

# Chapter 2

## Where Next in Object Recognition and how much Supervision Do We Need?

Sandra Ebert and Bernt Schiele

**Abstract** Object class recognition is an active topic in computer vision still presenting many challenges. In most approaches, this task is addressed by supervised learning algorithms that need a large quantity of labels to perform well. This leads either to small datasets (<10,000 images) that capture only a subset of the real-world class distribution (but with a controlled and verified labeling procedure), or to large datasets that are more representative but also add more label noise. Therefore, semi-supervised learning has been established as a promising direction to address object recognition. It requires only few labels while simultaneously making use of the vast amount of images available today. In this chapter, we outline the main challenges of semi-supervised object recognition, we review existing approaches, and we emphasize open issues that should be addressed next to advance this research topic.

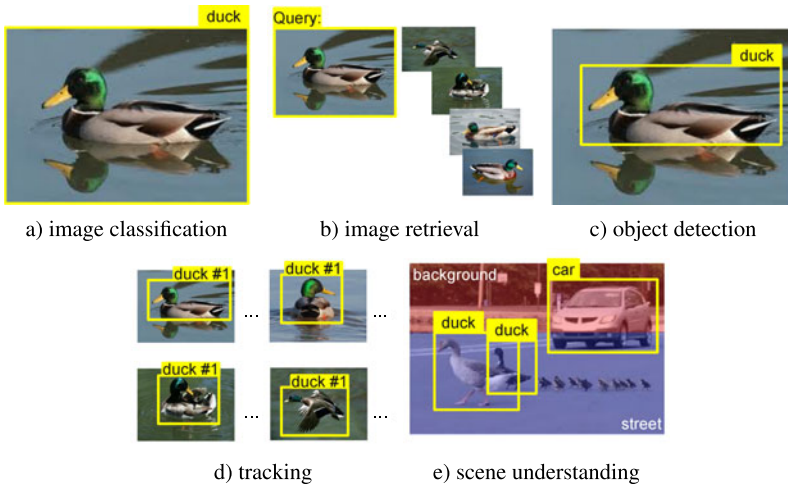
### 2.1 Introduction

Object recognition is one of the central topics in computer vision and an integral part of many computer vision tasks. To mention only a few, image classification (Fig. 2.1a) is one of the more basic tasks that includes object recognition to classify an image, for example, *duck*. Content-based image retrieval (CBIR, Fig. 2.1b) contains object recognition to search systematically for images that contain these objects. Object detection (Fig. 2.1c) must in addition specify the actual position of the recognized object in the image (marked as a bounding box), thus a clear separation between foreground and background is essential. Tracking (Fig. 2.1d) is based on object detection and tries to track the localized object across a sequence of frames. Finally, scene understanding (Fig. 2.1e) aims to capture the whole scene including all interactions among objects and the environment, for example, to

---

S. Ebert (✉) · B. Schiele  
Max Planck Institute for Informatics, Saarbrücken, Germany  
e-mail: [ebert@mpi-inf.mpg.de](mailto:ebert@mpi-inf.mpg.de)

B. Schiele  
e-mail: [schiele@mpi-inf.mpg.de](mailto:schiele@mpi-inf.mpg.de)

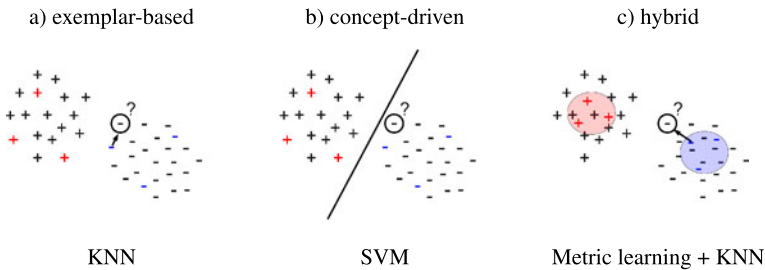


**Fig. 2.1** Computer vision applications with object recognition as an integral part

warn the car driver before an accident happens with the ducks. This list of computer vision tasks could be continued arbitrarily. But although object recognition is a crucial part it is surprising that even image classification, which only aims to provide the image label, does not provide satisfactory results on more challenging datasets.

In contrast, humans can quickly and accurately recognize objects in images or video sequences. They can categorize them into thousands of categories [10]. Beyond that they can track objects in videos and they are able to interpret the entire scene and to infer subsequent events. Of course, human perception, recognition, and inference also have their limits but they might serve as a good starting point to improve upon. Therefore it is not surprising that computer vision, in particular the machine learning part of it, is mainly driven by cognitive science—the science of understanding the learning and thinking of humans. But as controversial theories in cognition are, as diverse are the approaches in computer vision and machine learning.

One of these long-lasting debates in cognition focuses on the question whether a human learns exemplar-based or concept-driven. The exemplar-based model [67, 74, 129] assumes that humans store a list of exemplars for each category. A category decision will be made based on a similarity to existing exemplars. One of the most prominent representative in machine learning is the  $k$  nearest neighbor classifier visualized in Fig. 2.2a. This classifier looks for the nearest labeled neighbor in the training set marked as red and blue data point and uses the label of this training sample for classification. In contrast, concept-driven learning assumes that people abstract to a model or a prototype that is used for classifying objects [68, 69, 84]. This paradigm can be found in many algorithms that learn a model of a category and do any kind of generalization such as SVM that learns a decision boundary (Fig. 2.2b). More recently, there is a tendency towards the theory that humans use



**Fig. 2.2** Illustration of several learning principles that are used to classify the marked unlabeled data point: **(a)** exemplar-based learning, e.g., KNN classification, **(b)** concept-driven, e.g., SVM, or **(c)** hybrid approach, e.g., that groups exemplars around a general concept by transformation with metric learning and then applying KNN. *Blue and red points* are the labels of two different classes, and *black points* are the unlabeled data

either multiple learning systems in parallel [4, 35] or a hybrid version that groups exemplars around a general concept. This approach can be found for example in a combination of metric learning and KNN that first maps the exemplars to a more discriminative description, that is, examples of the same class are closer together visualized as red and blue area in Fig. 2.2c, and then applies KNN.

Another equally controversial but much older debate revolves around the question how we gain the insight that forms the base of our decisions. This leads to one of the most fundamental questions in machine vision: whether and how much supervision do we need? On one hand, a human learns provably faster with supervised feedback [3]. Therefore, it is not surprising that state-of-the-art performance is achieved by supervised algorithms. However, this success has to be put into perspective as the dataset construction itself contains an enormous amount of supervision. Each method is only as good as the underlying training data. If the learner sees only the side view of a car during training, then the resulting classifier will fail on cars shown in front views or from above. This aspect is often neglected in the subsequent evaluation and leads to datasets that are either small and strongly biased [82, 113] or large and error-prone [124]. On the other hand, it is also clear that many human decisions are driven to some extent by intuition, that is, more or less unsupervised, particularly in unfamiliar or risky situations [51]. But although unsupervised learning is an important research area [103, 123], for example, for object discovery or novelty detection, a minimum level of supervision is required at the end to judge the quality and to gain insight. Therefore, semi-supervised learning (SSL) has to be turned out the paradigm that combines the advantages of supervised learning and unsupervised learning by using both labels as well as the underlying structure or geometry in the data.

In the following, we discuss the large potential of semi-supervised learning in Sect. 2.2. After that, we outline in Sect. 2.3 the main challenges of object recognition with respect to semi-supervised learning and how it is already addressed in previous work. Finally, we point out open issues and we give recommendations how we could tackle those in Sect. 2.4.

## 2.2 How much Supervision Do We Need?

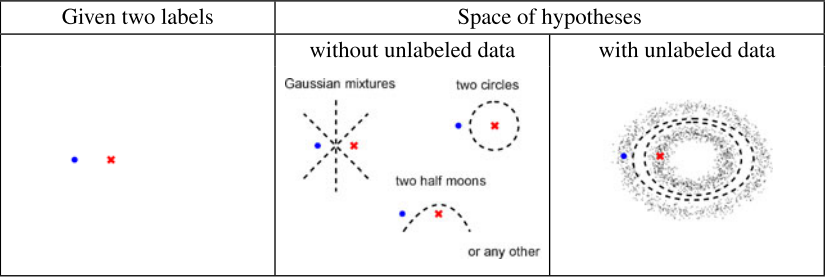
*Mind without structure is empty and  
perception without labels is blind.*

Translated from [52]

Imitation, intuition, and experience play an important role in human decision making [23] especially under risk [51]. But only a small fraction of this accumulated knowledge is labeled. The discussion around the question on *whether and how much supervision a human need* can be traced back at least to the philosophical theories of the 17th century. [52] was the first who argued that labels (knowledge) and structure (perception or observation) are closely intertwined, summarized in one of his key phrases (see beginning of this section). His theory fundamentally changed and influenced our way of thinking and acting. After 200 years of research, some of his basic assumptions and argumentations might be obsolete. But the underlying theory and the main argumentation itself is still up-to-date. Indeed this theory seems to be obvious because in addition to the things we learn supervised at home or in school, there are many other things that we learn without any teacher. For example, how we move, how we grab a glass, how we use language before we start school, how we make fast decisions and so on. Of course, many of those things are learned feedback-driven in the sense that an action is completed successfully or not. However, there are still many actions or feelings that we cannot explain let alone derive solely based on knowledge. In that sense, semi-supervised learning seems a natural choice to address object recognition as it allows us to improve state-of-the-art supervised learning approaches with more data without the need of correct annotations.

