Preface

Visual attributes are generally defined as mid-level semantic visual concepts or properties that are shared across categories, e.g., furry, striped, metallic, young. They have recently gained significant popularity in computer vision, finding applications in zero-shot classification (where a machine can recognize a concept even without having seen it before), image ranking and retrieval, fine-grained categorization, human-machine interaction, and many others.

This book provides an overview of and summarizes recent advances in machine learning and computer vision related to visual attributes, while exploring the intersection with other disciplines such as computational linguistics and human-machine interaction. It contains a collection of chapters written by world-renowned scientists, covering theoretical aspects of visual attribute learning as well as practical computer vision applications.

We would like to express our sincere gratitude to all chapter contributors for their dedication and high-quality work, as well as to Simon Rees and Wayne Wheeler from Springer for their support and help throughout the book's preparation.

Yorktown Heights, Klosterneuburg, Blacksburg
September 2016

Rogerio Feris
Christoph Lampert
Devi Parikh

Contents

Chapter 1
Introduction to Visual Attributes

Rogerio Feris, Christoph Lampert, and Devi Parikh

Visual recognition has significantly advanced in recent years, particularly through the widespread adoption of deep convolutional neural networks [22, 28] as the main tool for solving computer vision problems. The recognition accuracy recently obtained in standard benchmark datasets, such as Imagenet [7], has even surpassed human-level performance [15].

The fuel to power up these neural network models is training data. In fact, current methods often require at least thousands of manually annotated training examples for learning robust classifiers for new categories. While it is easy to obtain a large number of example images for common categories, such as images of vehicles or dogs, it is not straighforward to obtain annotated training sets for other infrequent categories, such as a particular vehicle model or a specific dog breed. There are tens of thousands of basic categories in the world (and significantly more subordinate categories) [3]. For many of them, only a few or *no examples at all* are available.

Zero-data or zero-shot classification refers to the problem of recognizing categories for which no training examples are available [26, 30]. This problem happens in many practical settings. As an example, for the task of predicting concrete nouns from neural imaging data [30], many nouns may not have corresponding neural image examples because of the costly label acquisition process. In the visual surveillance domain, while conducting a criminal investigation, the police may have only eyewitness descriptions available for searching a targeted suspect, instead of example images [13, 40]. Many *fine-grained* visual categorization tasks have classes for

Rogerio Feris
IBM T. J. Watson Research Center,
e-mail: rsferis@us.ibm.com

Christoph Lampert
Institute of Science and Technology Austria,
e-mail: chl@ist.ac.at

Devi Parikh
Virginia Tech,
e-mail: parikh@vt.edu

which only a few or no training images exist. For instance, the ImageNet dataset has 30 mushroom synsets, each with 1000 images, whereas there are more than ten thousand mushroom species found in nature. The zero-shot classification problem is also common in other fields. In large vocabulary speech recognition systems, it is infeasible to acquire training samples for each word. Recommender systems face issues when new apps are released without any user ratings (also known as the cold-start problem [35]).

Visual attributes, which are generally defined as mid-level semantic properties that are shared across categories (e.g., furry, yellow, four-legged), provide an effective way of solving the zero-shot classification problem. As initially demonstrated by Lampert et al. [25, 26], a novel unseen category with an associated description based on semantic attributes (either provided by experts or mined from language sources, such as Wikipedia [33, 34]) can be recognized by leveraging visual attribute classifiers, which can be learned using existing training data from known categories. This process is aligned with human capabilities of identifying objects only based on descriptions. For example, when given a sentence like "large gray animals with long trunks", we can reliably identify elephants [26]. Currently, the highest-performing methods for zero-shot learning rely on visual attributes, often in connection with other forms of semantic embedding such as distributional word vector representations [1, 2, 14, 33].

Visual attributes are both semantic (human-understandable) and visual (machine-detectable). In addition to zero-shot learning, they have proven effective in various other applications. As a communication channel between humans and machines, attributes have been used for interactive recognition of fine-grained categories [4], active learning [21], and image search with humans in the loop [20]. Attributes discretize a high dimensional feature space into a simple and readily interpretable representation that can be used to explain machine decisions to humans [16] and predict user annoyance [5]. Conversely, humans can provide rationales to machines as a stronger form of supervision through visual attributes [10]. Along this direction, attributes can serve as a form of privileged information [36] for improving recognition, especially when only a few training examples are available.

Another area in which attributes have recently played a major role is visual analysis of people. In the visual surveillance domain, state-of-the-art person re-identification systems [27, 37, 39] benefit from human attributes as features for improving matching of people across cameras. The extraction of face and clothing attributes enable search for suspects or missing people based on their physical description [13, 40]. In e-commerce applications, attributes are very effective in improving clothing retrieval [17] and fashion recommendation [29]. It has also been shown that facial attribute prediction is helpful as an auxiliary task for improving face detection [42] and face alignment [43]. Methods for image ranking and retrieval also benefit from attributes as a compact and semantic image representation [11, 23, 38].

