Pattern Recognition Letters 33 (2012) 218–226

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

IAIR-CarPed: A psychophysically annotated dataset with fine-grained and layered semantic labels for object recognition

Yang Wu [a,b,*], Yuanliu Liu [a], Zejian Yuan [a,*], Nanning Zheng [a]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, 28 West Xianning Road, Xi'an 710049, Shaanxi, PR China
[b] Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan

ARTICLE INFO

Article history:
Received 12 July 2010
Available online 20 October 2011
Communicated by S. Sarkar

Keywords:
Object recognition
Image database
Object detection
Pedestrian detection
Psychophysical experiments

ABSTRACT

Unlike many other object recognition datasets which provide either category-level or within-category annotations, we introduce a novel dataset called "IAIR-CarPed" with layered semantic labels ranging from categories to fine-grained subcategories. These labels are collected from 20 subjects via strict psychophysical experiments. To the best of our knowledge, it is the first time that an object recognition dataset is built in this way to represent the adaptive and in-depth interpretations of objects in human vision. This dataset focuses on "car" and "pedestrian" which are two representative categories important in real applications. It contains 3132 images collected from pictures taken under various conditions and 8567 objects carefully annotated by all the 20 subjects. Besides fine-grained and layered semantic labels, five types of detailed visual difficulties of these objects are also provided, which can be adopted to evaluate the representation and generalization abilities of the recognition systems against individual difficulties. We present here the details of building this dataset, its statistics and properties, and then discuss possible applications of it with some primary experimental results.

1. Introduction

Visual object recognition is an active research topic in computer vision, with great achievement made on both basic-level object categorization (e.g. car vs. pedestrian) and some specific object identification (e.g. face recognition). However, little has been done on the semantic recognition lain in between, which is called "**fine-grained**" categorization (Yao et al., 2011). Though it is worthy of looking into this largely unexplored subordinate categorization problem which can be clearly defined and constrained, we would like to step back and look at how humans recognize objects. As shown in Fig. 1, even for the same object category (car), different instances are usually perceived to be different, varying from categorical labels to detailed interpretations according to their perceptual information. Therefore, we think it is more interesting to go a step further from pure fine-grained categorization to "**layered**" recognition, i.e., coarse or fine-grained categorization depending on the visual appearance of objects.

Since the fine-grained and layered recognition is a new idea, there is no existing datasets directly suitable for that. Therefore, we propose here a benchmark dataset named "IAIR-CarPed"[1] for promoting the research along this direction. As its name shows, IAIR-CarPed focuses on two representative object categories (car and pedestrian) with 8567 annotated object instances in 3132 images, labeled by 20 subjects through psychophysical experiments, which are inspired by the research on human rapid object recognition in psychophysics (Serre et al., 2005, 2007). Briefly speaking, IAIR-CarPed dataset has the following three contributions to the research community.

- It provides fine-grained and layered semantic labels for all the applicable objects in the dataset.
- It contains independent annotations of five different types of visual difficulties, which can be used for evaluating the robustness and the data-modeling ability of different features and/or learning algorithms.
- The votes from 20 subjects to different semantic labels of the same object instance reveal the confusions of human vision among such concepts, which are valuable for guiding the learning of human-like recognition algorithms and evaluating the similarity of them to human vision.

As a by-product, this dataset is also good for the traditional task of object detection due to the large within-class variations of the data.

2. Related work

For object recognition, undoubtedly category-level semantic labels have been extensively explored, resulting in numerous

* Corresponding authors at: Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, 28 West Xianning Road, Xi'an 710049, Shaanxi, PR China. Tel.: +86 29 8266 8802x8038; fax: +86 29 8266 8672 (Y. Wu).
E-mail addresses: wuyang0321@gmail.com (Y. Wu), liuyuanliu@stu.xjtu.edu.cn (Y. Liu), zjyuan@aiar.xjtu.edu.cn (Z. Yuan), nnzheng@mail.xjtu.edu.cn (N. Zheng).

[1] http://mm.media.kyoto-u.ac.jp/members/yangwu/research/FGLR.html.

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.10.003

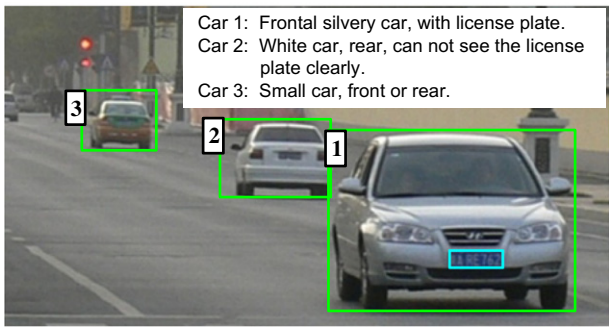


Fig. 1. Illustration of the human perception results of different object instances, which tend to be fine-grained and layered.

