally Aggregated Descriptors (VLAD) [43], which are some of the early successful image representations.



(a) Figure source: Felzenszwalb et al. [34] (b) Figure source: Gupta et al. [92]

Fig. 2 (a) Two components of the deformable part-based model learned for the person category. The "root" and "part" templates are show using the HOG feature visualization (left and middle) and the spatial model is shown on the right. (b) Examples of discriminative patches discovered for various classes in the PASCAL VOC dataset.

Geometric information can be added during the clustering process to account for spatial consistency, e.g., by coarsely quantizing the space using a spatial pyramid [55], or by appending the coordinates of the local patches (called "spatial augmentation") to the appearance before clustering [90, 91]. Parts may also be discovered via correspondences between pairs of instances obtained by some low-level matching procedure. For instance, Berg et al. [5] discover important regions in images by considering geometrically consistent feature matches across instances.

Another example of a model that combines appearance and geometry for part learning is the DPM of Felzenszwalb et al. [34]. The model has been widely used for object detection in cluttered scenes. A category is modeled as a mixture of components, each of which is represented as a "root" template and a collection of "parts" that can move independently relative to the root template. The tree-like structure of the model allows efficient inference through distance transforms. The parameters of the model are learned through an iterative procedure where the component membership, part positions, and appearances models are updated in order to obtain good separation between positive examples and the background. Figure 2a shows two components learned for person detection on the PASCAL VOC dataset [32]. The compositional architecture of the DPM led to significant improvements over the monolithic template-based detector of Dalal and Triggs [25].

Another example for task-driven part discovery is the "discriminative patches" approach of Singh et al. [92]. Here parts are initialized by clustering appearance, and through a process of positive and hard-negative mining the part appearances are iteratively refined. Finally parts that are *frequent* and help *discriminate* among classes are selected. Figure 2b shows example discriminative patches discovered for the PASCAL VOC dataset. The authors demonstrate good performance on image

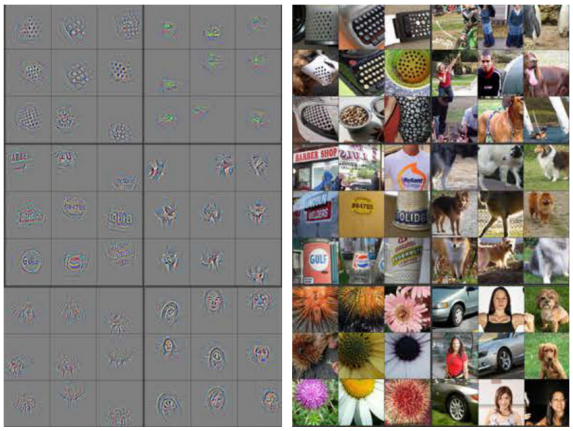


Fig. 3 Visualizations of the top activations of six *conv5* units of the AlexNet CNN [48] trained on ImageNet dataset [28]. For each image patch on the left the locations of where that are responsible for the activations are also shown on the left. The units strongly respond to parts such as dog and human faces, as well as attributes such as "grid-like" and "text". *Figure source:* Zeiler and Fergus [110]

classification datasets, such as PASCAL VOC, MIT Indoor scenes [83], using a representation that records the activation of discriminative patches at different locations and scales (similar to a bag-of-visual-words model [24]).

Since these methods primarily rely on appearance and geometric consistency, the discovered parts may not be aligned to semantics. For instance, the DPM requires that each object have the same set of parts even if the object is partially occluded. Hence the model uses a part to both recognize a part of the object or its occluder. Similarly, discriminative patches are visually consistent parts according to the underlying *Histograms of Oriented Gradient* (HOG) features [25] and hence two patches that are visually dissimilar but belong to the same semantic category are unlikely to be grouped as the same part. For example, two kinds of car wheels, or two styles of windows, will be represented using two or more parts.

Convolutional Neural Networks (CNNs) can be seen as part-based model trained in an end-to-end manner, i.e. starting from a pixel representation to class labels. The hierarchy of convolution and max-pooling layers resemble the computation of a deformable part-based model. Indeed, the DPM can be seen as a particular instantiation of a CNN since both HOG (see Mahendran and Vedaldi [62]) and the DPM computations (see Girshick et al. [38]) can be written as shallow CNNs. However, after the recent breakthrough result of Krishevsky et al. [48] on the ImageNet classification dataset [28], CNNs have become the architecture of choice for nearly all visual recognition tasks [12, 23, 39, 44, 60, 87, 94, 111, 112].

CNNs trained in a supervised manner can be seen to simultaneously learn parts and attributes. For instance, visualizations of the "AlexNet CNN" [48] by Zeiler and Fergus [110], as seen in Figure 3, reveal units that activate strongly on parts such as human and dog faces, as well as attributes such as "text" and "grid-like". Recent works, such as the *bilinear CNNs* [57] show that discriminative localized attributes emerge when these models are fine-tuned for fine-grained recognition tasks. Figure 4 shows example filters learned when these mdoels are trained on birds [100], cars [47], and airplane [64] datasets. The remarkable performance of CNNs shows that considering part and attribute discovery *jointly* can have significant benefits.

