Discovering a Lexicon of Parts and Attributes

Subhransu Maji

Toyota Technological Institute at Chicago,
Chicago, IL 60637, USA
smaji@ttic.edu

Abstract. We propose a framework to discover a lexicon of visual attributes that supports fine-grained visual discrimination. It consists of a novel annotation task where annotators are asked to describe differences between pairs of images. This captures the intuition that for a lexicon to be useful, it should achieve twin goals of discrimination and communication. Next, we show that such comparative text collected for many pairs of images can be analyzed to discover topics that encode *nouns* and *modifiers*, as well as *relations* that encode attributes of parts. The model also provides an ordering of attributes based on their discriminative ability, which can be used to create a shortlist of attributes to collect for a dataset. Experiments on Caltech-UCSD birds, PASCAL VOC person, and a dataset of airplanes, show that the discovered lexicon of parts and their attributes is comparable to those created by experts.

1 Introduction

A lexicon that supports fine-grained visual recognition provides an effective language-based interface for humans to query particular instances of a category. Some successful applications include searching faces with desired attributes [1], shopping websites that support structured search, etc. From the computer vision perspective, such a lexicon 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 [2,3]. However, to build such systems one requires a large dataset of images annotated with attributes. This work addresses the issue of deciding the set attributes to annotate for an object category, in order to enable fine-grained discrimination.

One may derive such a lexicon from "field guides" – books that help identify particular species of animals, birds, etc. These exist for some categories such as birds, but for a vast majority of object categories, there aren't any such sources. Moreover, even when a field guide is available, it may not be quite suited to the set of images in hand – one may have a field guide for military airplanes, but not for passenger planes or bi-planes. As we scale to thousands of object categories, it becomes desirable that the process of discovering such a lexicon be as automated as possible.

In this work we build on the intuition that a good lexicon should achieve twin goals of communication and discrimination, i.e., it should be easy to describe

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 21–30, 2012.

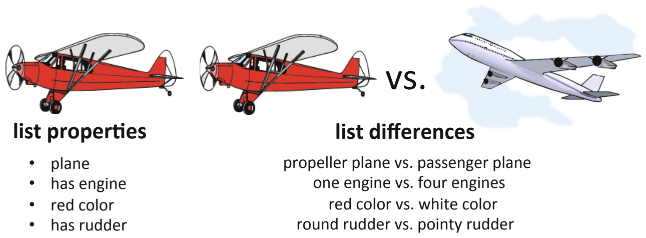

vs.
list properties
plane
has engine
red color
has rudder
list differences
propeller plane vs. passenger plane
one engine vs. four engines
red color vs. white color
round rudder vs. pointy rudder

Fig. 1. Fine-grained attributes are better revealed in the discriminative description task (right), than in the traditional description task (left)

instances, as well as sufficient to distinguish instances from one another using the lexicon. To this end we propose a novel annotation task of discriminative description, where one is asked to describe differences between pairs of images. As shown in Figure 1, in our interface, the annotator is shown a pair of images and is asked to describe in free-form English, a few differences between the two. Our annotation task can reveal properties that are more fine-grained than in the traditional annotation task of listing properties of objects, one at a time. Moreover, the frequency with which a certain attribute is used to distinguish pairs within a set provides an indication of its discriminative power. We require that the annotations be structured – each description be of the form "sentence a" vs. "sentence b", where "sentence a" and "sentence b" describe a property of the left and right images respectively. This provides us with a corpus of sentence pairs that can be analyzed to discover a lexicon of parts and attributes using Natural Language Processing (NLP) techniques.

We propose a novel generative model of sentence pairs across the corpus to extract a set of parts, *shared* topics that capture semantic properties such as "color" or "cardinality", as well as relations that encode part-specific attributes. The model also provides an ordering of these attributes based on their frequency within the corpus, which can provide a shortlist of attributes to collect for a dataset. We perform experiments using such annotations collected on Amazon Mechanical Turk [4] for images from Caltech-UCSD birds, PASCAL VOC person, and a dataset of airplane images, and show that the framework can be used to obtain a lexicons of parts and attributes that matches those created by experts.