But in defense of the more skeptical people, one has to state that this large unlabeled part of semi-supervised learning is almost impossible to grasp. Actually, it is even not clear whether humans will ever be capable to understand it in their completeness. For the simple reason that in the course of evolution, we only had to understand and infer simple causalities, for example, *I take the glass of water because I am thirsty*. But we were never forced to understand the entire chain of actions and decisions that leads to this final action, for example, *grasping the glass and drinking*. In fact, every physical movement is a highly individual action that is based on imitation and experience. Although this might lead to sub-optimal movements or actions, the acquired knowledge is quite sufficient to survive in everyday life—even if we notice some limitation and ask for more supervision, for example, to get rid of pain induced by suboptimal movements or to run faster in a marathon. Most advices only give a direction and do not describe a muscle-induced action in its full complexity.

One might argue that these examples are indeed more physical. But even if we limit these considerations to our decisions that could be purely driven by our knowledge, we still observe many decision based on the so called *gut feeling* or other feelings that we cannot explain. Why do we know that someone is annoyed or sad or impatient although this person uses exactly the same words as he does every day? Why is it impossible to imagine in advance how we will react or feel after a certain event, for example if we loose a competition. Even more complicated is to infer how other people will react on an event. The reason is simple and devastating at the same



**Fig. 2.3** Unlabeled data reduce the space of hypotheses if there are only few labels

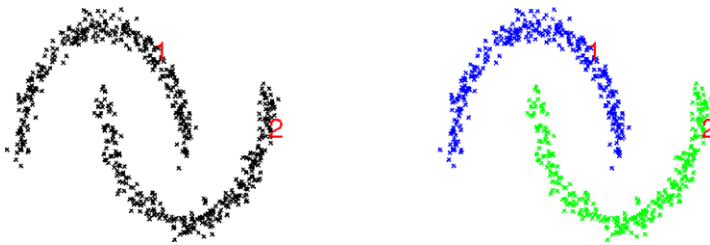
time: We are overwhelmed by millions of sensations per minute of which only a small fraction of impressions are processed consciously and the rest subconsciously and this is only a tiny fraction of what the entire world perceives. Thus, to answer at least the questions with respect to our own, we have to assimilate all sensations. With this insight, it becomes clear why research in semi-supervised learning (SSL) is still not where it should be. To bring a machine into the closer range of human thinking, we have to tap into the vast amount of unlabeled data meaning a ratio of 1 labeled to 1 million unlabeled data points and not the common ratio of today’s task descriptions of 1 labeled to 100 unlabeled observations.

Besides these more philosophical considerations, there are also many practical reasons for SSL. One obvious argument is the reduction of the hypotheses space [5] in particular if there are only few labels. Figure 2.3 shows one point per class in the leftmost image. Without any additional information, the space of possible hypotheses is less goal-oriented (Fig. 2.3 middle) while unlabeled data reduces this space as shown in the right visualization. This speed up of concept learning through relevant prior knowledge has been also verified in cognition by [70, 78]. A mechanical engineer will be proceed faster and more goal-oriented when assembling a machine in comparison to a layman because he already knows how to use the tools and where the single items should be approximately placed.

Another reason for SSL is the low amount of supervision. In particular for tasks such as semantic image labeling or image understanding, where we need pixel-wise annotations, this advantage becomes increasingly important. But also for tasks such as object detection or recognition, we observe a substantial improvement the more data are used. The most image descriptions are high dimensional resulting in a strong demand for data. But the labeling process is not always reliable—for example, when using Mechanical Turk [124]. Finally, some applications need a continuous update, for example, for separating spam emails from valid emails.

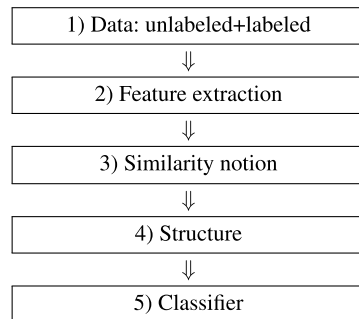
### 2.3 Challenges of Object Recognition

Semi-supervised learning (SSL) seems a reasonable approach to tackle object recognition as it makes use of both supervision in terms of labels and structure (geom-



**Fig. 2.4** Two half moons dataset with exactly one label per class (marked by a *red number*): before classification (*left*) and after classification (*right*) with 100 % accuracy

**Fig. 2.5** Workflow of semi-supervised learning algorithms in vision

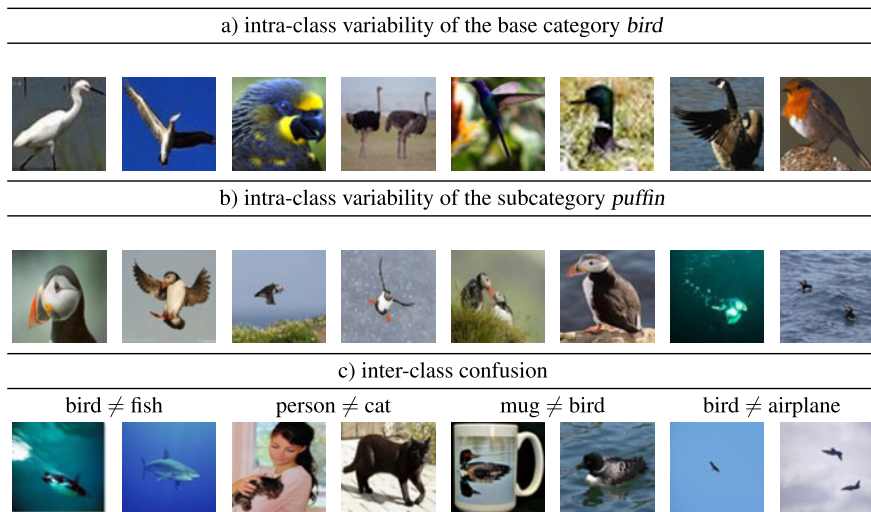


etry) that comes with the unlabeled and labeled data. Therefore, it is obvious that both parts strongly influence the performance of the classifier. In the following, we discuss the challenges of both components separately starting with the structural problems in Sect. 2.3.1 followed by the difficulties in getting representative labels in Sect. 2.3.2.

### 2.3.1 Structural Problems

The apparently most promising but also much more complicated direction for SSL is the improvement of the structure itself. Imagine a dataset like the *two half moons* shown in Fig. 2.4 with two labels marked with red numbers. It does not matter which label is used for classification. The used SSL algorithm [132] achieves always a classification performance of 100 %. Although this is an artificial example it still reflects our common sense assumption that there is exactly one concept for each class and each instance of this class is organized around this central prototype [19, 75]. But often, there is a large gap between our base assumption and today's computer vision task descriptions and solutions.

Figure 2.5 visualizes the general workflow of object recognition with SSL algorithms: Based on our dataset that consists of labeled and unlabeled data, we compute image descriptors for each image. After that we compute the similarities between each image pair with some measure. The resulting structure is used by SSL algo-



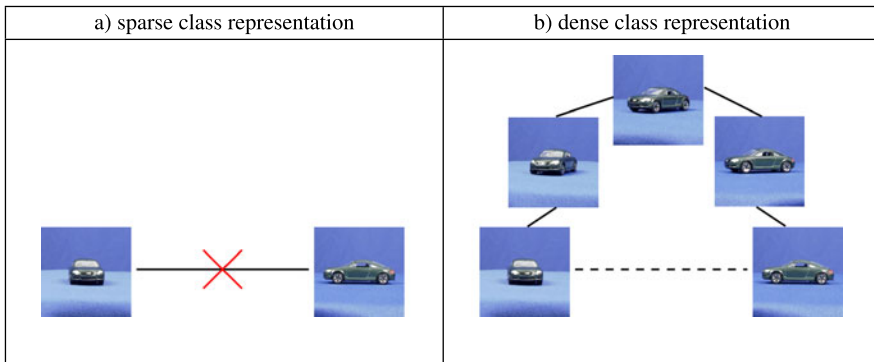
**Fig. 2.6** Examples for (a) large intra-class variability for the base category *bird* (top row) and (b) for the subcategory *puffin* of the category *bird*, and (c) small inter-class differences (bottom row)

rithms, for example, for EM clustering [73], as a regularization term to improve SVM [50, 98], or to build a graph structure and to find a solution with Mincut [12] or by label spreading [132]. However, each of these classifiers can be only as good as the extracted geometry of the data and this strongly depends on three main sources: (1) data, (2) image description, and (3) similarity notion. Furthermore, the quality of each single source is dependent on both the approaches that are used for these steps but also on the quality of the previous steps. This means, information loss for example through an incomplete dataset will be propagated to the classifier and cannot be compensated by one of the intermediate steps. Similar argument holds for image description: if one aspect is neglected, for example, color, the best similarity measure will be not able to properly distinguish between *green apples* and *red tomatoes*. In the following, we discuss each of these three components separately.

### 2.3.1.1 Data

Most common datasets for image classification like Caltech 101 [38], PASCAL VOC [36], Animals with Attribute [59], LabelMe [114], animals on the web [9], or ImageNet [27], are generated for supervised classification. They provide full label information that might be error-prone in particular when crowd-source services like Mechanical Turk are used [124]. They contain a large intra-class variety within a base category such as *bird* (Fig. 2.6a) but also within one specific subcategory of this class such as *puffin* (Fig. 2.6b). Without a good description and understanding of the concept, it will be difficult to connect those examples and group them together to one class. Small inter-class variation is the other end of the scale (Fig. 2.6c) leading





**Fig. 2.7** Example of (a) a sparse class description that makes it difficult to find a connection between both images and of (b) a dense class representation that makes it possible to find a way from the front view to the side view of a car over several viewpoints

to many overlapping and confusing areas that are even for humans difficult to learn. Finally and most unfortunately, they contain usually a limited amount of images because labeling is expensive.