Other applications of visual attributes include describing unfamiliar objects [12], scene analysis [32], material classification [6], and image virality prediction [8]. Beyond semantics, attributes have been used for understanding and predicting the

memorability and aesthetics of photographs [9, 18, 19]. Finally, attributes have been recently used for image editing (e.g., allowing users to adjust the attributes of a scene to be "snowy" or "sunset") [24] and for conditional image generation in the context of generative adversarial networks [41].

This book's goal is to summarize the main ideas related to visual attributes that were proposed in the past few years, and to cover recent research efforts related to this emerging area in an accessible manner to a wider research community. Next, we provide an overview of the chapters of the book, which comprise both theoretical aspects of attribute learning and practical applications.

1.1 Overview of the chapters

Part I: Attribute-based Recognition

The first part of the book covers attribute-based methods for *recognition of unseen classes* for which training examples are unavailable (i.e., zero-shot classification), *recognition of seen classes*, where attributes are used as privileged information during the training stage, and methods for *multi-task attribute learning*.

Chapter 2, by Bernardino Romera-Paredes and Philip H. S. Torr, introduces the problem of zero-shot learning and proposes a general framework that models the relationships between features, attributes, and classes, so the knowledge learned at the training stage can be transferred to the inference stage. The method is easily implemented: one line of code for training and another for inference; yet, it achieves impressive results on standard benchmark datasets.

In Chapter 3, Viktoriia Sharmanska and Novi Quadrianto consider the problem of visual recognition of categories when their attributes are used as privileged information during training time. In particular, they address whether attributes are still useful privileged data when modern deep convolutional features are used for visual classification. Their analysis shows that the answer to this question depends on the classification task's complexity.

In Chapter 4, Chao-Yeh Chen, Dinesh Jayaraman, Fei Sha, and Kristen Grauman address the problem of multi-task attribute learning, exploring when and to what extent sharing is useful for attribute learning. They introduce the idea of selective sharing during multitask learning of attributes, using semantic knowledge to decide what to share and what not to share during learning.

Part II: Relative Attributes and their Application to Image Search

The second part of the book introduces the concept of relative attributes [31], which consists of measuring the relative strength of properties (for example, "bears are furrier than giraffes") instead of simply determining whether they are present or not, and demonstrates the effectiveness of modeling relative attributes in image search applications.

In Chapter 5, Adriana Kovashka and Kristen Grauman show how semantic attributes can be effectively used for interactive image search with user feedback based on relative attribute comparisons. They present a system called "Whittle-Search", which can answer queries such as "show me shoes like these, but more formal". This chapter also covers techniques for actively selecting images for feedback and adapting attribute models for personalized user queries.

Chapter 6, by Aron Yu and Kristen Grauman, addresses the problem of fine-grained visual comparisons with attributes, which is valuable for sophisticated image search systems that may need to distinguish subtle properties between highly similar images. They develop computational models based on *local learning* for fine-grained visual comparisons, where a predictive model is trained on the fly using only the data most relevant to a given input. They also address the problem of determining when an image pair is indistinguishable in terms of a given attribute.

In Chapter 7, Fanyi Xiao and Yong Jae Lee introduce a weakly-supervised method for automatically discovering the spatial extent of relative attributes in images. This is achieved by mining a set of local, transitive connections ("visual chains") that establish correspondences between the same object parts across images. They show that the proposed localized approach better models relative attributes than baselines that either use global appearance features or stronger supervision.

Part III: Describing People based on Attributes

Automatically describing people based on their fine-grained semantic attributes is important for many application domains, such as visual surveillance and e-commerce. The third part of the book covers state-of-the-art methods for estimation of human attributes and their use in different applications.

Chapter 8, by Chen Change Loy, Ping Luo, and Chen Huang, presents recent progress and cutting-edge methods based on deep learning for solving problems in estimating facial attributes such as gender, age, presence of facial hair, eyewear, hairstyle, and others. They cover approaches for handling class imbalance in attribute prediction, and demonstrate the use of facial attribute classification as an auxiliary task for improving face detection and face alignment.

In Chapter 9, Si Liu, Lisa Brown, Qiang Chen, Junshi Huang, Luoqi Liu, and Shuicheng Yan introduce methods that leverage facial and clothing attributes as a mid-level representation for applications related to fashion. In particular, they show that modeling attributes is crucial for fashion recommendation systems. In addition, they show that attributes play a major role in a system for clothing retrieval from online shopping catalogs.

Part IV: Defining a Vocabulary of Attributes

After covering multiple uses of visual attributes, as described earlier, we address the problem of discovering them, i.e., how to define a vocabulary of attributes.

In Chapter 10, Subhransu Maji surveys recent methods and defines a taxonomy of techniques for discovering a vocabulary of parts and attributes. The approaches discussed in this survey consider a vocabulary of attributes defined by experts and based on discovery methods, such as non-semantic embeddings, text mining, similarity comparisons, and others.

In Chapter 11, Genevieve Patterson and James Hays use crowdsourcing to generate a vocabulary of discriminative scene attributes related to affordances, materials, and spatial layout. After the attributes are discovered, they annotate more than ten thousand images with individual attribute labels, and show that attribute models derived from this data serve as an effective intermediate representation for zero-shot learning and image retrieval tasks.