datasets which can be further categorized into subgroups for different recognition tasks including presence vs. absence image classification (e.g. PASCAL VOC (Everingham et al., 2010)), object detection/localization (e.g. CMU-MIT frontal face images (Schneiderman and Kanade, 2000) and INRIA person dataset (Dalal and Triggs, 2005)), object categorization (e.g. COIL-100 (Nene et al., 1996) and Caltech-101 (Fei-Fei et al., 2006)), and semantic segmentation or namely image parsing (e.g. LabelMe (Russell et al., 2008) and Lotus Hill dataset (Yao et al., 2007)). It is worth mentioning that some recent datasets (Yao et al., 2007; Deng et al., 2009) take into account the hierarchical relationships of the categories based on the ontology of WordNet (Fellbaum, 1998), trying to follow the way humans organize such concepts. Another important direction is the research on video-based object recognition, for which a carefully annotated database on real street scenes has recently been published (Brostow et al., 2009).

Meanwhile, within-category semantic labels attracts more and more attention in the past few years. They are important middle-level representations for category-level recognition, while at the same time meaningful for in-depth interpretations of objects. As a typical type of within-category labels, attributes have already been proven to be helpful for category-level recognition, as shown by Lampert et al. (2009) and Farhadi et al. (2009). The rich literature on using the divide-and-conquer philosophy for detecting objects with large within-category variations implicitly uses the subcategorial concepts like the pose of the face (Huang et al., 2007). There are also some prior research on within-category classification/description, or recently proposed fine-grained categorization (Yao et al., 2011). The work on multiplicative kernels (Yuan et al., 2008) manages to classify the view angles of cars, while another work directly targets at within-object classification (Aghajanian et al., 2009). Farhadi et al. (2009) proposes learning to describe the objects by their attributes so that unseen objects can be somehow interpreted and the unusualities of objects can be identified. Besides of global description, part-level semantics are also helpful for in-depth understanding of objects. A representative example is the And-Or decomposition in Lotus Hill dataset (Yao et al., 2007).

Unlike these efforts, we are focusing on the layered interpretation of objects in human vision, whose results can be categorial, subcategorial or even going down to object parts. Therefore, we are not aiming at providing a larger dataset, more categories, or deeper annotations, but focusing on the layered annotation of objects belonging to the same category even when they appear in the same image. We try to make it represent the real perceptual results in human vision as well as possible, and provide the community a good starting point on the research of fine-grained and layered object recognition.

3. Dataset construction

3.1. Fine-grained and layered semantic labels

For each category that we are interested in, there are many different visual aspects (or attributes) which can be used for defining fine-grained subcategories, such as brand, model, color, function for cars, and race, gender, age, occupation for pedestrians. Though straight-forward and valuable to human beings, many of such semantic labels are hard for collecting representative data and also hard to current object recognition algorithms. Thus, we would like to start with some easy-to-operate and visually distinctive attributes, such as the relative orientation of the objects to the camera and the clearness of some key parts of the objects. Fig. 2 presents the detailed categorization labels that have been used for annotating the proposed dataset.

3.2. Data collection and preprocessing

We were trying to make the dataset as representative as possible when we collected the data. There are two key points which have guided our collection: (a) to make sure there are sufficient examples for each output state (node on the semantic tree structure), and (b) to make the dataset as generic and natural as possible (i.e., large within-class variation).

More concretely, the images are mostly from pictures taken from natural scenes in a big city including campus, park, street, mountain, and rural areas, while some others are from photos taken under unusual weather conditions (like thick fog, snow and sand storm) or special shooting angles (like sunrise/sunset and backlight) either from the Internet or taken by us in controlled environments. We preprocessed these images to make sure they are of sufficient quality, properly cropped and resized (to a fixed size of 512×384 for consistency and convenience).

4. Annotation

Each of the image in IAIR-CarPed contains at least one instance of car or pedestrian, and we label all the object instances (cars and/or pedestrians) satisfying our demands: visually tellable, with less than two thirds' occlusion, truncated less than one half, and above the minimum sizes (car 30-pixel wide and pedestrian 45-pixel high).

Our annotations are mainly two types: semantic labels and geometric labels. The semantic labels are the nodes shown in Fig. 2 while the geometric ones are the bounding boxes of the objects and their key parts (if clear). Fig. 3 shows some examples of the annotated images.

4.1. Visual difficulties and object localization

As stated in Section 3.2, the dataset has great variations in the data, including significant visual difficulties. We follow the suggestion of Sven Dickinson (2009) to isolate the visual difficulties by categorizing them into five different types to aid the analysis of the robustness of different features and/or recognition algorithms and thus indicate improving directions. We check the five types one by one for each object instance, resulting in a 5-dimensional vector. Each element of it is a binary value, for which 1 indicates that the corresponding difficulty exists while 0 means the opposite. The five difficulties are defined as follows.

- **Occlusion**. Occluding less than 2/3 of the object area is considered, while larger occlusion or very little occlusion (less than 1/5) is ignored.