An ideal dataset for SSL should be *dense* enough that means each class should be densely sampled that allows to find compact and well separated clusters. In this dataset, we might be able to connect the front view of a car and a side view of car as in Fig. 2.7. But on the other hand, this dataset should be also *sparse* enough to avoid overheads due to space and time complexity. Usually, SSL approaches such as graph-based algorithms come with a quadratic time complexity in the number of images. Because of this complexity, the *the-more-data-the-better* strategy can usually not be applied.

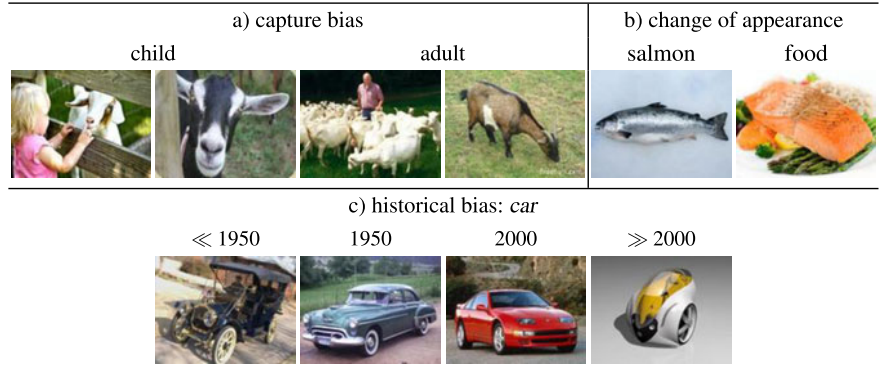
A second reason why *the-more-data-the-better* strategy does not work well in practice is that only a small fraction of the data, for example, added from the internet, is helpful for our classification task. Furthermore, these data sources often have a certain bias in terms of image type. One example is the data source bias. Flickr, that is the base for PASCAL VOC [36], contains mostly holiday pictures. Therefore *person* is with 31.2 % the most common object in this dataset. The second most frequent object is *chair* with 8.5 %. Another bias can be also seen in Fig. 2.8 (top row) that shows the first results of approx. 5.8 million results for the query *car* in Flickr. In contrast, Google shows more professional images that are often generated for marketing purpose as you can see in Fig. 2.8 (bottom row).

Another problem comes with the capture bias. This is usually a result from human properties such as body height. Most images are taken from an adult person in a standing position thus from an average height of 1.6–1.8 meters (Fig. 2.9a). A simple change to a child position, that is, <1 meter, leads to a different viewpoint and thus perception, for example, some things appear larger (buildings) or more frightening (animals). Furthermore, most people are right handed resulting, for example, in many images of mugs with the handle on the right side. Some objects have a quite different appearance (Fig. 2.9b), e.g., salmon considered as an animal vs. food, and





**Fig. 2.8** Data source bias: First results for the query *car* with (a) Flickr that contains more holiday pictures of cars (*top row*), and (b) Google images with focus on racing cars (*bottom row*)



**Fig. 2.9** Dataset bias due to (a) the sighting angle, (b) the semantic of an object, or (c) because of historical trends

other categories might change their appearance over time (Fig. 2.9c). Of course, it seems likely that massive amounts of data also contain relevant images but to find these images we have to process over millions of images for this single class.

**Related Work** Many SSL algorithms in particular graph-based algorithms come at least with a runtime of  $O(n^2m)$  with  $n$  the number of data and  $m$  the number of feature dimensions. This runtime is needed to compute all similarities between image pairs and to construct the graph. Thus, the applied algorithm depends strongly on the number of data and the dimensionality of the features but also on the approach itself. For example, [132] provide both a closed form solution that would need the inversion of a  $n \times n$  matrix and an iterative procedure that is faster and often avoids over-fitting. In general, there are two different strategies to reduce the runtime: (i) a reduction of the data space ( $\ll n$ ) to a representative subset of unlabeled data or (ii) an approximation either of the similarity matrix or the eigenvectors.

- (i) *Data reduction.* The most common approach to data reduction is clustering to find representative unlabeled data either by hierarchical clustering [64], or by k-means clustering [65, 100]. [26] propose a Greedy approach that starts with the labels only and successively add unlabeled samples farthest away from the

current set of labeled and unlabeled data. [37] find similar nodes by spectral decomposition and merge these together. Another technique is to treat this task as an optimization problem that considers each point in a data set as a convex combination of a set of archetypical or prototypical examples either with a fixed number of archetypes [21] or with an automatically learned number of these prototypes [85]. These techniques are used, e.g., to find typical poses [8], or to summarize a video sequence [34].

- (ii) *Approximation.* In contrast, Nystroem approximation is employed to approximate the entire kernel matrix. This approximation is estimated also on a subset of data that are retrieved either by random sampling [131] or with k-means clustering [130]. This approximation can then be used to find a segmentation [42], for similarity search [122], or face recognition [109]. To speed up the algorithms, [40] propose an approximation of the eigenvectors of the normalized graph Laplacian. [115] solve the dual optimization problem by introducing a sparsity constraint, and [54] use stochastic gradient descent to solve the TSVM.

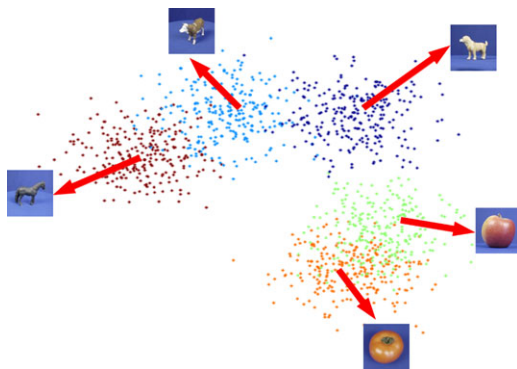
In [33], we focus mainly on the question if more unlabeled data have really a positive impact on the classification performance. In particular, we challenge the *the-more-data-the-better* strategy that is common sense in the computer vision community but also comes with an increase in runtime and space. This question is difficult to answer as adding unlabeled data leads to a different field of research due to the dataset bias [82, 113] and the data source bias (Sect. 2.3.1). Therefore, we focus on ILSVRC 2010 with 1 million images and reduce this large amount of data to a representative subset of unlabeled data showing that this representative subset leads to a better graph structure than using all unlabeled data. We compare our approach to [26] and [65] that can be considered state-of-the-art methods to reduce the graph size. But in comparison to previous work, we analyze also the effect of more unlabeled data. Additionally, our work is the first attempt to process more than one million data points that is far more than 30,000 data points used in previous work [26, 65, 130]. But this can be only seen as good starting point to improve on.

### 2.3.1.2 Image Description

Suppose we have a dataset that captures the broad variety of each class, for example, different viewpoints, several contexts and so on. Thus, there is a chance to build a compact cluster structure similar to Fig. 2.10. Then it does not automatically mean that we are also able to exhaust this potential. Today's image descriptors are far away from capturing all these different aspects that humans can easily recognize. In the following, we list briefly most of the common problems. See also [43] for a short overview of today's problem in computer vision.

**Intra-class/Inter-class Variability** As it mentioned before, many classes come with a large intra-class variance in their appearance and their surrounding environment. The class *bird* is one of the extreme cases where even the height varies from few centimeters like the hummingbird to almost 2 meters like the flightless ostrich

**Fig. 2.10** Structure with dense representation but still overlapping regions



(Fig. 2.6a) not to mention the large variation in shape and in color. Of course, a limitation to one species (Fig. 2.6b) might constrain the general appearance of an object but not the number of different poses or the context around this object. On the other side, there are classes that look similar to each other in particular in some poses, for example, a bird and an airplane in the sky (Fig. 2.6c), or when two classes jointly appear in an image, for example, a cat in the arms of a person or a sticker from an animal at a mug.

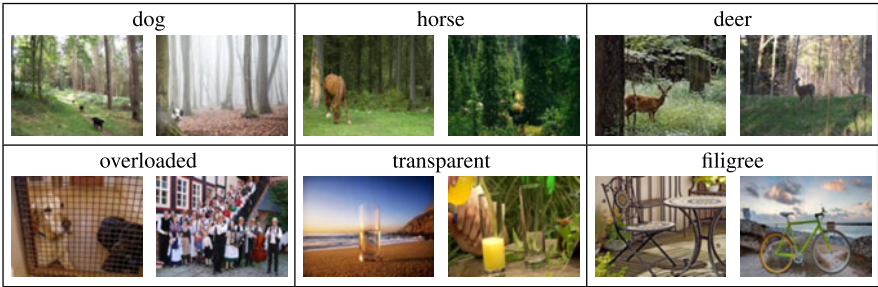
**Background Clutter** Some images are dominated by their background as can be seen in Fig. 2.11 where it is almost impossible to see some objects because of the trees. Often, these images are confused because of their similar-looking background. There are images with an overloaded background structure that are similar to many other images in a dataset. Finally, some objects are difficult to distinguish from the background because they are transparent (like glass) [44], or filigree like a bicycle or a chair.

