Additional Remarks on Designing Category-Level Attributes for Discriminative Visual Recognition*

Felix X. Yu†, Liangliang Cao§, Rogerio S. Feris§, John R. Smith§, Shih-Fu Chang†
† Columbia University § IBM T. J. Watson Research Center
{yuxinnan@ee, sfchang@cs}.columbia.edu {liangliang.cao, rsferis, jsmith}@us.ibm.edu

Abstract

This is the supplementary material for *Designing Category-Level Attributes for Discriminative Visual Recognition* [3]. We first provide an overview of the proposed approach in Section 1. The proof of the theorem is shown in Section 2. Additional remarks of the proposed attribute design algorithm are provided in Section 3. We show additional experiments and applications of the designed attributes for zero-shot learning and video event modeling in Section 4. Finally, we discuss the semantic aspects of automatic attribute design in Section 5. All the figures in this technical report are best viewed in color.

1 Overview of the Proposed Approach

Figure 1 provides an overview of the proposed approach.

In the offline phase, given a set of images with labels of pre-defined categories (a multiclass dataset), our approach automatically learns a category-attribute matrix, to *define* the category-level attributes. Then a set of attribute classifiers are learned based on the defined attributes (not shown in the figure). Unlike the previous work [2], in which both the attributes and the category-attribute matrix are pre-defined (as in the "manually defined attributes"), the proposed process is fully automatic.

In the online phase, given an image from the novel categories, we can compute the designed category-level attributes. The computed values of three attributes (colored

*Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

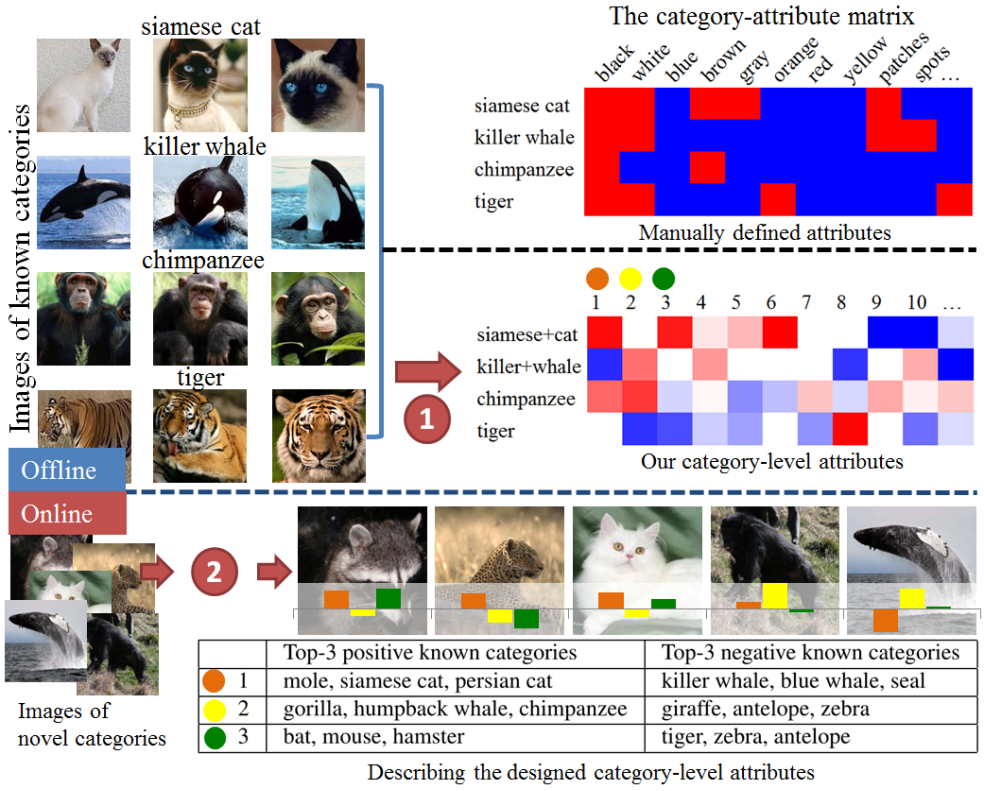


Figure 1: Overview of the proposed approach. ①: Designing the category-attribute matrix. ②: Computing the attributes for images of novel categories.

as orange, yellow and green) are shown in Figure 1. For example, the first image (of a raccoon) has positive responses of the orange and green attributes, and negative response of the yellow attribute. Because the category-level attributes are defined based on a category-attribute matrix, they can be interpreted as the relative associations with the pre-defined categories. For example, the orange attribute has positive associations with mole and siamese cat, and negative associations with killer whale and blue whale.