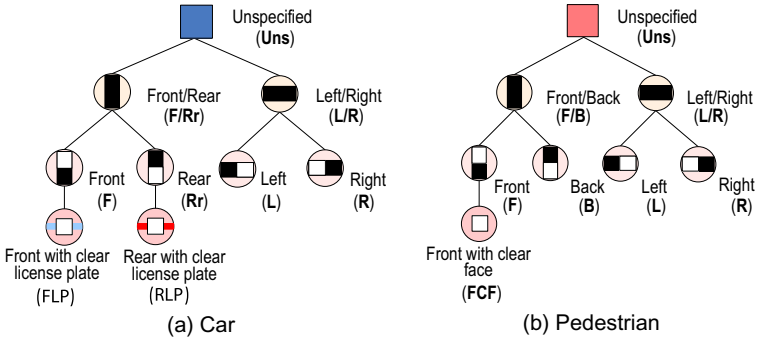


Fig. 2. The tree structure of the fine-grained and layered semantic labels for each category. From the top to the third layer, the relative orientation of the objects to the camera becomes more and more certain and specific (the root just indicates the category of the object but the orientation is totally unspecified), while the fourth layer is on the clearness of the key parts when the object is facing certain orientations. The shape of each node is designed to illustrate the semantic meaning associated to it. Proper abbreviations are provided beside the nodes for easier references of them in this paper.

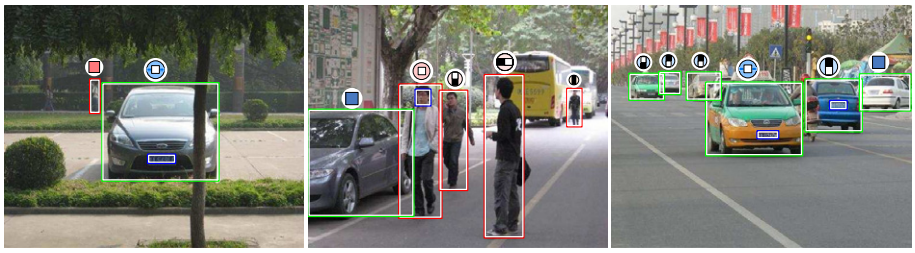


Fig. 3. Example images and their annotations. Green and red bounding boxes represent annotated cars and pedestrians respectively, while the sign above each object illustrates its semantic label. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- **Truncation**. 1/5–1/2 of the object is truncated, based on the area of the bounding box.
- **Nonuniform illumination**. Caused by shadows of other objects like trees, or uneven lightings at night.
- **Low contrast**. Under conditions of low illumination (e.g. in the rain), or motion blur.
- **Infrequent shape**. Car types besides sedan are treated as infrequent ones, and pedestrians with unusual pose or small children are of this type. Actually, in the IAIR-CarPed dataset, they are infrequent.

Fig. 4 shows some annotated objects with one of the five visual difficulties. Though most of the time occlusion and truncation are treated as the same in the literature, we would like to differentiate them so that they can refer to different scenarios which can be handled in different ways. We label the bounding box of an object as tight as possible to just bound its visible part of without objective estimation of its missing parts. Then we define "occlusion" and "truncation" as: **occlusion** – some part(s) of the object in the labeled bounding box is/are occluded by other objects, and **truncation** – the labeled bounding box is considered to be smaller than the actual bounding box if the object is fully visible. Note that our definition of these two concepts are not mutually-exclusive. By separating these two concepts, it is expected that pure occlusion without truncation is relatively easier to handle since the extension of the object is still tellable, while the cases in which truncation is involved are very challenging because the real bounding boxes are hidden. Since the difficulty labels and the bounding boxes are deterministic, we had them labeled by one single person and checked by another. There were very few cases that the first person made mistakes or the cases were really hard to annotate, then these two people discussed about them and made an agreement together.

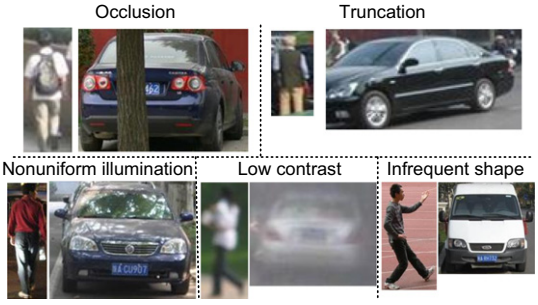


Fig. 4. Visual difficulties and their actual examples in IAIR-CarPed dataset.

4.2. Orientation

The orientation of objects is the main cue used by us to design the fine-grained and layered semantic labels. It can be seen that these labels cannot be easily separated for some ambiguous cases and different people may have different opinions. Therefore, we used the strategy of having a group of people (usually several dozens) to label them independently and then integrate the results, as widely used in psychophysical experiments.