2 Related Work

The task of discovering a lexicon of visual attributes has received some attention from the computer vision community in recent years. Berg et al. [5] use descriptions of products in shopping websites to mine phrases that appear frequently, which are sorted according to how well a computer vision algorithm can predict them. Parikh et al. [6] discover task-specific attributes with a user in the loop by considering discriminative directions in the data found automatically and asking users to name the variation along that direction. Recently Duan et al. [7] proposed a modification of the latter, which enables discovery of localized attributes such as ours. Both these approaches assume an explicit feature space where good discriminative directions can be easily found. In contrast, we aim to discover visual attributes that directly enable distinction of one instance from another independent of the features, or computer vision pipeline. This may be effective in discovering visual attributes that are otherwise hard to find without an intermediate step of part localization, which on many datasets can be quite difficult. We derive our attributes based on text collected using the proposed annotation task (Section 3), which to our knowledge has not been used previously, and rely on language modeling tools to discover attributes. We show that surface level statistics of the data derived from word alignments of such comparative text can be used to discover a lexicon of parts and attributes.

3 The Discriminative Description Task

Attributes should help to distinguish one instance from another within a category. We use this intuition to design an interface where the primary goal is to extract such attributes. Our annotation task consists of showing annotators pairs of images of the same category and asks 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 Figure 1. We also provide a few examples to guide the process. The pairwise comparison encourages annotators to list attributes that distinguish one instance from another. Thus, the lexicon that we elicit from the overall process is likely to be more specialized than what one might otherwise get by collecting properties of the instance one at a time. It also allows us to discover attributes that are relevant to the set of images in hand. For example, if all the planes in our dataset were propeller planes, we would discover attributes that distinguish propeller planes from one another.

The pairwise comparison is a general framework for collecting discriminative properties. In this work we focus on nameable parts and attributes, but with a simple modification of the interface we can also collect evidence for a property that is different by allowing the user to mark such regions in the image. This option might be more suited for categories which have un-nameable parts. One may also obtain even finer grained attributes by repeating the process for pairs of images within a sub-category. This process is natural because it is easier to list differences between objects that are similar as more parts can be put in correspondence and compared.

For a given set of images, one can sample pairs at random. The random sampling strategy 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 Discovering a Lexicon of Parts and Attributes

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 simple forms of comparative text such as those 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 "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" belong to the same semantic category, which in this case is "color". We propose a generative model of sentence pairs across the corpus that captures this structure. 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 topic may be modified by several modifier topics, and a single modifier topic may modify several noun topics. 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.

Given a corpus of sentence pairs $\mathbf{e}_s, \mathbf{f}_s$, the generative model is shown in Figure 2. Each sentence pair is generated according to an *attribute* z_s. The variable z_s encodes a bipartite relation between a *noun* and a *modifier* topic. This is enforced by modeling the topic distribution conditioned on z_s as a multinomial distribution Ω_{z_s} peaked at exactly one each of *noun* and *modifier* topics (see section 4.1). The topics for nouns and modifiers, Γ, are themselves multinomials that denote the probability of word given topic. A word in position i in the left sentence is generated by sampling a topic $t_{s,i}$ conditioned on z_s and a word conditioned on the topic. The right sentence is generated according to an alignment $\mathbf{a}$, which provides the locations of the words in $\mathbf{e}_s$ that generated $\mathbf{f}_s$. Each word in $\mathbf{f}_s$ is generated from the word in $\mathbf{e}_s$ and its topic, at the location given by $\mathbf{a}$, using a topic-specific multinomial Ψ. Let I_s and J_s, denote the lengths of the s^{th} left and right sentences respectively. Then the joint probability of the sentences and latent variables given the parameters $\Theta = \{\theta, \pi, \Omega, \Gamma, \Psi\}$ is given by:

$$\begin{aligned} P(z_s, \mathbf{e}_s, \mathbf{f}_s, \mathbf{a}_s, \mathbf{t}_s|\Theta) = \quad & \prod_{s=1}^{N} P(z_s|\theta) \prod_{i=1}^{I_i} P(t_{s,i}|z_s, \Omega) P(e_{s,i}|t_{s,i}, \Gamma) \\ & \times \prod_{j=1}^{J_i} P\left(a_j|\pi\right) P\left(f_{s,j}|e_{s,a_j}, t_{s,a_j}, \Psi\right) \end{aligned}$$