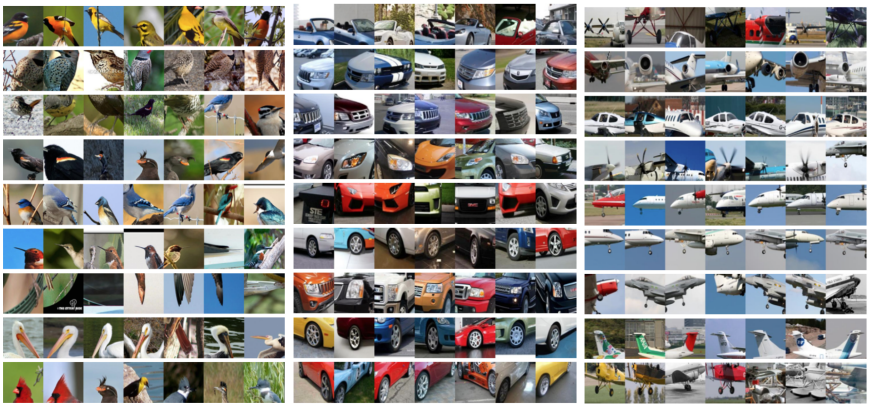


Fig. 4 Visualizions of the top activations of several units of the "bilinear CNN" (B-CNN [D,M] model) [57] fine-tuned on birds [100] (left), cars [47] (middle), and airplane [64] (right) datasets. Each row shows the patches in the training data with the highest activations for a particular unit of the "D network" (See [57] for details). The units correspond to various localized attributes ranging from yellow-red stripes (row 4) and particular beak shapes (row 8) for birds, wheel detectors (rows 6, 8, 9) for cars, to propeller (rows 1, 4) and vertical-stabilizer types (row 8) for airplanes.

3 Semantic language-based PnAs

Language is the source of categories for virtually all modern datasets in computer vision. The widely used ImageNet dataset reflects the hypernymy-hierarchy ("is a" relationships) of nouns in WordNet – a lexical database of words in English organized in a variety of ways [67]. Naturally, language is also a source of PnAs useful for a high-level description of objects, scenes, materials, and other visual phenomenon. For example, a cat can be described as a four-legged furry animal. This human-interpretable description of learned models provides a means for communication between a human and machine during learning and inference. Below we overview several applications of language-based PnAs from the literature.

3.1 Expert defined attributes

An early example of language-based attributes in the computer vision community was for describing texture. Bajscy proposed attributes such as orientation, contrast, size, and spacing of structural elements in periodic textures [2]. Tamura et al. [95] identified six visual attributes of textures namely *coarseness, contrast, directionality, linelikeness, regularity,* and *roughness*. Amadasun and King derived computational models for five properties of texture, namely, *coarseness, contrast, business, complexity,* and *texture strength* [1].

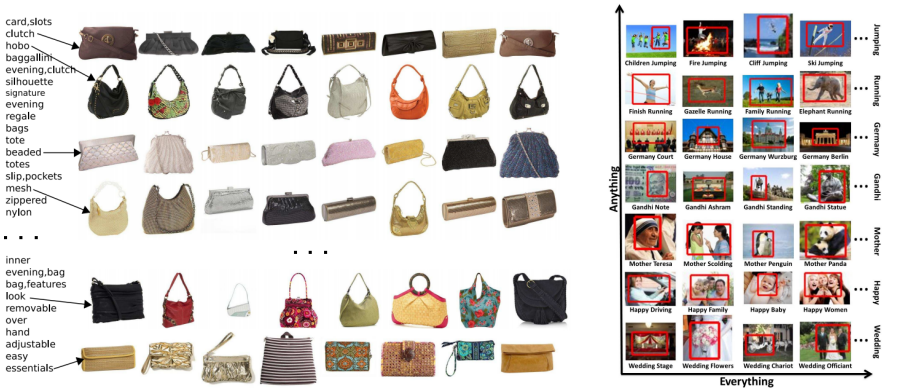
Recently, Cimpoi et al. [22] extended the set of describable attributes to include 47 different words based on the work of Rao and Lohse [85]. Other texture attributes such as material properties have been used to construct datasets such as *CUReT* [26], *UIUC* [54], *UMD* [105], *Outex* [69], *Drexel Texture Database* [71], *KTH-TIPS* [17, 41] and *Flickr Material Dataset* (FMD) [89]. In all the above cases experts identified the set of language terms as attributes based on domain knowledge, or in some cases through human studies [85].