**Illumination Changes** Another problem is the change in illumination depending on the time of the day and the season (Fig. 2.12). For example, a lake is susceptible to lighting conditions due to its surface and volume properties resulting in a wide color spectrum. But also many other objects look different during the day and at night, for example, trees are green during the day and dark at night.

**Truncation and Partial Occlusion** Partial occlusions are an omnipresent property. Herd animals like sheep or gazelles occur frequently in groups. As already mentioned before, some objects can be covered to a large extent by a person using that object like a bicycle or chair. And other object classes are very large so that they are only partly captured or truncated like a cathedral (Fig. 2.13).

**Shape Variation** Some categories have a large variation in shape and appearance, for example, *chair* (Fig. 2.14), *table*, or *lamp*. These categories can be often only described by their function such as *something to sit on*.

**Mimesis and Other** Another set of problems comes with the evolutionary adaptation of some species to their background so that they are difficult to recognize by



**Fig. 2.11** Examples with a dominating background that is shared among different classes (*top row*), and examples with overloaded background and object that are transparent or filigree so that background is always a part of the object (*bottom row*)



**Fig. 2.12** Examples of a lake with different illuminations depending on the time of day and the season



**Fig. 2.13** Examples of truncations and partial occlusions



**Fig. 2.14** Examples of the large variety in shape for the class *chair*

other animals, for example, the chameleon or the flounder. Other animals such as zebras are indeed visible but it is difficult to point out an individual animal due to their pattern structure (Fig. 2.15).

Basically, the ideal image descriptor should emerge with some prior knowledge about what and where the object is located in the image, how to separate the background from the main object, which color is trustable or rather how to adjust this color, and what are the possible and feasible poses of an object. Furthermore, this descriptor should have a general idea of the shape and texture of an object to infer which part is occluded or truncated. While a human focuses led on the main object,



**Fig. 2.15** Examples of objects that are difficult to distinguish from their background or to identify the object-specific shape



**Fig. 2.16** Examples of images that are difficult to understand without color information

many of today's image descriptors such as dense SIFT analyze every single blade of grass or every single leaf from a tree leading to an overcrowded image description that often considers only one aspect in the image like color or gradients. Of course a good similarity metric can handle this high dimensionality. But an information loss in this partial extraction propagates to the classifier. Figure 2.16 shows some examples with and without color information. Even for a human it is hard to follow a soccer game or to distinguish between eatable and poisonous mushrooms by just omitting color, not to mention information such as texture or shape.

**Related Work** As mentioned before, most image descriptors lack still expressiveness. They consider only one aspect when describing an image, for example, texture, shape or color. This leads inherently to an enormous information loss. Therefore, a combination of several features are essential, for example image-based features with geometry [14, 83, 125], shape with texture [20], several local appearances [94], local and global appearances [61], HOG with texture [121], multiple kernels for global as well as local features learned with an SVM [46, 105, 117], with boosting [28], or with conditional random fields [96].

In SSL, the combination of multiple graphs offers the possibility to capture different aspects in the data. For graph-based methods, there are few works that combine graph structures similar to multiple kernel learning (MKL) [2, 116]. In [55], they learn weights for combining graph Laplacian within an EM framework. [22] propose a method to find one graph from a set of graphs that best fits the data. [112] formulate this combination as a multi-modality learning problem that fuses different modalities either linearly or sequentially. [6] use domain knowledge to ex-



tract three different sources, that is, time, color and face features, that are combined with different hyperparameters. Finally, [47] and [110] combine a similarity and a dissimilarity graph and apply label propagation on this mixed graph structure. Most of these previous works are developed for applications in bioinformatics. But more importantly, they combine often only graphs based on different parameters, that is, a different number of neighbors  $k$  or different weight functions.

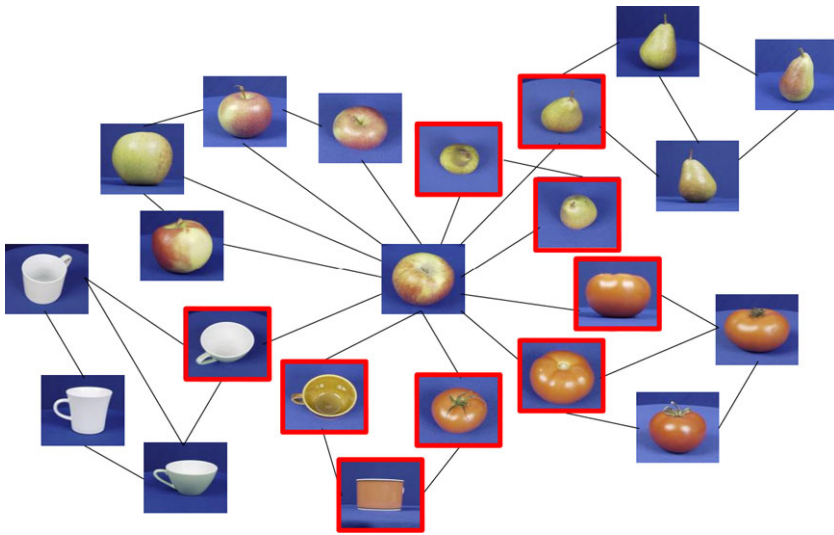
In [30], we show the strong influence on the graph quality when combining different image descriptors leading to a completely different and more powerful graph structure. Additionally, we use the SVM output to construct a new graph. But in comparison to [86] who use the SVM output to delete and insert existing edges, we build a complete new graph based on these decision values and combine this graph with our original graph. This leads to a richer and better connected graph structure than the graph structure in [86].

### 2.3.1.3 Similarity Notion

The final crucial part of the structure extraction is the similarity measure. Most frequently used is the Euclidean distance with a Gaussian kernel weighting. This is usually a good choice for feature vectors of low dimensionality ( $\ll 100$ ) containing only little noise. But as mentioned before, most image descriptors aggregate many not preprocessed information that leads to a high-dimensional vector ( $> 10,000$ ) from which only a small fraction of dimensions are relevant for a object class. In particular Euclidean distance is known to be sensitive to noise that becomes more prominent the more dimensions are used. One phenomenon that we observe with Euclidean distance is that some images are similar to almost all other images. The resulting structure (such as shown in Fig. 2.17) harms almost every classification algorithm because there is no clear separation between different classes [119].

Another problem comes with the missing weighting of the single dimensions in the feature space, that is, all dimensions are equally considered. But usually only a small fraction of this high dimensional feature vector is relevant for each class. Finally, we often consider image pairs instead of groups of images. This is easy to implement but seems suboptimal for good generalization. A human who has never seen a zebra before and only gets the first image from Fig. 2.18 will certainly have problems to build a general concept or model of a zebra because there is no information about the shape, the size, or the environment around this animal. Without these higher order relations extracted from a group of images, it might be difficult to distinguish the first image from the sofa shown in the last image.

**Related Work** Metric learning is a promising direction to tackle this problem. These methods find a better data representation such that examples within a class are close together and examples from different classes are far away, that is, small intra-class distances and large inter-class distances. Metric learning approaches can be split into unsupervised, supervised, and semi-supervised methods that are further divided into global and local learning methods. See also [126] and [29] who provide a more detailed exploration of metric learning methods.



**Fig. 2.17** Problem of Euclidean distance in a high dimensional space: The image in the *middle* is similar to many other images. *Red boxes* indicate false neighbors



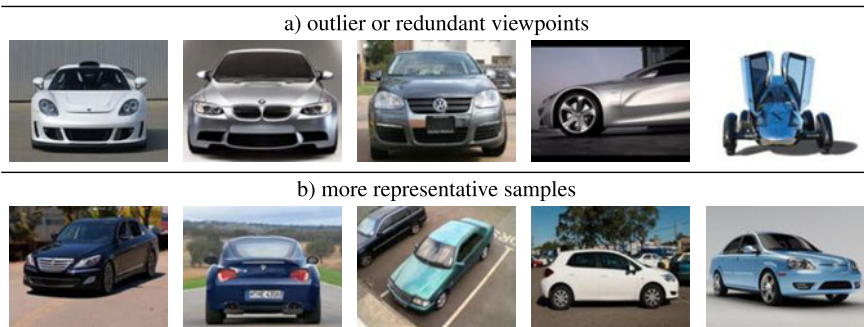
**Fig. 2.18** Pairwise image similarities might be problematic due to the missing generalization. From the *first image* it is not clear how to generalize so that this image do not get confused with the sofa in the *last image*

In [31] and [29], we analyze several supervised and unsupervised metric learning approaches with respect to a better graph construction. Particularly, we apply PCA [79] and LDA [41] to reduce the dimensionality of our feature representation and compare both methods. Furthermore, we also analyze ITML [24]. Instead of reducing the number of dimensions, it learns a weighting of the feature dimensions. The advantage of ITML is that it can be transformed into a kernelized optimization problem. Thus, the runtime depends only on the number of labels that is usually smaller than the number of dimensions ( $n \ll d$ ). Additionally, this approach shows state-of-the-art performance on Caltech 101 [58]. Finally, we integrate ITML in a semi-supervised metric learning scheme that leads to an increased performance.