However, we immediately found that human vision is so powerful that by careful examination with enough time, it can use very subtle difference to tell the exact orientations even though the objects are very small or look ambiguous at a glance. Nevertheless, we usually do the so called rapid recognition to efficiently recognize surrounding objects, as the results are usually sufficient though some of them are ambiguous. We followed the research in psychophysics and neuroscience on rapid recognition (Serre et al., 2007), and extended it from binary presence vs. non-presence image classification to our fine-grained and layered recognition. Unlike the assembling of a sequence of binary labeling processes for building ImageNet database (Deng et al., 2009), a

one-shot multiclass assignment was used for our task. Our experiments show that humans perform pretty well (compared to time-unlimited recognition) without much effort, which proves the effectiveness of fine-grained and layered recognition.

We hired 20 undergraduate or graduate students to perform the psychophysical experiment designed according to the one presented in (Serre et al., 2007). Fig. 5 illustrates its procedure. At each time, the machine randomly picks an unlabeled object instance (cropped from the original image using its bounding box with a small margin), and shows it against a black background for a stimulus time ranging from 20 ms to 50 ms, in inverse proportion to the scale of the object. When the stimulus time is less than 50 ms, an interstimulus interval (ISI) is shown with the black background only until the stimulus onset asynchrony (SOA), which is the stimulus plus the interstimulus interval, lasts as long as 50 ms. After that, a mask with random noises with the same size as the image shown before is displayed for 80 ms. This is designed to block possible back-projection from memory. Then the mask disappears, leaving only the background. The subject participating in the experiment is asked to tell the orientation of the object by assigning one of seven labels (given the object category) in the first three layers of the semantic tree to this object. This is done by pressing a predefined key and usually it takes less than 1s if the subject sees the object. After saving the label, the system transfers to the next one, and a new round begins.

The reason for varying the stimulus time is to compensate the response time of the eyes for adapting to the downsizing of the object. Though we have not found the psychophysical or biological evidence for the exact relationship between the scale of the object and the stimulus time needed for rapid recognition, experimentally we found that such a simple strategy can generate reasonably good results (compared to time-unlimited judgement).

Note that for objects with at least one of the three visual difficulties of occlusion, truncation, and low contrast, rapid recognition is very hard to generate reliable results. These objects seem to need selective visual attention or even visual reasoning which may be related to the feedback path in the cortex. Therefore, we designed a separate experiment for labeling these objects by extending the stimulus time to as long as 500 ms.

4.3. Key part clearness

In addition to the orientation, we also labeled the clearness of the key parts for objects with certain orientations: frontal/rear cars and frontal pedestrians. However, it is impossible to put the key part clearness labeling directly into the rapid recognition framework, because it is usually competing with the recognition of the object orientation as there is no time for saccade. Therefore, we chose to label the orientation first by forcing the subjects to focus on the global appearance of the object, and after getting the final orientation of the objects through the annotation integration (see Section 5.1), we asked them to label the key part clearness of frontal/rear cars and frontal pedestrians by just looking at the areas where the key parts may appear.

Note that it is not clear on how to define the clearness of the key parts properly, because we do not know the exact perceptual threshold of it and how this threshold can be used for consistent labeling by many subjects. After plenty of observation, we defined it as: for license plate the clearness means that most of the characters on it are readable, and for face it means that the facial features (especially the eyes) of it are tellable. The clearness check seems to be more complex than the simple presence classification, so rapid response may be unreliable. Therefore, enough stimulus time (1 s) was given to it.

5. Statistics

5.1. Voting for annotation integration

Since the key part clearness labeling is a binary classification problem, a simple majority voting was used to integrate the labels, i.e., the clearness is confirmed if and only if more than half of the labels are "yes". For object orientation labeling which is a multi-class classification problem, though normal voting is a simple and fair way to integrate the labels, it is not suitable for our problem. When the orientation is ambiguous, people may choose the semantic label in the two upper layers of the tree, but may also choose either one of its children (i.e., a more specific orientation) if they think they can somehow weakly tell its orientation. These uncertain judgements result in the diversity of votes for the visually ambiguous object, for example, 5 to "front/back", 6 to "front", 5 to "back", and 4 to the others, then it will be dangerous by choosing "front" as its ground truth, as most people do not agree with that. Therefore, we propose a transferred super-majority voting strategy as shown in Algorithm 1 to integrate the results.

Algorithm 1: Transferred Super-majority Voting:

Require:

The semantic tree structure $\mathcal{T} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{N_1, \ldots, N_n\}$ are nodes and $\mathcal{E}$ are edges (if $(N_i, N_j) \in \mathcal{E}$, then N_i is a parent of N_j);

The original votes $\mathbf{v} = (v_1, \ldots, v_n)^{\mathrm{T}}$ of the input object;

A group of predefined winning thresholds $\mathbf{t} = (t_1, \ldots, t_n)^{\mathrm{T}}$.

Ensure:

The final label index of the input object $l \in \{1, \ldots, n\}$.