The category-level attributes are more intuitive than mid-level representations defined on low-level features. In fact, our attributes can be seen as soft groupings of categories, with analogy to the idea of building taxonomy or concept hierarchy in the library science. We will further discuss the semantic aspects of the proposed method in Section 5.

2 Supplementary of the Learning Framework

2.1 Proof of Theorem 1

Theorem 1. *The empirical error of multi-class classification is upper bounded by* $2\epsilon/\rho$.

Proof. Given example $(\mathbf{x}, y)$, we assume the example is misclassified as some category $z \neq y$, meaning that

$$\| \mathbf{A}_{y\cdot} - \mathbf{f}(\mathbf{x}) \| > \| \mathbf{A}_{z\cdot} - \mathbf{f}(\mathbf{x}) \| . \tag{1}$$

Then

$$\| \mathbf{A}_{y\cdot} - \mathbf{f}(\mathbf{x}) \| > \frac{\| \mathbf{A}_{y\cdot} - \mathbf{f}(\mathbf{x}) \| + \| \mathbf{A}_{z\cdot} - \mathbf{f}(\mathbf{x}) \|}{2}. \tag{2}$$

From triangle inequality and the definition of ρ:

$$\| \mathbf{A}_{y\cdot} - \mathbf{f}(\mathbf{x}) \| + \| \mathbf{A}_{z\cdot} - \mathbf{f}(\mathbf{x}) \| \geq \| \mathbf{A}_{y\cdot} - \mathbf{A}_{z\cdot} \| \geq \rho. \tag{3}$$

So we know misclassifying $(\mathbf{x}, y)$ implies that

$$\| \mathbf{A}_{y\cdot} - \mathbf{f}(\mathbf{x}) \| > \frac{\rho}{2}. \tag{4}$$

Therefore given m examples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)$, the number of category recognition mistakes we make is at most

$$\frac{\sum_{i=1}^{m} \| \mathbf{A}_{y_i\cdot} - \mathbf{f}(\mathbf{x}_i) \|}{\rho/2} = \frac{2m\epsilon}{\rho}. \tag{5}$$

Thus the empirical error is upper bounded by $2\epsilon/\rho$. □

3 Supplementary of the Algorithm

3.1 Parameters of the Algorithm

There are two parameters in the attribute design algorithm, λ and η. Larger λ means smaller ρ for the category-attribute matrix, and larger η means less redundancy r for the designed attributes. Figure 2 visualizes the influence of the parameters based on a randomly generated visual proximity matrix $\mathbf{S}$.

4 Supplementary of the Experiments

4.1 Zero-shot Learning

Figure 3 visualizes the averaged similarity matrix based on the results of 10 users.

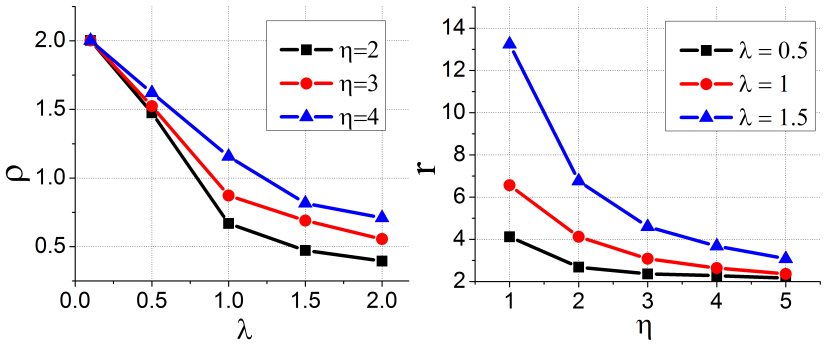


Figure 2: The influence of the two parameters. Left: the influence of λ: larger λ means smaller ρ for the category-attribute matrix. Right: the influence of η: larger η means less redundancy r for the designed attributes. The visual proximity matrix $\mathbf{S}$ used for this figure is a 50×50 randomly generated non-negative symmetric matrix.