Beyond textures, language-based attributes have since been proposed for a variety of other datasets and applications. Farhadi et al. [33] describe object categories with *shape*, *part-names* and *material attributes*. Lampert et al. [52] proposed the *Animals with Attributes* (AwA) dataset consisting of variety of animals with shared attributes such as color, food habits, size, etc. The *Caltech-UCSD Birds* (CUB) dataset [100] consists of hundreds of species of birds labeled with attributes such as the shape the beak, color of the wings, etc. The *OID:Airplanes* [98] dataset consists of airplanes labeled with attributes such as number of wings, type of wheels, shapes of parts, etc. Attributes such as gender, eye color, hair syle, etc., have been used by Kumar et al. [49] to recognize, describe, and retrieve faces. Other examples include attributes of people [10], human actions [58], clothing style and fashion [19, 106], urban tribes [50], and asthetics [30].

A challenge is using language-based attributes to the degree of specialization to be considered. For instance, while an attribute of airplane such as the *shape of the nose* can be understood by most people, an attribute such as the *type of the aluminum alloy used in manufacturing* can only be understood by a domain expert. Similarly, the scientific names of parts of animals are typically known only to a domain expert. While common attributes have the advantage that they can be annotated by "crowdsourcing", they may lack the precision needed for fine-grained discrimination between closely related categories. Bridging the gap between expert-defined and commonly-used attributes remains an open question. In the context of object categories this aspect has been studied by Ordonez et al. [70] where they learn common names ("entry-level categories") by analyzing the frequency of usage in text on the Internet, e.g. *grampus griseus* is translated to a *dolphin*.

3.2 Attribute discovery by automatically mining text

Language-based attributes may also be mined from large sets of images with captions. Ferrari and Zisserman [36] mine attributes of texture and color from descriptions on the web. Berg et al. [6] obtain attributes by mining frequently occurring phrases from captioned images and estimating if they are visually salient by training a classifier to predict the attribute from images (Figure 5a). In the process they also characterize if the attributes are localized or not. Text on the Internet from online books, Wikipedia articles, etc., have been mined to discover attributes for objects [31] (Figure 5b), semantic affordances of objects and actions [18], and other common-sense properties of the visual world [21].



(a) Figure source: Berg et al. [6] (b) Figure source: Divvala et al. [31]

Fig. 5 (a) Automatically discovered handbag attributes from [6], sorted by "visualness" measured as the predictability of the attribute based on visual features. (b) Automatically mined visual attributes for various categories from books [31].

3.3 Interactive discovery of nameable attributes

While captioned images are a great source of attributes, the vast majority of categories are not well represented in captioned images on the web. In such situations one can aim to discover nameable attributes *interactively*. Parikh and Grauman [73] show annotators images that vary along a projection of the underlying features and ask them to describe it if possible (Figure 6a). To be effective the method requires a feature space whose projections are likely to be semantically correlated.

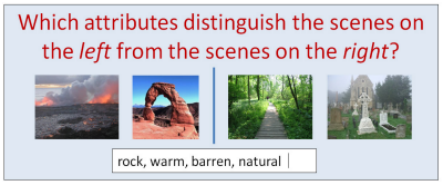
Patterson and Hays [80] start from a set of attributes mined from natural languge descriptions and ask annotators to select five attributes that distinguish images from various scene classes in the SUN database. Thus attributes suited for discrimination within the set of images can be discovered (Figure 6b).

A similar strategy was used in my earlier work [63] where annotators were asked to describe the visual differences between pairs of images (Figure 6c) revealing fine-grained properties useful for discrimination. The collected data was mined to discover a lexicon of parts and attributes by analyzing the *frequency* and *co-occurrence* of words in the descriptions (Figure 7).

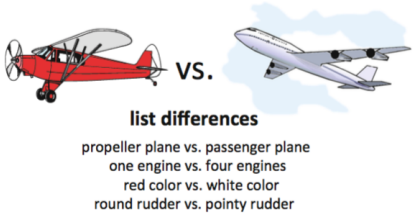
3.4 Expert defined parts

Like attributes, language-based parts have been widely used in computer vision for modeling articulated objects. An early example of this is *pictorial structure* model for detecting people in images where parts were based on the human anatomy [35].

(a) Figure source: Parikh and Grauman [73]



(b) Figure source: Patterson and Hays [80]



(c) Figure source: Maji [63]

Fig. 6 Interactive attribute discovery. Annotators are asked to (a) *name what varies* in the images from left to right [73], (b) *select attributes* that distinguish images on the left from the right [80], and (c) *describe the differences* between pairs of instances [63]. The collected data is analyzed to discover a set of *nameable* attributes.