### 2.3.2 Labeling Issues

The second import issue besides the structure is the label information. As mentioned before, supervision causes no problems if the structure itself perfectly separates





**Fig. 2.19** SSL is strongly dependent on the representativeness of the small training set: **(a)** less representative samples for the entire class *car* vs. **(b)** more representative samples in terms of viewpoints

the classes. But usually this is not the case. Therefore, the label information plays an important role in particular for semi-supervised learning where we have only few labels per class. While for supervised learning more data is labeled, in SSL we have to deal with a ratio of 1 %–5 % labeled to 95 %–99 % unlabeled data. Thus, there is a need for high quality labels that are representative for the class and allow a better generalization. Additionally, we have to ensure that there is at least one label for each mixture (e.g., different viewpoints or appearances) of one class otherwise it might be difficult to classify unseen viewpoints. Figure 2.19 shows five less representative samples for the class *car* in the first row, assuming that the test set contains also the back or the side view of a car. In contrast, the second row shows more representative samples of this class so that the main properties of this object will be apparent such as the shape and the surface.

Coming back to Fig. 2.17 if the image in the middle with these many false neighbors is labeled, then most of the neighboring images will be false classified (marked with a red bounding box) because of the strong impact on the direct neighbors. Another problem occurs when a class is split into separate clusters, for example, front view of a car and side view of a car, and there are only labels for one of these sub-clusters. The other sub-clusters cannot be classified correctly anymore. Ideally, we have labels that are representative or prototypical for a class that means they lie in a dense region and consider each aspect or viewpoint of a class.

**Related Work** Active learning is a well known strategy to reduce the amount of supervision to a small but representative subset and to improve the quality of the learner at the same time. This is also verified by cognitive science as [3] shows that a higher accuracy is achieved with feedback during the learning in comparison to the scenario where supervision is only provided at the beginning. In machine learning, active learning leads in most cases to better performance. [1] show that some NP-complete learning problems become polynomial in computation time. On the other side, active learning in combination with some classification algorithms might lead to poor performance, for example, SVM with few examples at the beginning [120]. Model selection is critical for these algorithms [108].

Most popular is pool-based active learning. These methods consider all unlabeled data as a pool from which samples are drawn to be labeled. In general, pool-based methods can be divided by their sampling strategy into three different types. Exploitation-driven methods [97, 111] focus mainly on uncertain regions during the learning process. In contrast, methods based on exploration sampling [13, 72] estimate the overall distribution of the entire data space and query samples that represent and cover this space. Finally, there are also strategies that combine both exploration and exploitation to get samples that are uncertain but also diverse. We refer also to [99] and [29] who provide a general overview on different active learning strategies.

In [32], we propose a novel sampling criteria for exploration that shows significant better performance in comparison to previous exploration criteria. Furthermore, we address active learning by proposing a reinforced active learning formulation (RALF) that considers the entire active learning sequence as a process. Our approach can deal with multiple criteria, is able to have time-varying trade-offs between exploration and exploitation, and is fast and efficient due to a compact parameterization of the model without dataset-specific tuning. In comparison to [7] who also use a reinforcement procedure, our model comes with fewer parameters, more flexibility in terms of sampling criteria, and provides always a linear combination of exploration and exploitation instead of switching between criteria. For the linear combination, we extend the work proposed by [15] to a time-varying combination that leads to a better adaptivity to dataset requirements. Finally, we show in [33] that a potentially large improvement is possible when applying metric learning with more representative labels. To this end, we combine active learning with metric learning and show improvements of more than 10 % over our previous publication [31] and more than 20 % improvement as compared to our first publication [30].

## 2.4 Future Perspective

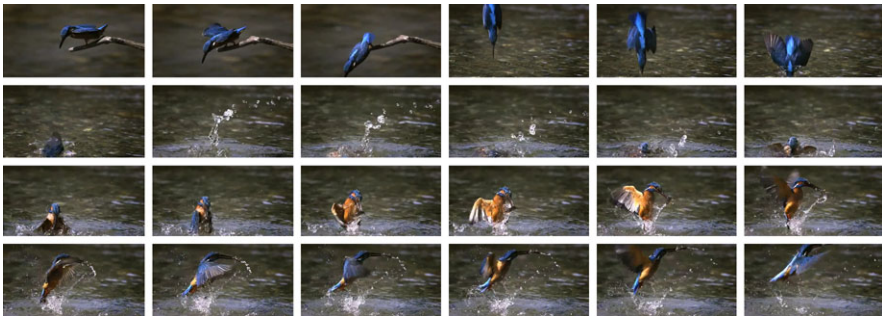
In this last section, we will mention open issues and how we could tackle those in Sect. 2.4.1. These suggestions are closely intertwined to cognitive science for the simple reason that particularly object recognition raises automatically the question how humans form class concepts. Thus, it is not surprising that this close relationship has been studied earlier, for example, in the *Roadmap of Cognitive Vision* [118]. Of course, the human object recognition should be only seen as a good starting point to improve on. Therefore, we finally discuss in Sect. 2.4.2 also the human weaknesses in perception, learning and inference and how we might overcome those with machine learning and computer vision.

### 2.4.1 What Can We Learn from Human Object Recognition?

Following the general structure of this chapter, we discuss open issues of each of the components of SSL separately, that is, (1) data, (2) image description, (3) sim-



**Fig. 2.20** Visualization of an incomplete object class description that makes it difficult to find relation between these images

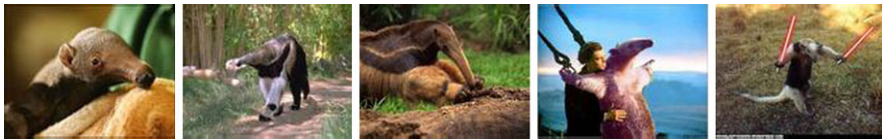


**Fig. 2.21** Visualization of a more complete object class description extracted from a video sequence

ilarity notion and (4) supervision. Additionally, we also challenge the use of label propagation in the last subsection called (5) exemplar-based vs. concept-driven learning.

#### 2.4.1.1 Data

[133] stated in their position paper about graph-based SSL that one of the limitations of these algorithms results from the common sense assumption that each class can be projected to a single manifold. Thus, they suggest to model classes by a mixture of multiple manifolds. This observation is particularly in computer vision not novel. Also [92] distinguish between visual classes and object classes as there are classes such as *chair* that cannot be modeled with one mixture. Even if this is an important aspect, an at least equally important problem is the incompleteness of today's datasets in terms of viewpoints and variations for a class to fully leverage the power of SSL algorithms. In [29], we assume that most classes can be described with one concept and all exemplars of these classes are grouped around this concept [19]. But often we have to deal with class descriptions as shown in Fig. 2.20 for the class *kingfisher*. Even for a human who have never seen this class before might have problems to merge these images in one compact mixture because color, shape, and appearance are quite different. But an image sequence might provide a path among those images as visualized in Fig. 2.21 and helps to extract an object models similar to the work of [90] shown for car tracking.



**Fig. 2.22** Visualization of feasible and unfeasible poses and scenes for the class *anteater* that will be more obvious with informations such as 1.5–1.8 m long

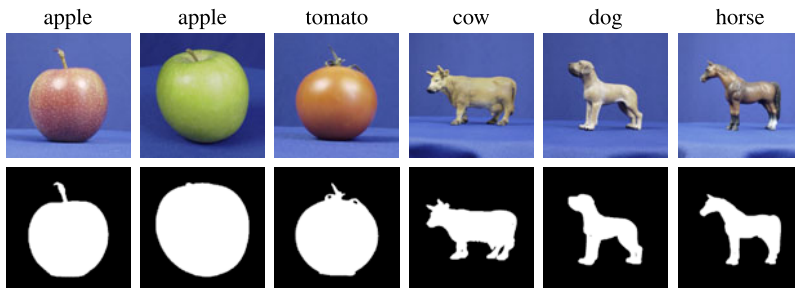
In [33], we analyze the problem of incomplete datasets by looking at existing datasets such as ImageNet with more than one million images and their behavior and performance when adding more unlabeled data. But this should be only considered as a first attempt towards a smooth manifold structure. As a next step, we have to get away from this controlled setting because we still depend on the quality of the datasets given by the limited (although larger) amount of images and the quality of labels. Instead, we have to tap into other sources. In general, there are three possibilities: (1) combining several datasets, (2) adding synthetic data or (3) browsing the internet.

Merging different datasets is problematic because most of the available datasets have an inherent bias attached to the dataset [82, 113]. Although there are works that try to undo the damage of dataset by estimating the bias for each dataset [56], combining itself seems an unsatisfactory strategy because each dataset has different classes and the amount of images is also strongly limited. In contrast, adding synthetic data is a more promising direction but is still in a early stage of development. There are only few works that either generate new training images [63, 81], or add synthetic data points (so-called *ghost points*) in the distance space itself [18, 127]. The former approach is currently bound to certain classes such as *people* for where 3D shape models exist that ensure the generation of feasible poses and shape variations. The latter one is hard to control because the semantic meaning of these *ghost points* is not clear and it can lead to a blending of different classes. To expand approaches suggested by [81] to other classes, we have to integrate more physical constraints to guarantee feasible poses and appearances of the object. Figure 2.22 shows some images of the class *anteater*. A human does not necessarily have to know and to observe this animal to decide if a pose is likely or not. Information such as size of 1.5–1.8 m or weight of  $\approx 60$  kg might be already a good advice.