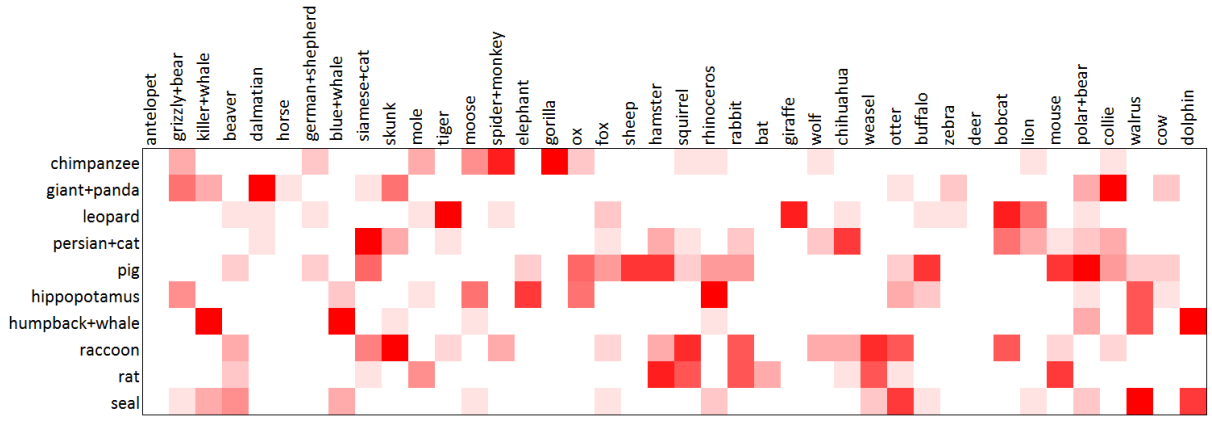


Figure 3: The manually built visual similarity matrix. It characterize the visual similarity of the 10 novel categories and the 40 known categories. This matrix is obtained by averaging the similarity matrices built by 10 different users. Each user is asked to build a visual similarity matrix, by selecting 5 most visually similar known categories for each novel category. The selected elements will be set as 1, and others as 0.

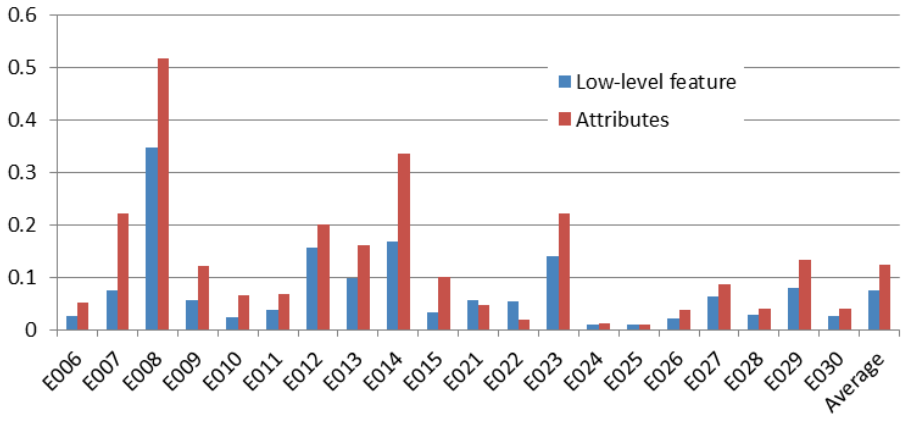


Figure 4: Average Precision results base on low-level feature and attributes for full exemplar task of TRECVID MED 2012. The results are evaluated on the internal threshold split containing 20% of the training data. Linear SVMs are used for event modeling. The same low-level features are used for training attributes.

4.2 Designing Attributes for Video Event Modeling

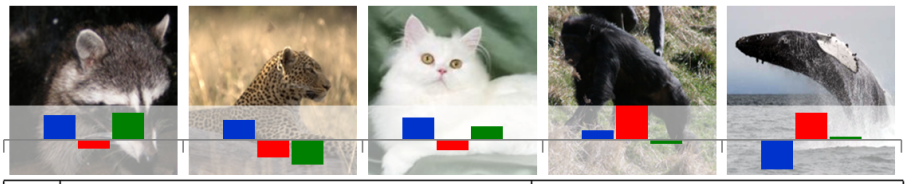
We show one additional application of using attributes for video event classification on the TRECVID 2012 MED task.

Traditionally, the semantic features for video event modeling are learned from the taxonomy with the labeled images [1]. The taxonomy is manually defined based on expert knowledge, and a set of images must be labeled by human experts. Similar to the manually specified attributes, the semantic features suffer from the following problems.