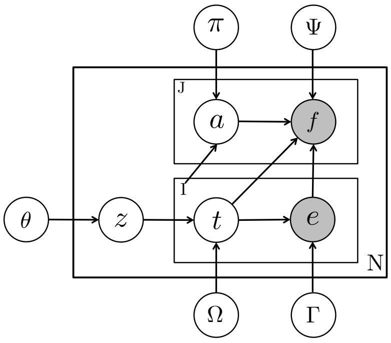
The generative process for the right sentence $P(\mathbf{f}_s|\mathbf{e}_s, \mathbf{t}_s)$, is similar to the IBM word alignment model [8], popular in machine translation to initialize translation tables across a pair of languages. The starting point for these models is a corpus of sentence pairs in two languages, say *French* and *English*, which are translations of one another. In the simplest IBM model, each word in the French sentence is generated independently from a word in the English sentence according to an alignment vector $\mathbf{a}_s = a_{s,1}, a_{s,2}, \ldots, a_{s,J_s}$, denoting the positions of the word(s) in the English sentence that generated each French word. The joint probability of the alignment vector and the French sentence given the English sentence is:

$$P(\mathbf{f}_s, \mathbf{a}_s | \mathbf{e}_s, I_s, J_s) = \prod_{j=1}^{J_s} P(a_{s,j} | I_s, J_s) P(f_{s,j} | e_{s,a_{s,j}}) \quad (1)$$

Each entry in the alignment vector $a_{s,j} \in \{1, \ldots, I_s\}$ is either chosen uniformly at random (IBM 1), or proportional to $\pi(|a_{s,j} - j|)$ (IBM 2). The distribution π is peaked at 0, which encourages words in the same position across a sentence pair to be aligned to one another.

Compared to the IBM models, we have also introduced topics in the source language, which can enable topic-specific word emissions. We also do not model the "NULL" word commonly used in these models since the source and target languages are the same in our setting. This model is related to the BiTAM model proposed by [9], which also models type-specific translations. The difference, however, is that our types (or topics) are attached to the words in the source language and *not* to the alignments.

The generative process for the left sentence is similar to LDA [10] and to its variants such as Correspondence-LDA [11]. The main difference lies in how we model the topics themselves. Our topics encode *nouns* and *modifiers*, and are estimated by relying on the structure of the task to cluster words (described in Section 4.1). In addition the topic proportions in each sentence are constrained to be *bipartite* and *not* drawn from a Dirichlet distribution, which allows us to model part-modifier topic correlations, and thereby discover the relations between parts and their attributes.



For each sentence pair $\mathbf{e}_s, \mathbf{f}_s$, $s \in \{1, \ldots, N\}$

- Sample relation $z_s \sim Multinomial(\theta)$
- For each word position $i \in 1, \ldots, I_s$ in $\mathbf{e_s}$
  - Sample topic $t_{s,i} \sim Multinomial\left(\Omega_{z_s}\right)$
  - Sample word $e_{s,i} \sim Multinomial\left(\Gamma_{t_{s,i}}\right)$
- For each word position $j = \{1, \ldots, J_s\}$ in $\mathbf{f_s}$
  - Sample $a_j \in \{1, \ldots, I\} \propto \pi(|a_j - j|)$,
  - Sample word $f_{s,j} \sim Multinomial\left(\Psi_{e_{a_j}, t_{a_j}}\right)$

Fig. 2. The generative model of the corpus consisting of sentence pairs $\{\mathbf{e}_s, \mathbf{f}_s\}$

4.1 Initialization and Parameter Estimation

We estimate all the parameters by maximizing the likelihood of the data using EM. Given an estimate of the parameters $\{z_s, \mathbf{t}_s, \mathbf{a}_s, \theta, \pi, \Omega, \Gamma, \Psi\}$, we update one parameter at a time, keeping others fixed, until convergence. Our model has many parameters, and careful initialization is necessary to obtain a good solution. Fortunately, in our case we can initialize the parameters of our model from simpler models which can be easily solved.

Initializing $\boldsymbol{\pi}, \mathbf{a}_{\boldsymbol{s}}, \boldsymbol{\Psi}$. We use the IBM model 2 in Equation 1 to learn the distortion probabilities π^{ibm} and topic independent Ψ^{ibm}. The alignments are initialized to the most likely ones according to the IBM model. We initialize $\pi \leftarrow \pi^{ibm}$ and $\Psi_{e,f,t} = \Psi^{ibm}_{e,f}$.