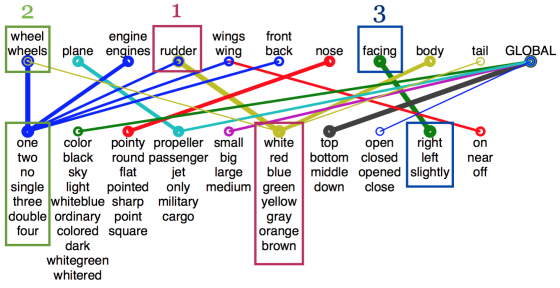


Fig. 7 The vocabulary of parts (top row) and their attributes (bottom row) discovered by from sentence pairs describing the differences between images in *OID:Airplanes* dataset [98]. The three most discriminative attributes are also shown. Figure source: Maji [63].

A modeling decision that is unique compared to attributes is the choice of the spatial extent, scale, pose, and other visual phenomenon, for a given semantic part.

Broadly, there are commonly used methods for collecting part annotations (Figure 8). The first is *landmark-based* where positions of landmarks, such as joint positions of humans, or fiducial points for faces are annotated. The second is *bounding-box-based* where part bounding-boxes are explicitly labeled to define the extent of each part. The bounding-boxes may be further refined to reflect the pixelwise support or segmentation of the parts.

When landmarks are provided one could simply assume that parts correspond to these landmarks. This strategy has been applied for modeling faces with fiducial points [113], articulated people with deformable part-based models [35, 108], etc. Another strategy is to discover parts that correspond to frequently occurring configuration of landmarks. The *poselets* approach combines this strategy with a procedure to select a set of *diverse* and *discriminative parts* for the task of person detection [9]. The discovered poselets are different from both landmarks and anatomical parts (Figure 9a). For instance, a part consisting of *half the profile face and the right shoulder* is a valid poselet. These patterns can capture distinctive ap-

pearances that arise due to self-occlusion, foreshortening, and other phenomenon which are hard to model in a traditional part-based model.

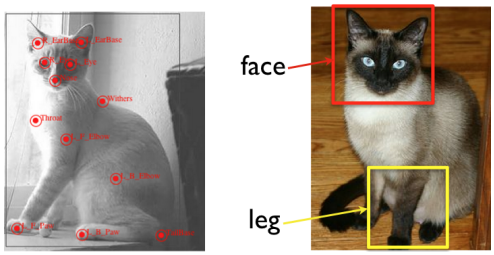


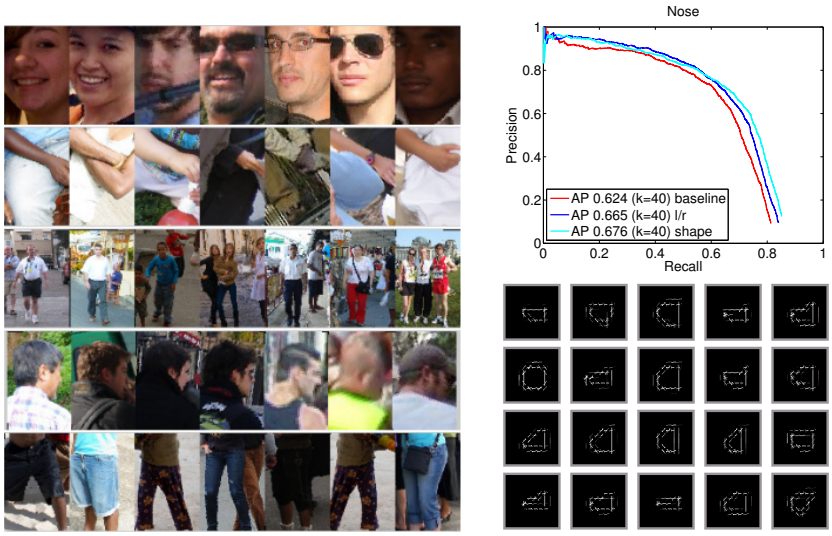
Fig. 8 Two methods for collecting part annotations. On the left, the positions of set of landmarks are annotated. On the right, bounding-boxes for parts are annotated.

When bounding-boxes are provided there is relatively little flexibility in part discovery. Much work in this setting has focused on effectively modeling appearance through a mixture of templates. Additional annotations, such as viewpoint, pose, or shape, can be used to guide mixture model learning. For instance, Vedaldi et al. [98] show that using shape and viewpoint annotations to initialize HOG-based parts improves detection accuracy compared to the aspect-ratio based clustering (Figure 9b).

4 Semantic language-free PnAs

Language-based PnAs, when applicable, provide a rich semantic representation of objects. However language alone may not be sufficient to capture the full range of visual phenomena. Consider the space of colors defined by the [R,G,B] values. Berlin and Kay in their seminal work [8] analyzed the words used to describe color across widely across languages. While languages like English have many words to describe color, there are languages that have very few words, including an extreme case of language with only only have two words ("bright" and "dull") to describe color leading to a gross simplification of the color space. Similarly, restricting one to nameable parts poses challenges in annotating categories that are structurally diverse. It would require significant effort to define a set of parts that apply to all chairs, or all buildings, since the resulting set of parts would have to very large to account for the diversity within the category. Moreover, the parts are unlikely to have intuitive names, e.g. "top-right corner of the left handle".