1. Normal voting: $l = \arg\max_i v_i$. Note that if l has multiple initial values due to equal votes, then go through the following steps until termination with each one of them, and pick the one lies deepest in the tree $\mathcal{T}$ as the final label index l from all the results
2. Thresholding and decision making: if $v_l / \sum_{i=1}^{n} v_i \geqslant t_l$, then terminate and return l
3. Subtree integration: $\forall N_i \in \mathcal{V}$, if $(N_l, N_i) \in \mathcal{E}$, then $v_l = v_l + v_i$, $v_i = 0$
4. Termination or label transferring: if $v_l / \sum_{i=1}^{n} v_i < t_l$ and $\exists i, (N_i, N_l) \in \mathcal{E}$, then $l = i$ and go back to step 3, else terminate and return l

The goal of the transferred super-majority voting algorithm is to find the most fine-grained label for the object which has enough supports from all the original votes compared to predefined thresholds. In our case, the thresholds for the nodes in the same level of the tree are set to be the same, and the ones for the three different semantic levels are set to be 0, 0.8 and 0.7, respectively from top to bottom. We found that such a setting results in semantically plausible final labels (by our checking afterwards) for the annotated objects. An example of the transferred super-majority vote is shown in Fig. 6.

We provide both the original labels and the final voted label in the annotation of each object, so that people can choose to use our proposed label as the ground truth or generate their own result from the original labels based on any other vote integration strategies when needed. Fig. 7 presents two typical examples of the final information for the annotated objects.

5.2. Semantic confusions of humans

The original votes not only provide evidence for the final label assignment, but also reveal the recognition ability of humans on

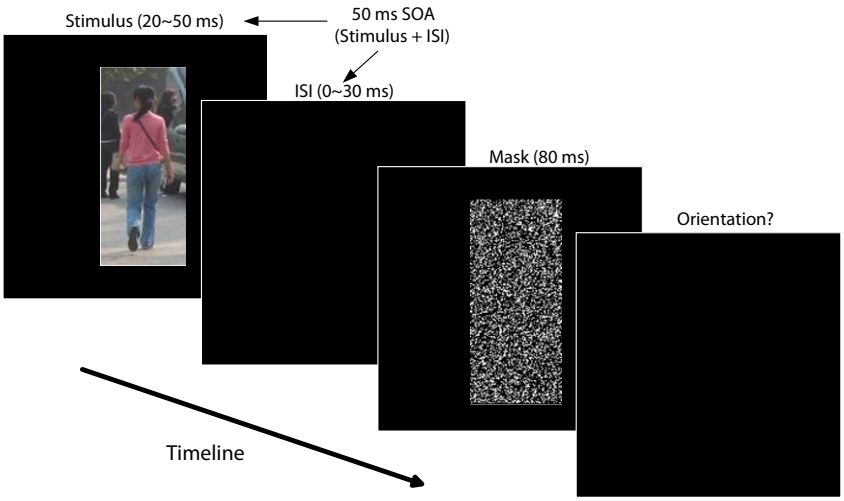


Fig. 5. Fine-grained and layered semantic annotation through rapid recognition. Please refer to Section 4.2 for details.

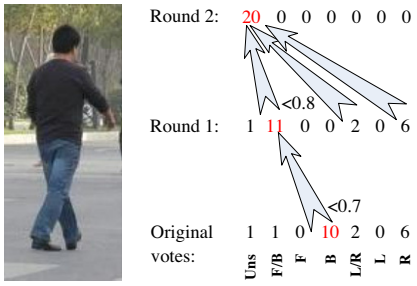


Fig. 6. A real example of the transferred super-majority voting for label integration. In this example, though the majority of the votes are "back", they cannot go beyond the threshold 0.7. After two rounds of label transferring and subtree integration, the final label is "unspecific", which is a suitable assignment with respect to the original votes.

distinguishing different semantic concepts. We treat the integrated label as the ground truth of the object, and the votes to other labels as pseudo mistakes. By computing the average percentage of the number of votes to node N_j while its voted ground truth is N_i, we can see how likely humans may mistake the semantic label associated with N_i for the one associated with N_j. We believe that the semantic confusions of humans are valuable for guiding the learned recognition model to generate results as close to those of the humans as possible.

We discuss here only the confusions between different orientational labels, as they can represent the human confusions more clearly and intuitively. When working on all the semantic labels including the clearness of the key parts, one only needs to separate the votes to the third layer and the fourth layer by the key part clearness label of these objects, then an augmented confusion matrix can be got.

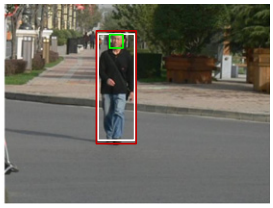
After getting the final labels from the voting, we compute the confusion matrices using the original votes, which are referred to as **original confusions**. To see how people make mistakes on different data, we compute the confusion matrices on two different subsets of the two categories respectively: set "S" contains objects without any visual difficulties, i.e., only simple examples, while set "D" is the complementary subset, containing objects with at least one special difficulty. Note that set "S" is labeled though rapid recognition only, while most of the examples in set "D" are not labeled by rapid recognition. The results are shown in the upper row of Fig. 8.