Initializing $\boldsymbol{\Gamma}$: *noun* and *modifier* Topics. This is a crucial step in our approach. In our application, we wish to learn topics that encode parts and attributes. A possible way of doing this would be to use language-specific knowledge such as part-of-speech tags. However, automatic tagging may not be accurate for words in a new domain, limiting its applicability. We instead use the observation made earlier in this section to initialize noun and modifier topics, i.e., words that repeat across pairs are likely to be from the same noun topic, whereas words that are different are likely to be from the same modifier topic. This provides a domain and perhaps even a language-independent way of characterizing nouns and their modifiers.

To initialize noun topics, we find words i with $f(i) > \tau_p$ and $\Psi_{i,i} > \rho_p$, where $f(i)$ is the number of times the word i appears in the corpus. $\Psi_{i,i}$ characterizes how frequently the word i aligns to itself across sentence pairs. We set $\tau_p = 5$ and $\rho_p = 0.6$, which gives us an initial set of noun topics.

To initialize modifier topics, we construct a similarity matrix between pairs of words $S(i,j) = f(i) \times \Psi^{ibm}_{i,j}$. The matrix S counts the number of times word j is aligned to the word i across the corpus. We zero out the rows and columns corresponding to the words already taken in the earlier step. We find the connected components of the graph G, where nodes i, j are connected if $S(i,j) > \tau_a$ and $\Psi_{i,j} > \rho_a$. Connected components of size at least two are assigned a modifier topic. The remaining words are assigned to a "COMMON" topic. We set $\tau_a = 10$ and $\rho_a = 0.2$ in all our experiments. The parameters ρ and τ control the number of noun and modifier topics desired. Setting τ or ρ higher would result in fewer topics. We initialize topic probabilities:

$$P(topic = t|word = i) = \frac{\sum_{j \in C_t} \Psi^{ibm}(j|i)}{\sum_t \sum_{j \in C_t} \Psi^{ibm}(j|i)} \tag{2}$$

Where C_t is the set of words in topic t, which can then be used to initialize Γ. As we will see in the experiments section, the process discovers semantic modifier categories such as *color*=\{*red, blue, green,...*\}, *cardinality*=\{*one, two,...*\}, etc., that capture the kinds of variation each part is likely to have.

Initializing $\boldsymbol{\Omega}, \boldsymbol{z_s}, \boldsymbol{\theta}$: Bipartite Relations between *nouns* and *modifiers*. The previous step discovers n noun and m modifier topics. We consider relations between all possible choices of noun and modifier topics. In addition, we also consider m relations which correspond to attributes of special part "GLOBAL".

This is to encode relations that describe global properties of the object, such as "gender" for the person category. This leads to a total of $(n+1) \times m$ relations.

For each relation, we initialize a multinomial peaked at the corresponding noun and modifier topics. In addition, the multinomial is allowed a fixed portion of the "COMMON" topic to model words in the sentence that are not from either part and attribute topic. The rest of the values are assigned a small constant value. Given this initialization, we run EM to estimate a mixture of multinomials and memberships. At each M step we ensure that the relations remain bipartite – multinomials are assigned small values for topics other than the ones in the relation and renormalized.

After EM converges, we drop clusters that have fewer than τ members, which is set to 1% of the data in our experiments. We run EM again with the remaining clusters until convergence and initialize z_s, θ, Ω to the values of cluster memberships, frequencies, and estimated multinomial mixture means, respectively.

5 Experiments

We experiment with images from Caltech-UCSD birds, PASCAL VOC person, and a dataset of airplane images. These categories are diverse and contain instances with different attributes. Our experiments were performed on Amazon Mechanical Turk [4] using the interface described in Section 3. For each category, we also provided a few examples of annotations while clearly indicating that the list is not exhaustive. Figure 3 shows example annotations collected using our interface overlaid on the images. Annotators provide natural language descriptions that differentiate the two images. By pairing the same image with others, different properties of the image can be revealed. Thus, our approach may also be used to discover instance-specific attributes that discriminate the instance from others.

The collected annotations can be noisy. These include formatting errors, e.g., empty sentences or sentences without the word "vs." for separation. There is also noise due to different ways of spelling the same word, synonyms, etc. Ignoring sentences with formatting errors typically leaves about $80-85\%$ of the sentences.

5.1 Caltech-UCSD Birds

The dataset [12] consists of 200 species of birds and was introduced for fine-grained visual category recognition. We sample 200 images, one random image from each category for our discovery process. For these images we sampled 1600 pairs uniformly at random and collected annotations.