- The human defining and labeling processes are very expensive, especially if we need large-amount of concepts, with enough clean training data.
- Though the taxonomy is semantically plausible, it may not be consistent to the visual feature distributions. Consequently, some dimensions of the semantic feature vector are difficult to be modeled.

Motivated by the above facts, we use the proposed category-level attributes as a data-consistent way of modeling "semantics". Specifically, we design attributes based on 518 leaf nodes of the taxonomy [1] (as the known categories).

To test the performance of the proposed approach, we have trained and extracted 2,500 dimensional attribute feature for the pre-specified task of MED. Figure 4 shows the performance of the low-level feature and the proposed attribute feature. Impressively, attributes have achieved relative performance gain over 60%, improving the mAP from 0.075 to 0.123.



	Positive	Negative
■	mole, siamese cat, persian cat	killer whale, blue whale, seal
■	gorilla, humpback whale, chimpanzee	giraffe, antelope, zebra
■	bat, mouse, hamster	tiger, zebra, antelope

Figure 5: Using category-level attributes to describe images of novel categories. In the table below, three attributes are described in terms of the corresponding top positive/negative known categories in the category-attribute matrix. Some designed attributes can be further interpreted by concise names: the first two can be described as small land animals *vs.* ocean animals, black *vs.* non/partial-black. Some may not be interpreted concisely: the third one is like like rodent *vs.* tiger and cloven hoof animals. The figure above shows the computed attribute values for images of novel categories.

5 Discussions about Semantics

5.1 Interpretations of the Category-Level Attributes

One unique advantage of the designed attributes is that they can provide interpretable cues for visualizing the machine reasoning process. In other words, the designed attributes can be used to answer not only "what", but also "why" one image is recognized as certain category. First, the attributes are designed on category level, the descriptions are readily available through weighted categories names (*e.g.*, the attribute that has high association with polar bear, and low association with walrus, lion). Second, the regularization term $J_2(\mathbf{A})$ in the attribute design formulation can in fact lead to human interpretable attributes, by inducing "similar" categories NOT to be far away in attribute space.

Some examples of using the computed attributes to describe the images of novel categories are shown in Figure 5.

5.2 Designing Semantic Attributes

Note that not all attributes designed can be semantically interpreted. We discuss one possible way of enhancing the semantics in the attribute designing process, with the help of human interactions.

The solution is to modify the attribute design algorithm with an additional *semantic projection* step: after getting each $\mathbf{a}$ (a column of the category-attribute matrix), make some changes to $\mathbf{a}$ to get $\varphi(\mathbf{a})$, such that $\varphi(\mathbf{a})$ is semantically meaningful. Figure 6

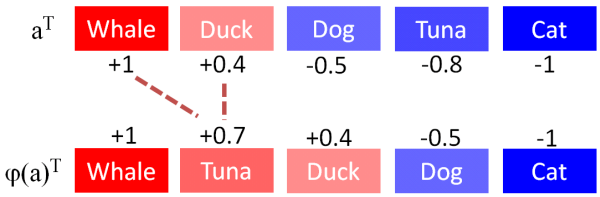


Figure 6: Semantic projection for designing attributes with stronger semantics.

shows an example of semantic projection (by human). In this example, by changing $\mathbf{a}$ to $\varphi(\mathbf{a})$, the designed attribute can be easily interpreted as "water dwelling animals".

Specifically, given an initial pool of pre-defined attributes, together with their manually specified category-attribute matrix, we can define some rules of *what* kinds of category-level attributes are semantically meaningful. For instance, it is intuitive to say *the union (black or white), intersection (black and white), and subset (chimpanzee kinds of black, attributes are often category-dependent) of the manually defined attributes are semantically interpretable.* The operations of union, intersection *etc.* can be modeled by operations on the manually specified category-attribute matrix. The designed attributes can then be projected to the nearest semantic candidate:

$$\varphi(\mathbf{a}) = \arg\min_{\mathbf{a}' \in \mathcal{A}} \parallel \mathbf{a}' - \mathbf{a} \parallel, \tag{6}$$

in which $\mathcal{A}$ is the semantic space defined by rules. This method can be used to efficiently *expand* the predefined semantic attributes. We will study this in our future work.

References

[1] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, J. Smith, and F. Yu. IBM Research and Columbia University TRECVID-2012 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), and Semantic Indexing (SIN) Systems. In *NIST TRECVID Workshop, Gaithersburg, MD*, December, 2012.

[2] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[3] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.