Though the original confusions directly represent the recognition results of humans, following them may end up with aggressive predictions which might have a high probability of making mistakes. Therefore, we propose using the conservative confusions computed by the adjusted votes after transferred super-majority voting. By doing so, the original confusion matrices are mapped into the ones shown in the second row of the same figure, which are referred to as **adjusted confusions**. Such an adjustment has significant effects on reducing the confusions from ambiguous labels to their child labels.

5.3. Data distributions

Scale distributions. Intuitively, objects belonging to higher layers tend to have smaller scales. To verify it and show the actual distributions of our data, we compute an object scale histogram for each semantic label as shown in Fig. 9, excluding the truncated objects whose actual size is unknown.

Bbox	[53 142 464 275]
Label	9
Label_hit	[0 0 0 0 0 0 0 0 20]
Original_label	<1x20 double>
Category	'car'
Keypart_bbox	[]
Difficulty label	[1 0 0 0 0]



Bbox	[182 50 255 270]
Label	4
Label_hit	[1 0 19 0 0 0 0 0]
Original_label	<1x20 double>
Category	'pedestrian'
Keypart_bbox	[202 56 233 87]
Keypart_label	<1x20double>
Difficulty label	[0 0 0 0 0]

Fig. 7. Examples of the final information provided for each object. "Label_hit" stands for the transferred votes for each semantic label in the tree structure in a depth first order including the key part clearness label, while "Label" is the index of the final label after vote integration. "Original_label" and "Keypart_label" are the original labels from subjects.

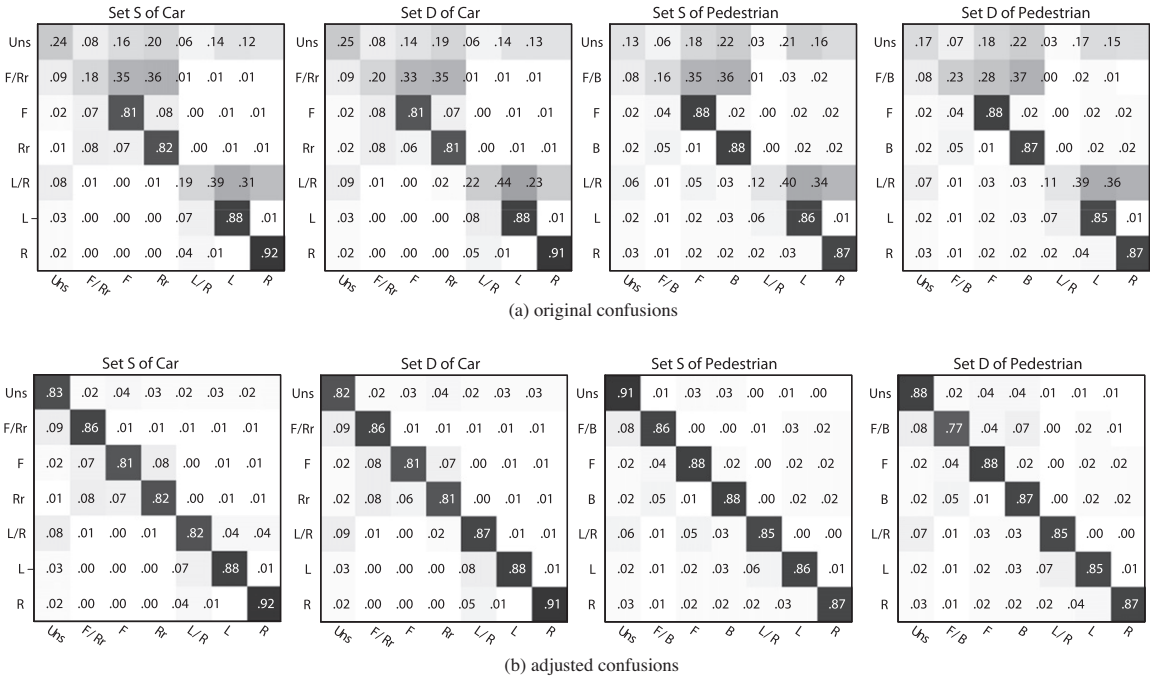


Fig. 8. The human confusions on object orientations through rapid recognition. The upper row shows the confusions computed using the original labels which are aggressive, while the lower row presents the adjusted ones which look much more conservative.