Another limitation comes with the fixed pool of unlabeled data in particular if we use existing datasets. This is similar to use the knowledge of a child for our entire life without any update. But in fact, our knowledge base will be permanently updated. That is why children regard a lost rabbit in a magic show as the reality while an adult knows that this can be only a trick. However, the internet provides us a large amount of images as well as videos and it is steadily updated with new data. A transformation of current SSL algorithms into on-line learning algorithms might be necessary [48, 89, 107] to benefit from these changes. Nevertheless, tapping into this data source poses many problems and questions: How do we get representative samples for each class out of these large amounts of data? How should we deal with



**Fig. 2.23** First examples for the query *fish* in Google images (out of 1.5 billion results) that are not representative for this class



**Fig. 2.24** Most confusing classes for ETH-80 although the conditions are optimal for SSL, i.e., smooth manifold structure and no background clutter

incorrect tags? Do we get enough images for each class, for example, endangered animals/plants or deep sea fish? But even the first question is of great importance if we look at the first examples of the query *fish* in Google (Fig. 2.23) that contains drawings, a robot fish (3rd image), body paintings and other atypical examples.

#### 2.4.1.2 Image Description

Missing data is clearly not the only bottleneck that can be seen in [29] for ETH-80. This dataset is well suited for semi-supervised learning because each object is photographed from different viewpoints and there is no background clutter, occlusion, or truncation of the object. But in our experiments, we achieve at most 80 % with 5 randomly drawn labels per class and a combination of three different descriptors. Figure 2.24 shows the most confusing classes for this dataset with the corresponding binary masks, that is, tomatoes are mixed up with apples and the animals (cow, dog, horse) are confused with each other. By using also a color descriptor, we are able to distinguish green apples and red tomatoes but the final improvement is only minor.

This poses many questions. Which information do we miss? Do we need a better texture description to distinguish the different surfaces of tomato and apple? Why can a human easily guess the object class for the animals by only looking at the binary masks in the second row in Fig. 2.24? Do we describe a concept of a class in terms of proportions? It seems obvious that a better shape description is needed. But it is still not clear how to extract, to store and to use this structural description. In [95], they show that only 15 % to 30 % of an object is required to recognize 100 classes correctly independent of the orientation and the view point of the object. But many of today's shape descriptors lack on this generalization. They extract



**Fig. 2.25** Representativeness of viewpoints for the class *armadillo*: images on the *left side* show less representative viewpoints as they miss important properties of this species while the viewpoints in the *right images* are most informative for this class

often too detailed and too specific contours of an object and store this description as a template that will be later used for classification by matching. Although this is a good starting point, this approach needs too many templates to perform reasonably well. One promising direction is to use 3D information [80, 106] that allows to extract structural information about the object and to make assumptions about unseen parts in the image. In these mentioned works, they use a CAD model of an object to get this information. In general, this information is difficult to get for the most object classes that we tackle in this work. Furthermore, they consider the detection of the object also as a matching problem. But ideally we would use these rich 3D models to extract structural invariances for a set of viewpoints to get a more general description assuming that we need only a representative set of salient points to maximize the discrimination between objects [91].

Apparently, there is a similar discussion in cognitive science [49]. Supporter of the viewpoint-based model theory believe that we can extract all information from different viewpoints assuming infinite many viewpoints, i.e., templates, that would clearly exceed our brain capacity. In contrast, advocates of the structural description models argue that each object can be explained by viewpoint-independent invariances. For some classes, this assumption might be true such as *bottle* or *orange*. But for many other classes it will be difficult to find such invariances over all viewpoints, for example, the table legs are invisible in the top view. More recently, there is a common agreement that we need both templates for completeness and structural description for generalization. But this trade-off is currently missing in computer vision. Thus, we need a set of representative and most discriminative viewpoints [93] as shown for the class *armadillo* in Fig. 2.25 but we also need a more general description, for example, properties, and proportions that are invariant over a set of different viewpoints.

Another issue in our experimental setting is that we compute the image descriptor on the entire image. In fact, this is a fast and simple way of extraction but it is not clear whether this is an advantage due to the additional context information or a disadvantage because of the background clutter. But this could be analyzed by using for example part-based models [39], or foreground extraction [60].

Finally, we also miss prior knowledge and context to speed up and enhance image description similar to the work of [45] for image segmentation. [70] show that humans learn three times faster and more accurate if the features of an object were related to each other. A human learns even more although this additional knowledge is not strictly necessary for accurate performance [53]. Thus, there is obviously



a strong correlation between associations among features and final performance, for example, animals with feathers are more likely to have also wings in comparison to animals with fur. Equally important is the grouping of features or objects otherwise it would be impossible to follow a soccer match if the player do not wear an uniform. But in many applications, the raw image description is fed directly into the classifier without any intermediate steps such as grouping, ranking, or finding associations. This is rather disappointing because we cannot really reconstruct and understand what went wrong during the classification. Apart from that, some categories are almost only defined by their function, for example, *chair*. Thus, to boost the recognition of those classes, we need associations for example with human poses as shown in [25].

### 2.4.1.3 Similarity Notion

Encouraged from the positive results of previous metric learning literature, we integrate in [29] several of those methods in the graph construction procedure. But the outcome did not meet our expectations. The main reason is that previous work almost exclusively compare their methods to the Euclidean distance (L2). In this work, we also observe a larger improvements for the L2 distance but these final numbers are lower than just applying Manhattan (L1) distance. In fact, it is almost impossible to improve L1 distance with any metric learning procedure. PCA decrease the performance of L1 and also the most supervised metric learning approaches decrease the performance or do not have any effect. Only with ITML [24], we observe a small improvement of approximately 1.5 %. But this benefit seems rather out of proportion if we consider the runtime and the tedious parameter search.

One problem is that the supervised approaches tend to over-fit due to the small amount of labeled data. [66] addresses this problem by including the geometry of the entire dataset as an additional regularization parameter. But this geometry is not updated during the learning that strongly limits the outcome of this algorithm. In principle, any change of the metric space should also cause a change in the geometry of the data. In [31], we tackled this issue by using an interleaved procedure that integrates successively unlabeled data with their highest prediction values. This method works fine for datasets with an already high graph quality. Otherwise the predictions are often incorrect so that the algorithm drifts to a worse solution. Additionally, we cannot control the label distribution leading to an unbalanced metric learning as some classes are more often requested than other classes. To further improve this approach, we have to incorporate a balancing factor and we should find a way to adjust and update the predictions.

In the long term, we require also different models and levels of granularity to express the similarity between objects. The properties and the description is completely different between base categories such as *cat* and *dog* and two species of the base category *dog*. Also [88] argue that basic level categories carry most information of a category and the categorization of objects into sub- or super-categories takes usually longer than the assignment of a base level category because super-classes ask for a generalization and sub-classes need a specification. Therefore, it is



not surprising that many learning algorithms do not improve their performance when using also a hierarchy for learning as shown in [87] because they assume always the same level of similarity description. A better approach would be to start with base level categories (mid-level of a hierarchy) and to switch the strategies when learning super- and sub-classes. The general benefit of a hierarchy should be more obvious as it allows to structure our data. Another important issue might be to integrate also relations into the similarity notion such as *larger head, more compact body, thinner legs* similar to the work of [77] that use relative attributes.

Finally, we also need a better visualization of the resulting graph structure. [11] visualized in their work a neural network to answer the questions what has the network learned and how is the knowledge represented inside this network. For graph structures, similar questions cannot be answered or only insufficiently. In [29], we look usually at the next nearest neighbors. But this is only one aspect of structure. It does not reflect the interactions in the entire graph. The shortest path between two nodes might be an interesting information. But usually this does not offer any valuable clue to the graph structure as the average shortest path length is  $\approx 2$  due to the previously mentioned *hub* nodes [119]. Also information visualization strategies such as multidimensional scaling do not produce revealing results.

#### 2.4.1.4 Supervision

In [32], we improve the quality of labels with active learning. This is a promising direction and should be always considered within semi-supervised learning due to the small amount of labels and the stronger dependency of the quality of those. However, our model, that automatically estimates the trade-off between exploration and exploitation and combines more than two criteria, has still some open issues. The trade-off is modeled with discrete states and not continuously. The feedback given by the overall entropy might be unreliable. The number of parameters is in comparison to previous work [7, 76] smaller but still to high. Finally, the initialization for this reinforcement learning is difficult and time-consuming as we start with no prior knowledge. Thus, one improvement could be the integration of domain knowledge or by using counterexamples [16].

#### 2.4.1.5 Concept-Driven vs. Exemplar-Based Learning

Graph-based algorithms are a popular choice for SSL. These algorithms reflect more or less the exemplar-based theory in cognition [57, 67, 74] assuming that humans store a list of exemplars and use a nearest neighbor approach to categorize objects. But this theory seems inconsistent as [84] and [128] show that people abstract to prototypes sometimes even without seeing those [68]. Thus, we possess a generalization ability from which the used algorithms are far away. In [31, 33], we approach this problem by combining label propagation with some prototype-based methods. Metric learning transforms the data space such that classes are more compact and in [33] we add prototypical unlabeled examples.

Although these approaches are a step in the right direction, they still miss a notion of the concept that is flexible enough to classify also unseen constellations and appearances of one object. Concepts allows us to go beyond the information given or visible [104], for example, if a human knows that an object is an apple then he also knows that there is most likely a core inside. This leads to one of the fundamental questions: “What makes a category seem coherent?” that is not yet satisfactory answered. [71] argue that similarity alone is not sufficient to describe a concept. We need also feature correlations, a structure of the attributes that are internal to a concept, and background knowledge as already discussed in the previous subsections. Beyond that we also require a relation of the concepts to each other. One possibility to get away from this purely similarity-driven approach of label propagation is to consider groups of images instead of pairwise similarities.