In this section we overview methods to discover semantically aligned PnA without restricting oneself to language-based interfaces. The underlying approach is to collect annotations relative to another. Such annotations provides constraints which can be utilized to guide the alignment of the representation to semantics. We describe several examples of such approaches.



(a) Discovered poselets for person detection (b) Detection using part mixtures

Fig. 9 Visual part discovery from annotations. (a) Poselets discovered for detecting people using landmark annotations on the PASCAL VOC dataset. Figure source: Bourdev et al. [9]. (b) Detection AP using $k = 40$ mixture components based on aspect-ratio clustering, left-right clustering, and supervised shape clustering. Nose shape clusters learned by EM are shown in the bottom. Figure source: Vedaldi et al. [98].

4.1 Attribute discovery from similarity comparisons

Similarity comparisons of the form *"A is more similar to B than C"*, can be used to obtain annotations without relying on language. These can be used to transform the data into an Euclidean space that respects the similarity constrains using methods for *distance metric learning* [104, 27], *large-margin nearest neighbor* learning [103], t-STE [61], *Crowd Kernel Learning* [96], etc.

Figure 10 shows a visualization of the categories in the CUB dataset using a two-dimensional embedding learned from crowdsourced similarity comparisons between images [101]. Each image-level similarity constraint is converted to a category-level similarity constraint by using the category labels of the images from which an embedding is learned using t-STE. A group of points on the bottom-right corresponds to perching birds, while another group on the bottom-left corresponds to gull-like birds.

Since a representation learned in such manner respects the underlying perceptual similarity, it can be used as a means of interacting with a user for fine-grained recognition. Wah et al. [101] build an interface where users interactively recognize bird species by selecting the most similar image in a display. The underlying perceptual embedding is used to select the images to be displayed in each iteration. The au-

thors show that the method requires fewer questions to get to the right answer than an attribute-based interface of Branson et al. [14].

A drawback of similarity comparisons is that there can be considerable ambiguity in the task since there are many ways to compare images. Most methods for learning embeddings do not take this into account and hence are less robust to annotations collected via "crowdsouring" which can have significant noise. A number of approaches aim to reduce this ambiguity by providing additional instructions to the annotators.

The *relative attributes* approach of Parikh and Grauman [74] guides similarity comparisons by focusing on a particular describable attribute. An example annotation task is: *is A smiling more than B*, as seen in Figure 11a. Such annotations are used to learn a ranking function, or a one dimensional embedding, of images corresponding to the attribute. Relative attributes bridge the gap between categorical attributes and low-dimensional semantic embeddings, and have been used for interactive search and learning of visual attributes [75, 46].

Wah et al. [101] guide similarity comparisons by restricting the image to a *part of the object*, as seen in Figure 11b, to obtain a semantic embedding of parts. The authors use parts discovered using the *discriminative patches* approach [92], but part annotations can be used instead when available. The authors show that localized perceptual similarities provides a richer way of indicating closeness to a test image and leads to better efficiency during interactive recognition tasks.

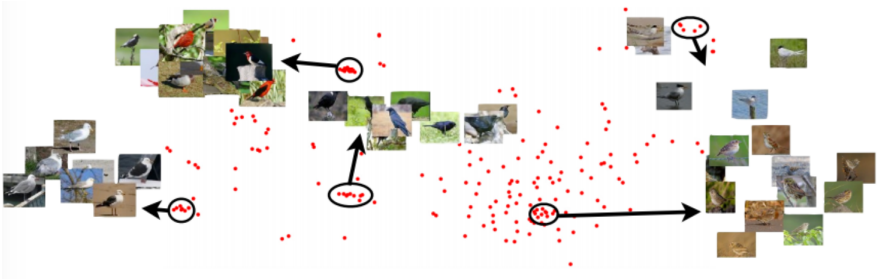


Fig. 10 A visualization of the first two dimensions of the 200-node category-level similarity embedding. Visually similar classes tend to belong to coherent clusters (circled and shown with selected representative images). Figure source: Wah et al. [101] (Best viewed digitally with zoom).

4.2 Part discovery from correspondence annotations

Traditional methods for annotating parts require a set of nameable parts. When such parts are not readily available one can instead label correspondences between pairs of instances. Maji and Shakhanrovich [65, 66] show that when annotators are asked to mark correspondences between image pairs within a category, the result is fairly

(a) Relative attributes (b) Non-localized and localized similarity

Fig. 11 (a) In the *relative attributes* framework an attribute is measured relative to other images, e.g. is the person in the image smiling more, or less, than the other images. Figure source: Parikh and Grauman [74]. (b) *Global* or *localized similarity comparisons* are used to learn a perceptual embedding of the entire object or parts respectively. Figure source: Wah et al. [102].