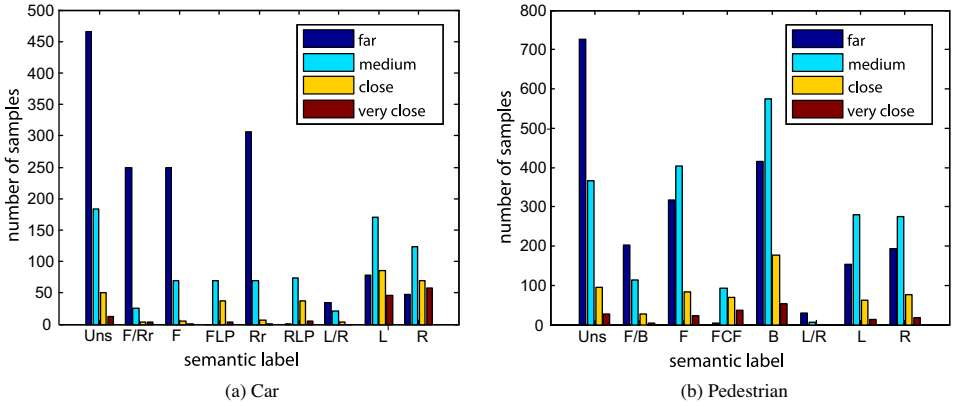


Fig. 9. The scale distributions of the objects on the semantic subsets. We briefly quantize the scales into four groups (far, medium, close, and very close) according to the width of the car and the height of the pedestrian respectively. More specific, for cars, the four groups are [30,128], (128,256], (256,384] and (384,512], while for pedestrians they are [45,96], (96,192], (192,288], and (288,384].

Difficulty distributions. Since we labeled the presence of the five difficulties independently, there are totally 2^5 combinational states. We list eight of them with their distributions on the semantic subsets in Table 1. The five sets with only one single type of difficulty are denoted by "D1" to "D5", respectively, while the whole dataset is denoted by "SD", and "SD-D2" means that the truncated examples ("D2") are excluded. As mentioned before, due to the fact that we have only labeled the tight bounding boxes for truncated objects, it is not easy to use these examples as the actual extents of them need to be estimated during training. Therefore, the six sets without truncated samples are recommended for common experiments. Briefly speaking, for these six sets, the distributions of examples belonging to different semantic subsets are similar, though the total number of examples varies.

Training and testing data. To make the dataset ready for performance comparison, we split the whole image set (containing 3132 images) into two equally sized subsets for training and testing by random sampling. The indices of images for training and testing are released along with the dataset.

6. Applications

IAIR-CarPed may have three applications. The first one is within-category object classification using its fine-grained and layered semantic labels which is called "fine-grained and layered object classification" in this paper. In this application, the bounding boxes of the objects are supposed to be given. Another application is the traditional object detection using only the category labels and bounding boxes. The third one is to do simultaneous classification and localization, called "fine-grained and layered object recognition" here. In this section we present some primary experiments on the first two with our analyses, while at the same time discuss possible solutions to the third, leaving their implementations for future exploration due to its complexness.

6.1. Fine-grained and layered object classification

To model the structural relationships between the fine-grained labels, one can use any maximum-margin based discriminative

Table 1
Distributions of the number of objects in subsets with various difficulties. 'S' and 'D' in the second column stand for "simple" and "difficult" respectively. The abbreviations are set notations, while the numbers in the parentheses are the corresponding difficulty labels. '1' means the difficulty exists while '0' means not, and 'x' means '0' or '1'.

Category	Set	Uns	F/Rr (B)	F	FLP (FCF)	Rr (B)	RLP	L/R	L	R	All
Car	S (00000)	388	167	233	84	265	100	20	253	190	1700
	D1 (10000)	136	24	34	4	36	3	25	57	74	393
	D2 (01000)	67	14	16	5	36	2	5	40	30	215
	D3 (00100)	24	14	12	11	21	4	1	6	2	95
	D4 (00010)	41	27	7	0	11	0	3	10	7	106
	D5 (00001)	28	14	14	7	27	6	0	17	5	118
	SD-D2 (x0xxx)	711	283	326	109	384	118	60	380	299	2670
	SD (xxxxx)	953	320	362	116	454	123	106	476	382	3292
Pedestrian	S (00000)	591	130	569	158	760	0	15	316	336	2895
	D1 (10000)	198	24	89	27	276	0	3	84	114	815
	D2 (01000)	41	3	50	12	43	0	2	23	18	192
	D3 (00100)	6	1	7	3	9	0	1	0	2	29
	D4 (00010)	265	164	91	1	75	0	5	62	51	714
	D5 (00001)	46	9	31	7	36	0	5	31	36	201
	SD-D2 (x0xxx)	1215	348	824	201	1219	0	35	507	562	4911
	SD (xxxxx)	1309	355	910	216	1302	0	37	544	602	5275

structured learning algorithms like SVMstruct (Joachims et al., 2009). However, we choose to use a relatively more efficient learning algorithm named SOnline, for which the details can be found in (Wu et al., 2009). There are two key components in this model: the joint input–output feature map and the loss function. Instead of designing a complex joint feature map and learn a common model for all possible outputs, we decouple the joint representation by learning different models for different outputs and thus let the feature representation only depends on the input sample. By doing so, we only need to make sure the features are able to differentiate the fine-grained labels. Note that the orientational information and key part clearness information are better to be represented by different features as they are very different patterns. To make things easier and clearer, we choose to do some primary experiments on the orientational labels (i.e., the first three layers) only,[2] and use dense HOG features (denser gridding than the original one (Dalal and Triggs, 2005)) followed by a PCA dimension reduction (the same as (Felzenszwalb et al., 2010)) as input feature representation. About the loss function, instead of the traditional zero-one loss and many hand-designed tree-structured losses, we propose to utilize the human confusion statistics (confusion matrix) for designing a new one which is called "**human confusion loss**". Details on this loss function and extensive comparisons between it and other losses can be found in our previous study on cost-sensitive object classification (Wu et al., 2011),[3] which presents initial experiments on the proposed dataset. Experimental details can also be found there.