### 2.4.2 Beyond Human Perception, Learning, and Inference

In the previous subsection, we discussed some future work strongly based on the insight of cognitive science. This focus on human object recognition might serve as a good starting point. But also human perception and inference has their weaknesses that might be tackled by computers. One of these shortcomings is the selective attention also known as *change blindness*. There are several studies showing that a human does not recognize large differences such as a complete different clothing of a person in a video sequence of the same situation when focusing on the conversation [62]. In [102], one person is exchanged by another while the other person explains the direction without noticing the exchange. Most famous is the *invisible gorilla* [17] that runs through a video sequences and most people overlook this disguised person. But 78 % of the people are sure to recognize unexpected objects [101] that is also called *memory illusion* meaning that we have the feeling of continuous attention because we cannot remember the unconscious moments. In this point, computers are trustworthy and this is one reason why most of the assembly line work or other production steps are done by a machine. Also in computer vision we can benefit from this advantage by completely analyzing video sequences (not partially like a human) or by scanning through millions of images to find prototypical examples of one class.

Another problem comes with the limited knowledge base of a human. Even if a person learns day and night, he will never be capable to acquire the entire knowledge and experience existing in our world. Also in the case that we bound this knowledge to a particular area for example a lawyer who read all cases to his topic or a doctor who is specialized to one organ. We cannot be sure that this specialist will remember the appropriate precedent or the disease pattern if it is needed. In contrast, with a computer we are able to get more information at the same time and to remind humans on the existence of some facts, e.g. to assist the diagnosis. This ability is also in computer vision of great importance as we can acquire more and better knowledge from the internet that might be helpful for semi-supervised learning.



**Fig. 2.26** Visualization of rare categories and their effect on our inference: (a) Wolpertinger a fake object, and (b) duckbill platypus a real object that seems like a fake as it mix up properties of different species

Finally, also human inference is highly dependent on the knowledge of a person. Sure we infer quickly the position of a glass and can grasp it within few seconds and we immediately recognize the *Wolpertinger*—a bavarian mythical creature—shown in Fig. 2.26 as a fake because no hare has a deer head and bird wings. But on the other side, rare species such as the duckbill platypus (Fig. 2.26 right) looks also like an elaborate fraud to us as if someone stick the duckbill on this animal. In fact, this species comes with an unusual appearance and atypical properties for a mammal such as laying eggs like a bird or a reptile, having a tail like a beaver, a bill like a duck, and foots like an otter. Assuming that we can collect more knowledge with a computer then this added information should also improve the inference beyond that of a human. In particular in the shown case from Fig. 2.26, a computer should be in a better position to decide which one is a fake. First, each imitation of the *Wolpertinger* looks different in comparison to images of the duckbill platypus. Second, we can also take into account the trustability of the source.

## References

1. Angluin D, Laird P (1988) Learning from noisy examples. *Mach Learn* 2:343–370
2. Argyriou A, Herbster M, Pontil M (2005) Combining graph Laplacians for semi-supervised learning. In: *NIPS*
3. Ashby FG (1992) Multidimensional models of categorization. In: *Multidimensional models of perception and cognition*. Erlbaum, Hillsdale, pp 449–483
4. Ashby FG, Todd WT (2011) Human category learning 2.0. *Ann NY Acad Sci* 1224:147–161
5. Balcan M-F, Blum A (2005) A PAC-style model for learning from labeled and unlabeled data. In: *COLT*
6. Balcan M-f, Blum A, Pakyan Choi P, Lafferty J, Pantano B, Rwebangira MR, Zhu X (2005) Person identification in webcam images: an application of semi-supervised learning. In: *ICML WS*
7. Baram Y, El-yaniv R, Luz K (2004) Online choice of active learning algorithms. *J Mach Learn Res* 5:255–291
8. Bauckhage C, Thureau C (2009) Making archetypal analysis practical. In: *DAGM*
9. Berg TL, Forsyth DA (2006) Animals on the web. In: *CVPR*
10. Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychol Rev* 94(2):115–147
11. Bischof H, Pinz A, Kropatsch WG (1992) Visualization methods for neural networks. In: *IAPR*

12. Blum A, Chawla S (2001) Learning from labeled and unlabeled data using graph mincuts. In: ICML
13. Buhmann JM, Zöller T (2000) Active learning for hierarchical pairwise data clustering. In: ICPR
14. Burl MC, Perona P (1996) Recognition of planar object classes. In: CVPR
15. Cebron N, Berthold MR (2009) Active learning for object classification: from exploration to exploitation. *Data Min Knowl Discov* 18(2):283–299
16. Cebron N, Richter F, Lienhart R (2012) “I can tell you what it’s not”: active learning from counterexamples. In: *Progress in artificial intelligence*
17. Chabris C, Simons D (2010) *The invisible gorilla: how our intuitions deceive us*. Crown Publishing Group
18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:341–378
19. Cohen B, Murphy GL (1984) Models of concepts. *Cogn Sci* 8(1):27–58
20. Cootes TF, Edwards GJ, Taylor CJ (1998) Active appearance models. In: ECCV
21. Cutler A, Breiman L (1994) Archetypal analysis. *Technometrics* 36(4):338–347
22. Daitch SI, Kelner JA, Spielman DA, Haven N (2009) Fitting a graph to vector data. In: ICML
23. Damasio A (1994) *Descartes’ error: emotion, reason, and the human brain*. Penguin Group
24. Davis J, Kulis B, Jain P, Sra S, Dhillon I (2007) Information-theoretic metric learning. In: ICML
25. Delaitre V, Fouhey DF, Laptev I, Sivic J, Gupta A, Efros AA (2012) Scene semantics from long-term observation of people. In: ECCV
26. Delalleau O, Bengio Y, Le Roux N (2005) Efficient non-parametric function induction in semi-supervised learning. In: AISTATS
27. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: CVPR, June 2009. IEEE
28. Dubout C, Fleuret F (2011) Tasting families of features for image classification. In: ICCV
29. Ebert S (2012) *Semi-supervised learning for image classification*. PhD thesis, Saarland University
30. Ebert S, Larlus D, Schiele B (2010) Extracting structures in image collections for object recognition. In: ECCV
31. Ebert S, Fritz M, Schiele B (2011) Pick your neighborhood—improving labels and neighborhood structure for label propagation. In: DAGM
32. Ebert S, Fritz M, Schiele B (2012) Active metric learning for object recognition. In: DAGM
33. Ebert S, Fritz M, Schiele B (2012) Semi-supervised learning on a budget: scaling up to large datasets. In: ACCV
34. Elhamifar E, Sapiro G, Vidal R (2012) See all by looking at a few: sparse modeling for finding representative objects. In: CVPR
35. Erickson MA, Kruschke JK (1998) Rules and exemplars in category learning. *J Exp Psychol Gen* 127(2):107–140
36. Everingham M, Van Gool L, Williams CK (2008) The PASCAL VOC
37. Farajtabar M, Shaban A, Reza Rabiee H, Rohban MH (2011) Manifold coarse graining for online semi-supervised learning. In: ECML
38. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. *IEEE Trans Pattern Anal Mach Intell* 28(4):594–611
39. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9):1627–1645
40. Fergus R, Weiss Y, Torralba A (2009) Semi-supervised learning in gigantic image collections. In: NIPS
41. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7:179–188
42. Fowlkes C, Belongie S, Chung F, Malik J (2004) Spectral grouping using the Nystrom method. *IEEE Trans Pattern Anal Mach Intell* 26(2):214–225

43. Freeman WT (2011) Where computer vision needs help from computer science. In: ACM-SIAM symposium on discrete algorithms
44. Fritz M, Black M, Bradski G, Darrell T (2009) An additive latent feature model for transparent object recognition. In: NIPS
45. Fussenegger M, Roth PM, Bischof H, Pinz A (2006) On-line, incremental learning of a robust active shape model. *Pattern Recognit* 4174:122–131
46. Gehler P, Nowozin S (2009) On feature combination for multiclass object classification. In: ICCV
47. Goldberg AB, Zhu X, Wright S (2007) Dissimilarity in graph-based semi-supervised classification. In: AISTATS
48. Grabner H, Leistner C, Bischof H (2008) Semi-supervised on-line boosting for robust tracking. In: ECCV
49. Hayward WG (2003) After the viewpoint debate: where next in object recognition? *Trends Cogn Sci* 7(10):425–427
50. Joachims T (1999) Transductive inference for text classification using support vector machines. In: ICML
51. Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2):263–291
52. Kant I (1781) *Kritik der reinen Vernunft*. Johann Friedrich Hartknoch Verlag. English edition: Kant I (1838) *Critique of pure reason* (trans: Haywood F)
53. Kaplan AS, Murphy GL (2000) Category learning with minimal prior knowledge. *J Exp Psychol* 26(4):829–846
54. Karlen M, Weston J, Erkan A, Collobert R (2008) Large scale manifold transduction. In: ICML. ACM Press, New York
55. Kato T, Kashima H, Sugiyama M (2009) Robust label propagation on multiple networks. *IEEE Trans Neural Netw* 20(1):35–44
56. Khosla A, Zhou T, Malisiewicz T, Efros AA, Torralba A (2012) Undoing the damage of dataset bias. In: ECCV
57. Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. *Psychol Rev* 99(1):22–44
58. Kulis B, Jain P, Grauman K (2009) Fast similarity search for learned metrics. *IEEE Trans Pattern Anal Mach Intell* 31(12):2143–2157
59. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: CVPR
60. Lee YJ, Grauman K (2009) Foreground focus: unsupervised learning from partially matching images. *Int J Comput Vis* 85:143–166
61. Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: CVPR. IEEE
62. Levin DT, Simons DJ (1997) Failure to detect changes to attended objects in motion pictures. *Psychon Bull Rev* 4(4):501–506
63. Li W, Fritz M (2012) Recognizing materials from virtual examples. In: ECCV
64. Li Y-F, Zhou Z-H (2011) Improving semi-supervised support vector machines through unlabeled instances selection. In: AAAI
65. Liu W, He J, Chang SF (2010) Large graph construction for scalable semi-supervised learning. In: ICML, pp 1–8
66. Lu Z, Jain P, Dhillon IS (2009) Geometry-aware metric learning. In: ICML
67. Medin DL, Schaffer MM (1978) Context theory of classification learning. *Psychol Rev* 85(3):207–238
68. Minda JP, Smith JD (2001) Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *J Exp Psychol Learn Mem Cogn* 27(3):775–799
69. Murphy GL (2002) *The big book of concepts*
70. Murphy GL, Allopenna PD (1994) The locus of knowledge effects in concept learning. *J Exp Psychol Learn Mem Cogn* 20(4):904–919