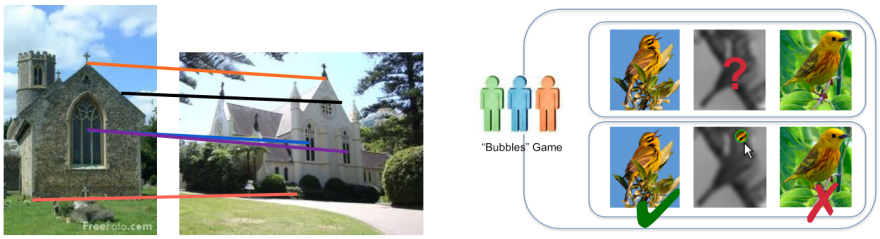
consistent across annotators, even when the names of parts are not known (Figure 12a). Annotators rely on semantics beyond visual similarity to mark correspondences – two windows are matched even though they are visually different.

Methods for part discovery that rely on appearance and geometry can be extended to take into account the pairwise constraints obtained from such correspondence annotations. The authors propose an approach were the patches corresponding to a semantic part are iteratively updated while respecting the underlying matches between image pairs. The resulting discovered patches are both visually and semantically aligned and can be used for rich part-based analysis of objects, including for detection and segmentation [66].

Another method that implicitly obtains correspondences is the *BubbleBank* approach of Deng et al. [29]. Annotators are shown two images A and B, and asked which of the two is the category of the third image (Figure 12b). The caveat is that the third image is blurry, but the user can click on parts of the image to reveal what is underneath. Since, in order to accurately recognize the category corresponding parts have to be compared such annotations reveal the salient regions or parts for a given category. These clicks are used to create the *BubbleBank* representation, a set of parts centered around the frequently clicked locations, and applied for fine-grained recognition.

5 Conclusion

The chapter summarizes the current techniques for PnA discovery by categorizing them into three broad categories. The methods described are most relevant for describing and recognizing fine-grained categories, but this is by no means a complete account of existing methods. Unsupervised part-based methods alone have a rich history and even within the DPM family methods vary on how they model part



(a) Pairwise correspondence annotations (b) The "Bubbles" game

Fig. 12 (a) Annotators click on *corresponding regions* between to indicate parts [65, 66]. (b) The *Bubbles game* shows annotators a blurry image in the middle and asks which one of the two categories, left or right, does it belong to. The user can click on a region of the blurry image to reveal what is underneath. These clicks reveal the discriminative regions within an image which is used to learn a part-based representation called the *BubblesBank*. Figure source: Deng et al. [29]

appearance and geometric relationships between parts. See Ramanan [84] for a excellent survey of classical part-based models.

Similarity, a sub-field of Human-Computer Interaction (HCI) designs "games with purpose" to annotate properties of images including attributes and part labels. A well known example is the *ESP game* [99] where a pair of annotators *independently* tag images and get rewarded only if the tags match. This make it competitive encouraging participation and reduces vandalism. Some frameworks discussed in this chapter such as pairwise correspondence for part annotations, describing the differences for attribute discovery, and the *Bubbles* game, fall into this category. For a good overview of such techniques see the lecture notes by Law and Ahn [53].

We also did not cover methods that discover the structure of objects by analyzing its motion over time. This has been well studied in *robotics* to discover the kinematic structure of articulated objects [15, 93]. Although this works best at the instance-level, the strategy has been used to discover parts within a category [88].

Finally, a number of recent works discover PnAs within the framework of deep CNNs for fine-grained recognition [12, 57, 111, 112]. Although these methods have been very successful, they bring a new set of challenges. One of them is training models for a new domain when limited labeled data is available. Factorization of the appearance using parts and attributes, either using labels provided explicitly through annotations, or implicitly in the model, continues to be the method of choice for such situations.

Acknowlegements

Subhransu Maji acknowledges a UMass Amherst startup grant, and thanks Gregory Shakhnarovich, Catherine Wah, Serge Belongie, Erik Learned-Miller, and Tsung-Yu Lin for helpful discussions.