Figure 4(a), shows the learned topics and attributes for birds category. The learned parts for each category are shown on the top row, modifiers on the bottom row, and the bipartite relation between parts and attributes is shown using edges connecting them. The thickness of the edge indicates the frequency of the relation in the dataset. The discovered parts and modifiers correctly 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 $\{tail\} \leftrightarrow \{long, short, small, ..\}$, i.e. the size of the tail. Other distinguishing features are colors of various body parts such as body, tail and head, and beak shape, such as pointy vs. round, etc. An interesting relation that is discovered is $\{like\} \leftrightarrow \{sparrow, duck, crow, eagle, dove, \ldots\}$. Even though we had 200 species of birds, the annotators choose to describe each bird based on their similarity to a commonly-seen set of birds. Similarity to prototypes is a discriminative visual attribute for birds 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 [12]. Out of the 12 parts of birds, which are *forehead, crown, bill, eye, throat, nape, breast, back, wing, belly, leg* and *tail*, we discover 6 of them. We miss parts such as crown and nape, which are sub-parts of the head region and unfamiliar to non-experts. This brings up an important aspect of the problem which is that the lexicon familiar to experts may be quite different from that of annotators commonly available on crowdsourcing platforms such as Amazon Mechanical Turk [4]. Nevertheless, the pairwise comparison is an attractive framework to obtain such lexicons regardless of the expertise of the annotator.

5.2 Airplanes

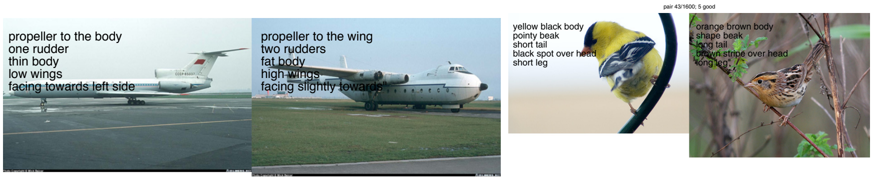
We collect 200 images from a website of airplane photographs[1]. We sampled 1000 pairs uniformly at random and collected annotations.

Figure 4(b), shows the discovered attributes. Here, the most frequent attribute is the color of the rudder. Our dataset has many passenger planes, and they all tend to have different rudder colors corresponding to different airlines. The number of wheels is the second most distinguishing feature. This actually corresponds to sentence pairs "one front wheel vs. two front wheels", which distinguishes *propeller* and other smaller planes from bigger jet planes. The correct relation is $\{front\ wheel\} \leftrightarrow \{one, two\}$. We instead discover two relations $\{front\} \leftrightarrow \{one\}$ and $\{wheel\} \leftrightarrow \{one\}$, because we don't model phrases. The next most important relation is the facing direction, which roughly distinguishes one half of the dataset from another. Other discovered relations include 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 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.

5.3 PASCAL VOC Person

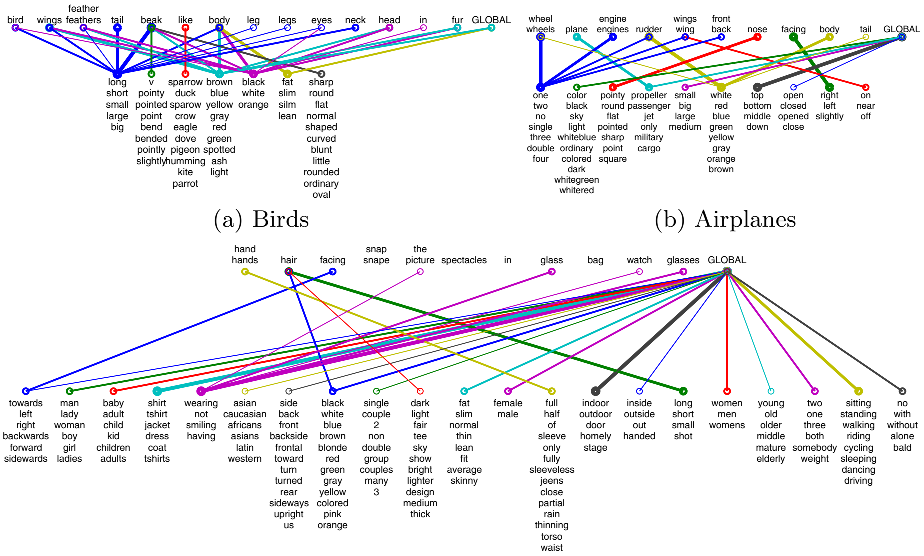
A dataset consisting of attributes of people from the PASCAL Visual Object Challenge (VOC) dataset was introduced by Bourdev et al. [3]. We sample 400