71. Murphy GL, Medin DL (1985) The role of theories in conceptual coherence. *Psychol Rev* 92(3):289–316
72. Nguyen HT, Smeynders A (2004) Active learning using pre-clustering. In: *ICML*
73. Nigam K, McCallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. *Mach Learn* 39:103–134
74. Nosofsky RM (1984) Choice, similarity, and the context theory of classification. *J Exp Psychol* 10(1):104–114
75. Osherson DN, Smith EE (1981) On the adequacy of prototype theory as a theory of concepts. *Cognition* 9(1):35–58
76. Osugi T, Kun D, Scott S (2005) Balancing exploration and exploitation: a new algorithm for active machine learning. In: *ICDM*
77. Parikh D, Grauman K (2011) Relative attributes. In: *ICCV*, November 2011. *IEEE*
78. Pazzani MJ (1991) Influence of prior knowledge on concept acquisition: experimental and computational results. *J Exp Psychol* 17(3):416–432
79. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag* 2(6):559–572
80. Pepik B, Stark M, Gehler P, Schiele B (2012) Teaching 3D geometry to deformable part models. In: *CVPR*
81. Pishchulin L, Jain A, Andriluka M, Thormählen T, Schiele B (2012) Articulated people detection and pose estimation: reshaping the future. In: *CVPR*
82. Ponce J, Berg TL, Everingham M, Forsyth DA, Hebert M, Lazebnik S, Marszalek M, Schmid C, Russell BC, Torralba A, Williams CKI, Zhang J, Zisserman A (2006) Dataset issues in object recognition. In: Ponce J, Hebert M, Schmid C, Zisserman A (eds) *Towards category-level object recognition*. LNCS. Springer, Berlin, pp 29–48
83. Pope A, Lowe DG (1996) Learning appearance models for object recognition. In: *Object representation in computer vision II*
84. Posner MI, Goldsmith R, Welton KE (1967) Perceived distance and the classification of distorted patterns. *J Exp Psychol* 73(1):28–38
85. Prabhakaran S, Raman S, Vogt JE, Roth V (2012) Automatic model selection in archetype analysis. In: *DAGM*
86. Rohban MH, Rabiee HR (2012) Supervised neighborhood graph construction for semi-supervised classification. *Pattern Recognit* 45(4):1363–1372
87. Rohrbach M, Stark M, Schiele B (2011) Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: *CVPR*
88. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. *Cogn Psychol* 8:382–439
89. Saffari A, Godec M, Pock T, Leistner C, Bischof H (2010) Online multi-class LPBoost. In: *CVPR*, June 2010. *IEEE*
90. Schiele B (2000) Towards automatic extraction and modeling of objects from image sequences. In: *Int sym on intelligent robotic systems*
91. Schiele B, Crowley JL (1996) Where to look next and what to look for. In: *IROS*
92. Schiele B, Crowley JL (1997) The concept of visual classes for object classification. In: *Scand conf image analysis*
93. Schiele B, Crowley JL (1998) Transinformation for active object recognition. In: *ICCV*
94. Schiele B, Crowley JL (2000) Recognition without correspondence using multidimensional receptive field histograms. *Int J Comput Vis* 36(1):31–52
95. Schiele B, Pentland A (1999) Probabilistic object recognition and localization. In: *ICCV*
96. Schnitzspan P, Fritz M, Roth S, Schiele B, Berkeley Eecs UC (2009) Discriminative structure learning of hierarchical representations for object detection. In: *CVPR*
97. Schohn G, Cohn D (2000) Less is more: active learning with support vector machines. In: *ICML*
98. Seeger M (2001) Learning with labeled and unlabeled data. Technical report, University of Edinburgh

99. Settles B (2009) Active Learning Literature Survey. Technical report, University of Wisconsin–Madison
100. Simon I, Snaveley N, Seitz SM (2007) Scene summarization for online image collections. In: ICCV. IEEE
101. Simons DJ, Chabris CF (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28(9):1059–1074
102. Simons DJ, Levin DT (1998) Failure to detect changes to people during a real-world interaction. *Psychon Bull Rev* 5(4):644–649
103. Sivic J, Russell BC, Efros AA, Zisserman A, Freeman WT (2005) Discovering object categories in image collections. In: ICCV
104. Smith E, Medin DL (1981) Categories and concepts. Harvard University Press, Cambridge
105. Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B (2006) Large scale multiple kernel learning. *J Mach Learn Res* 7:1531–1565
106. Stark M, Goesele M, Schiele B (2010) Back to the future: learning shape models from 3D CAD data. In: BMVC
107. Sternig S, Roth PM, Bischof H (2012) On-line inverse multiple instance boosting for classifier grids. *Pattern Recognit Lett*, 33(7):890–897
108. Sugiyama M, Rubens N (2008) Active learning with model selection in linear regression. In: DMKD
109. Talwalkar A, Kumar S, Rowley H (2008) Large-scale manifold learning. In: CVPR, June 2008, pp 1–8
110. Tong W, Jin R (2007) Semi-supervised learning by mixed label propagation. In: AAAI, vol 22
111. Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. *J Mach Learn Res* 2:45–66
112. Tong H, He J, Li M, Zhang C, Ma WY (2005) Graph based multi-modality learning. In: ACM multimedia
113. Torralba A (2011) Unbiased look at dataset bias. In: CVPR
114. Torralba BA, Russell BC, Yuen J (2010) LabelMe: online image annotation and applications. In: Proc IEEE
115. Tsang IW, Kwok JT (2006) Large-scale sparsified manifold regularization. In: NIPS
116. Tsuda K, Shin H, Schoelkopf B (2005) Fast protein classification with multiple networks. *Bioinformatics* 21:59–65
117. Vedaldi A, Gulshan V, Varma M, Zisserman A (2009) Multiple kernels for object detection. In: ICCV, pp 606–613
118. Vernon D (2005) A research roadmap of cognitive vision. Technical report, ECVision: the European research network for cognitive computer vision systems
119. Von Luxburg U, Radl A, Hein M (2010) Getting lost in space: large sample analysis of the commute distance. In: NIPS
120. Wang L, Chan KL, Zhang Z (2003) Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In: CVPR. IEEE Comput. Soc., Los Alamitos
121. Wang X, Han TX, Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: ICCV, September 2009. IEEE
122. Wang G, Wang B, Yang X, Yu G (2012) Efficiently indexing large sparse graphs for similarity search. *IEEE Trans Knowl Data Eng* 24(3):440–451
123. Weber M, Welling M, Perona P (2000) Unsupervised learning of models for recognition. In: ECCV
124. Welinder P, Branson S, Belongie S, Perona P (2010) The multidimensional wisdom of crowds. In: NIPS, pp 1–9
125. Wiskott L, von der Malsburg C (1993) A neural system for the recognition of partially occluded objects in cluttered scenes. *Int J Pattern Recognit Artif Intell* 7(4):935–948
126. Yang L (2006) Distance metric learning: a comprehensive survey. Technical report, Michigan State University



127. Yang X, Bai X, Köknar-Tezel S, Latecki LJ (2013) Densifying distance spaces for shape and image retrieval. *J Math Imaging Vis* 46:12–28
128. Zaki SR, Nosofsky RM (2007) A high-distortion enhancement effect in the prototype-learning paradigm: dramatic effects of category learning during test. *Mem Cogn* 35(8):2088–2096
129. Zaki SR, Nosofsky RM, Stanton RD, Cohen AL (2003) Prototype and exemplar accounts of category learning and attentional allocation: a reassessment. *J Exp Psychol Learn Mem Cogn* 29(6):1160–1173
130. Zhang Z, Zha H, Zhang M (2008) Spectral methods for semi-supervised manifold learning. In: *CVPR*
131. Zhang K, Kwok JT, Parvin B (2009) Prototype vector machine for large scale semi-supervised learning. In: *ICML*. ACM Press, New York
132. Zhou D, Bousquet O, Navin Lal T, Weston J, Schölkopf B (2004) Learning with local and global consistency. In: *NIPS*
133. Zhu X, Goldberg AB, Khot T (2009) Some new directions in graph-based semi-supervised learning. In: *ICME*

<http://www.springer.com/978-1-4471-5519-5>

Advanced Topics in Computer Vision

Farinella, G.M.; Battiato, S.; Cipolla, R. (Eds.)

2013, XIV, 433 p. 218 illus., 180 illus. in color.,

Hardcover

ISBN: 978-1-4471-5519-5