References

1. Amadasun, M., King, R.: Textural features corresponding to textural properties. Systems, Man, and Cybernetics **19**(5) (1989) 8
2. Bajcsy, R.: Computer description of textured surfaces. Morgan Kaufmann Publishers Inc. (1973) 8
3. Bansal, A., Farhadi, A., Parikh, D.: Towards transparent systems: Semantic characterization of failure modes. In: European Conference on Computer Vision (ECCV) (2014) 2
4. Bengio, Y.: Learning deep architectures for AI. Foundations and trends in Machine Learning **2**(1), 1–127 (2009) 5
5. Berg, A.C., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005) 6
6. Berg, T., Berg, A., Shih, J.: Automatic attribute discovery and characterization from noisy web data. European Conference on Computer Vision (ECCV) (2010) 3, 9, 10
7. Bergamo, A., Torresani, L., Fitzgibbon, A.W.: Picodes: Learning a compact code for novel-category recognition. In: Neural Information Processing Systems (NIPS) (2011) 5
8. Berlin, B., Kay, P.: Basic color terms: Their universality and evolution. Univ of California Press (1991) 12
9. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: European Conference on Computer Vision (ECCV) (2010) 1, 2, 3, 11, 13
10. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: International Conference on Computer Vision (ICCV) (2011) 1, 3, 9
11. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 2
12. Branson, S., Horn, G.V., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. In: British Machine Vision Conference (BMVC) (2014) 7, 16
13. Branson, S., Van Horn, G., Wah, C., Perona, P., Belongie, S.: The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization. International Journal of Computer Vision **108**(1-2), 3–29 (2014) 2
14. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: European Conference on Computer Vision (ECCV) (2010) 2, 14
15. Broida, T., Chellappa, R.: Estimating the kinematics and structure of a rigid object from a sequence of monocular images. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 497–513 (1991) 16
16. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 1
17. Caputo, B., Hayman, E., Mallikarjuna, P.: Class-specific material categorisation. In: International Conference on Computer Vision (ICCV) (2005) 9
18. Chao, Y.W., Wang, Z., Mihalcea, R., Deng, J.: Mining semantic affordances of visual object categories. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 9
19. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: European Conference on Computer Vision (ECCV) (2012) 9
20. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., et al.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 3
21. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: International Conference on Computer Vision (ICCV) (2013) 9
22. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 9

23. Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A.: Deep filter banks for texture recognition, description, and segmentation. International Journal of Computer Vision pp. 1–30 (2016) 7
24. Csurka, G., Dance, C.R., Dan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. ECCV Workshop on Stat. Learn. in Comp. Vision (2004) 5, 7
25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005) 6, 7
26. Dana, K.J., van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real world surfaces. ACM Transactions on Graphics **18**(1), 1–34 (1999) 9
27. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: International Conference on Machine Learning (ICML) (2007) 13
28. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 5, 7
29. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2013) 15, 16
30. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 9
31. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 9, 10
32. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015) 6
33. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 2, 9
34. Felzenszwalb, P.F., Grishick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2010) 1, 2, 3, 6
35. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1), 55–79 (2005) 10, 11
36. Ferrari, V., Zisserman, A.: Learning visual attributes. In: Neural Information Processing Systems (NIPS) (2007) 9
37. Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: VLDB, vol. 99, pp. 518–529 (1999) 5
38. Girshick, R., Iandola, F., Darrell, T., Malik, J.: Deformable part models are convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 7
39. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 7
40. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007) 5
41. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.O.: On the significance of real-world conditions for material classification. European Conference on Computer Vision (ECCV) (2004) 9
42. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology **24**(6), 417 (1933) 5
43. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010) 5
44. 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) 2, 7

45. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2013) 1
46. Kovashka, A., Parikh, D., Grauman, K.: WhittleSearch: Image search with relative attribute feedback. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 2, 14
47. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on (2013) 7, 8
48. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS) (2012) 2, 3, 7
49. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(10), 1962–1977 (2011) 9
50. Kwak, I.S., Murillo, A.C., Belhumeur, P.N., Kriegman, D., Belongie, S.: From bikers to surfers: Visual recognition of urban tribes. In: British Machine Vision Conference (BMVC) (2013) 9
51. Lad, S., Parikh, D.: Interactively guiding semi-supervised clustering via attribute-based explanations. In: European Conference on Computer Vision (ECCV) (2014) 2
52. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 2, 9
53. Law, E., Ahn, L.v.: Human computation. Synthesis Lectures on Artificial Intelligence and Machine Learning **5**(3), 1–121 (2011) 16
54. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **28**(8), 2169–2178 (2005) 9
55. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2006) 6
56. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998) 3
57. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. International Conference on Computer Vision (ICCV) (2015) 7, 8, 16
58. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 9
59. Lloyd, S.P.: Least squares quantization in PCM. Information Theory, IEEE Transactions on **28**(2), 129–137 (1982) 5
60. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 7
61. van der Maaten, L., Weinberger, K.: Stochastic triplet embedding. In: Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on, pp. 1–6. IEEE (2012) 13
62. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 7
63. Maji, S.: Discovering a lexicon of parts and attributes. In: Second International Workshop on Parts and Attributes, ECCV 2012 (2012) 3, 10, 11
64. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 7, 8
65. Maji, S., Shakhnarovich, G.: Part Annotations via Pairwise Correspondence. In: 4th Workshop on Human Computation, AAAI (2012) 4, 14, 16
66. Maji, S., Shakhnarovich, G.: Part discovery from partial correspondence. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2013) 14, 15, 16
67. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995) 8