Part V: Attributes and Language

We conclude our volume with a forward-looking topic: the connection of visual attributes and natural language.

In Chapter 12, Marcus Rohrbach discusses using visual attributes as semantic units between natural language and visual recognition. In particular, he covers methods for mining attributes from language resources, generating sentences from images and video, grounding natural language in visual content, and visual question answering.

In Chapter 13, Carina Silberer states that distributional models of word meaning have been criticised as "disembodied" in that they are not grounded in perception, and show that visual attributes predicted from images can be used as a way of physically grounding word meaning. Silberer introduces a new large-scale dataset of images annotated with visual attributes and a neural network-based model, which learns higher-level meaning representations by mapping words and images, represented by attributes, into a common embedding space.

References

[1] Akata, Z., Malinowski, M., Fritz, M., Schiele, B.: Multi-cue zero-shot learning with strong supervision. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[2] Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label embeddings for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **38**, 1425–1438 (2016)

[3] Biederman, I.: Recognition by components - a theory of human image understanding. Psychological Review **94**, 115–147 (1987)

[4] 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)

[5] Christie, G., Parkash, A., Krothapalli, U., Parikh, D.: Predicting user annoyance using visual attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
[6] 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)
[7] 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)
[8] Deza, A., Parikh, D.: Understanding image virality. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
[9] Dhar, S., Ordonez, V., Berg, T.: High level describable attributes for predicting aesthetics and interestingness. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
[10] Donahue, J., Grauman, K.: Image recognition with annotator rationales. In: International Conference on Computer Vision (ICCV) (2011)
[11] Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
[12] Farhadi, A., Endres, I., Hoiem, D., , Forsyth, D.A.: Describing objects by their attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
[13] Feris, R., Bobbit, R., Brown, L., Pankanti, S.: Attribute-based people search: Lessons learnt from a practical surveillance system. In: International Conference on Multimedia Retrieval (ICMR) (2014)
[14] Gan, C., Yang, T., Gong, B.: Learning attributes equals multi-source domain generalization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
[15] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: International Conference on Computer Vision (ICCV) (2015)
[16] Hendricks, L., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European Conference on Computer Vision (ECCV) (2016)
[17] Huang, J., Feris, R., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: International Conference on Computer Vision (ICCV) (2015)
[18] Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A.: What makes a photograph memorable? IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36**, 1469–1482 (2014)
[19] Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European Conference on Computer Vision (ECCV) (2016)

[20] Kovashka, A., Parikh, D., Grauman, K.: WhittleSearch: Interactive image search with relative attribute feedback. International Journal of Computer Vision (IJCV) **115**, 185–210 (2015)
[21] Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: International Conference on Computer Vision (ICCV) (2011)
[22] Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Conference on Neural Information Processing Systems (NIPS) (2012)
[23] Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A search engine for large collections of images with faces. In: European Conference on Computer Vision (ECCV) (2008)
[24] Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. In: ACM SIGGRAPH (2014)
[25] Lampert, C., 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)
[26] Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36**, 453–465 (2013)
[27] Layne, R., T., H., Gong, S.: Re-id: Hunting attributes in the wild. In: British Machine Vision Conference (BMVC) (2014)
[28] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**, 2278–2324 (1998)
[29] Liu, L., Xu, H., Xing, J., Liu, S., Zhou, X., Yan, S.: Wow! you are so beautiful today! In: International Conference on Multimedia (ACM MM) (2013)
[30] Palatucci, M., Hinton, G., Pomerleau, D., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Conference on Neural Information Processing Systems (NIPS) (2009)
[31] Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision (ICCV) (2011)
[32] Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
[33] Qiao, R., Liu, L., Shen, C., van den Hengel, A.: Less is more: zero-shot learning from online textual documents with noise suppression. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
[34] Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where and why? semantic relatedness for knowledge transfer. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
[35] Schein, A., Popescul, A., Ungar, L., Pennock, D.: Methods and metrics for cold-start recommendations. In: International Conference on Research and Development in Information Retrieval (ACM SIGIR) (2002)

[36] Sharmanska, V., Quadrianto, N., Lampert, C.: Learning to rank using privileged information. In: International Conference on Computer Vision (ICCV) (2013)
[37] Shi, Z., Hospedales, T., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
[38] Siddiquie, B., Feris, R., Davis, L.: Image ranking and retrieval based on multi-attribute queries. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
[39] Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: European Conference on Computer Vision (ECCV) (2016)
[40] Vaquero, D., Feris, R., Brown, L., Hampapur, A.: Attribute-based people search in surveillance environments. In: Winter Conference on Applications of Computer Vision (WACV) (2009)
[41] Yan, X., Yang, J., Sohn, K., Le, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision (ECCV) (2016)
[42] Yang, S., Luo, P., Loy, C., Tang, X.: From facial part responses to face detection: A deep learning approach. In: International Conference on Computer Vision (ICCV) (2015)
[43] Zhang, Z., Luo, P., Loy, C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **38**, 918–930 (2015)