[1] `http://airliners.net`



Airplanes Birds

Fig. 3. Example annotations collected using our interface.


bird
wings
feather feathers
tail
beak
like
body
leg
legs
eyes
neck
head
in
fur
GLOBAL
long short small large big
v pointy pointed point bend bended pointly slightly
sparrow duck sparow crow eagle dove pigeon humming kite parrot
brown blue yellow gray red green spotted ash light
black white orange
fat slim silm lean
sharp round flat normal shaped curved blunt little rounded ordinary oval
wheel wheels
plane
engine engines
rudder
wings wing
front back
nose
facing
body
tail
GLOBAL
one two no single three double four
color black sky light whiteblue ordinary colored dark whitegreen whitered
pointy round flat pointed sharp point square
propeller passenger jet only military cargo
small big large medium
white red blue green yellow gray orange brown
top bottom middle down
open closed opened close
right left slightly
on near off
hand hands
hair
facing
snap snape
the picture
spectacles
in
glass
bag
watch
glasses
GLOBAL
towards left right backwards forward sidewards
man lady woman boy girl ladies
baby adult child kid children adults
shirt tshirt jacket dress coat tshirts
wearing not smiling having
asian caucasian africans asians latin western
side back front backside frontal toward turn turned rear sideways upright us
black white blue brown blonde red green gray yellow colored pink orange
single couple 2 non double group couples many 3
dark light fair tee sky show bright lighter design medium thick
fat slim normal thin lean fit average skinny
female male
full half of sleeve only fully sleeveless jeens close partial rain thinning torso waist
indoor outdoor door homely stage
inside outside out handed
long short small shot
women men womens
young old older middle mature elderly
two one three both somebody weight
sitting standing walking riding cycling sleeping dancing driving
no with without alone bald

(a) Birds (b) Airplanes

(c) Person

Fig. 4. Learned *parts* (top row), *modifiers* (bottom row), and *attributes* (edges) for birds, airplanes and person category. The thickness of the edge is proportional to the frequency of the attribute in the corpus.

random images from the training/validation subset of the dataset. For these images we sampled 1600 pairs uniformly at random and collected annotations.

Figure 4(c) shows the discovered attributes for this dataset. We discover attributes such as *gender, hair style, hair length, dress type, wearing glasses, hats, etc* which are present in [3]. In addition, we discover new ones such as the action being performed – sitting, standing, dancing, etc.

6 Conclusion

We propose a framework for discovering a lexicon of fine-grained visual attributes of object categories that achieves the twin goals of communication and discrimination. We show that text generated from pairwise comparisons of instances within object categories provides a rich source of attributes that are discriminative and task-specific by design. We also propose a generative model of the sentence pairs and show that it discovers topics corresponding to parts and modifiers and relations between them on three challenging datasets.

Although in this task we focus on the lexicon aspect, the same interface can be modified so that the user can provide evidence of the difference, for example, by drawing bounding boxes around the region of interest. This can enable discovery of visual parts which are otherwise hard to name. The pairwise comparison framework may be used to discover attributes in a coarse-to-fine manner by recursively applying the framework within each sub-category.

Acknowledgments. The author was partially supported by a 2012 JHU Summer Workshop. The airplane images are courtesy Mick Bajcar and http://airliners.net.

References

1. Kumar, N., Belhumeur, P., Nayar, S.: FaceTracer: A Search Engine for Large Collections of Images with Faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008)
2. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: CVPR (2010)
3. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV (2011)
4. MTurk: Amazon mechanical turk, http://www.mturk.com
5. Berg, T.L., Berg, A.C., Shih, J.: Automatic Attribute Discovery and Characterization from Noisy Web Data. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 663–676. Springer, Heidelberg (2010)
6. Parikh, D., Grauman, K.: Interactive discovery of task-specific nameable attributes. In: Workshop on Fine-Grained Visual Categorization, CVPR (2011)
7. Duan, K., Bloomington, I., Parikh, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: CVPR (2012)
8. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Comput. Linguist. (1990)
9. Zhao, B., Xing, E.P.: Bitam: bilingual topic admixture models for word alignment. In: COLING-ACL (2006)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. (2003)
11. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR (2003)
12. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology (2010)