68. Ng, A.Y., Jordan, M.I., Weiss, Y., et al.: On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems **2**, 849–856 (2002) 5
69. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **24**(7), 971–987 (2002) 9
70. Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: Predicting entry-level categories. International Journal of Computer Vision pp. 1–15 9
71. Oxholm, G., Bariya, P., Nishino, K.: The scale of geometric texture. In: European Conference on Computer Vision (ECCV) 9
72. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: International Conference on Computer Vision (ICCV) (2011) 1
73. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 10, 11
74. Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision (ICCV) (2011) 14, 15
75. Parikh, D., Kovashka, A., Parkash, A., Grauman, K.: Relative attributes for enhanced human-machine communication. In: Twenty-Sixth AAAI Conference on Artificial Intelligence (2012) 14
76. Parikh, D., Zitnick, C.: Human-debugging of machines. NIPS WCSSWC **2**, 7 (2011) 2
77. Parizi, S.N., Oberlin, J.G., Felzenszwalb, P.F.: Reconfigurable models for scene recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 1
78. Parkash, A., Parikh, D.: Attributes for classifier feedback. In: European Conference on Computer Vision (ECCV) (2012) 2
79. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3
80. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3, 10, 11
81. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2007) 5
82. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: European Conference on Computer Vision (ECCV) (2010) 5
83. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 7
84. Ramanan, D.: Part-based models for finding people and estimating their pose. In: Visual Analysis of Humans, pp. 199–223. Springer (2011) 16
85. Rao, A.R., Lohse, G.L.: Towards a texture naming system: Identifying relevant dimensions of texture. Vision Research **36**(11), 1649 – 1669 (1996) 9
86. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute discovery via predictable discriminative binary codes. In: European Conference on Computer Vision (ECCV) (2012) 5
87. Razavin, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: DeepVision workshop (2014) 2, 7
88. Ross, D.A., Tarlow, D., Zemel, R.S.: Learning articulated structure and motion. International Journal of Computer Vision **88**(2), 214–237 (2010) 16
89. Sharan, L., Rosenholtz, R., Adelson, E.H.: Material perception: What can you see in a brief glance? Journal of Vision **9:784**(8) (2009) 9
90. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: British Machine Vision Conference (2013) 6
91. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Fisher networks for large-scale image classification. In: Advances in Neural Information Processing Systems (2013) 6
92. Singh, S., Gupta, A., Efros, A.: Unsupervised discovery of mid-level discriminative patches. European Conference on Computer Vision (ECCV) (2012) 1, 6, 14

93. Sturm, J.: Learning kinematic models of articulated objects. In: Approaches to Probabilistic Model Learning for Mobile Manipulation Robots, pp. 65–111. Springer (2013) 16
94. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: International Conference on Computer Vision (ICCV) (2015) 7
95. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. Systems, Man and Cybernetics, IEEE Transactions on **8**(6), 460 –473 (1978) 8
96. Tamuz, O., Liu, C., Belongie, S., Shamir, O., Kalai, A.T.: Adaptively learning the crowd kernel. In: International Conference on Machine Learning (2011) 13
97. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience **3**(1), 71–86 (1991) 5
98. Vedaldi, A., Mahendran, S., Tsogkas, S., Maji, S., Girshick, R., Kannala, J., Rahtu, E., Kokkinos, I., Blaschko, M.B., Weiss, D., Taskar, B., Simonyan, K., Saphra, N., Mohamed, S.: Understanding objects in detail with fine-grained attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1, 3, 9, 11, 12, 13
99. Von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006) 16
100. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 3, 7, 8, 9
101. Wah, C., Horn, G.V., Branson, S., Maji, S., Perona, P., Belongie, S.: Similarity comparisons for interactive fine-grained categorization. In: Computer Vision and Pattern Recognition (2014) 4, 13, 14
102. Wah, C., Maji, S., Belongie, S.: Learning localized perceptual similarity metrics for interactive categorization. In: IEEE Winter Conference on Applications of Computer Vision, WACV (2015) 15
103. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Neural Information Processing Systems (NIPS) (2006) 13
104. Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Neural Information Processing Systems (NIPS) (2002) 13
105. Xu, Y., Ji, H., Fermuller, C.: Viewpoint invariant texture description using fractal analysis. International Journal of Computer Vision **83**(1), 85–100 (2009) 9
106. Yamaguchi, K., Kiapour, M.H., Berg, T.: Paper doll parsing: Retrieving similar styles to parse clothing items. In: International Conference on Computer Vision (ICCV) (2013) 9
107. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.C.: Layered object models for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(9), 1731–1743 (2012) 1
108. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(12), 2878–2890 (2013) 11
109. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2013) 5
110. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV) (2014) 7
111. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: European Conference on Computer Vision (ECCV) (2014) 1, 7, 16
112. Zhang, N., Paluri, M., Rantazo, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 7, 16
113. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 11