Besides the research on loss function, it is also interesting to see how the algorithm performs on the subsets with different visual difficulties. To complement the experiments on the whole dataset (Wu et al., 2011), we present here more results in this direction. We train the model on set "S" and test it on the subsets with only one visual difficulty in them, such as "D1", "D3", "D4" and "D5" ("D2" is not used due to its special ground truth), in comparison to the test performance on set "S". As shown in Table 2, the robustness to certain difficulty not only depends on the feature representation, but also depends on the object category, i.e., the data itself. For the four different difficulties, low contrast (in set "D4") consistently reduces the performance as it weaken the image gradients for HOG features, while occlusion (in set "D1") has much larger influence on cars than pedestrians. Note that the occlusion defined by us does not contain truncation, so that cars are usually occluded by pedestrians or trees which may influence the orientation classification a lot, while pedestrians are more often to be occluded by bags which have little impact on that. The other two visual difficulties have opposite impacts on these two categories: nonuniform illumination (in set "D3") hurts cars a lot but little to pedestrians, while infrequent shape (in set "D5") is a big problem for pedestrians but not as disturbing for cars. Such contrasting results mainly due to the properties of the data but not the features and the algorithms. Nonuniform illumination has so few examples for pedestrians (as shown in Table 1) that its results are not statistically meaningful, while the infrequent shape for cars is not challenging as it stands for those vehicles having very similar shape to sedans. This is just a demo on how these difficult subsets can be used, it will be more meaningful if different features and/or different algorithms are compared on them, which is open to the public.

Table 2
Robustness of the proposed algorithm against individual visual difficulties. The model is trained on set "S" using human confusion loss and evaluated using the same loss.

Category	Test set				
	S	D1	D3	D4	D5
Car	0.1659	0.3146	0.1938	0.3430	0.1357
Pedestrian	0.2905	0.3150	0.1748	0.4086	0.4598

6.2. Object detection

There are a lot of publicly available datasets for car and pedestrian detection, such as the widely used UIUC side-view car detection dataset (Agarwal et al., 2004), a multiview car detection dataset (Yuan et al., 2008) collected from other general purpose databases, the small scale pedestrian detection datasets collected in different scenarios (Dalal and Triggs, 2005; Wu and Nevatia, 2009), the recently proposed large scale pedestrian detection datasets captured in street scenes only (Enzweiler and Gavrila, 2009; Ess et al., 2007; Dollar et al., 2009), and the challenging unconstrained car and people data in the PASCAL VOC challenges (Everingham et al., 2010). Even though, the IAIR-CarPed dataset is still a valuable benchmark for car and pedestrian detection which cannot be replaced, as it has two remarkable advantages: (a) **fine-grained within-category labels** which can be explored for boosting detection performance; and (b) **visual difficulty labels** which enables the in-depth study on the robustness of features/algorithms.

[2] The model can be easily extended to deal with all of the semantic labels as long as the features and the loss function are properly defined.

[3] In that paper "loss" is called "cost" due to the context of cost-sensitive classification.

IAIR-CarPed, unlike the other datasets, contains no specific negative images for training and testing, which are considered by us to be unnecessary. Actually, the subwindows which have less than 50% overlap with the labeled ground truths can be used as negative examples. Such a strategy not only differentiates the objects and their backgrounds, but also covers possible false positives due to misalignments, like the region within two cars.

We report experimental results of two detection algorithms: one is a baseline algorithm using HOG features with both dense grids (Wu et al., 2011) and sparse grids (Felzenszwalb et al., 2010) followed by a linear SVM classifier, and the other is the part-based deformable model (Felzenszwalb et al., 2010), which is the state-of-the-art. We trained our baseline algorithm on set "SD-D2" which contains all the applicable object instances, and tested the trained model on six different subsets: "S", "D1", "D3", "D4", "D5" and "SD-D2". For the part-based deformable model, we used the pre-trained model on the PASCAL VOC 2008 challenge dataset to directly test on the six subsets. Because the minimum size of the annotated objects in IAIR-CarPed is about twice smaller than that of the HOG templates, we upscaled the images by two times for training and testing.

The experimental results are shown in Fig. 10, which is measured by the per-image Detection Error Tradeoff (DET) curves on a log–log scale, i.e., miss rate versus false positive per image (FPPI), as proposed in (Dollar et al., 2009). Besides the per-image DET curve, the Precision-Recall (PR) curves are also good for measuring the detection performance, which have been used in PASCAL VOC challenges (Everingham et al., 2010). Note that for the purpose of comparing detection performances of one algorithm on different subsets of the dataset, one should look into the per-image DET curves, because the PR curves bias on larger subsets.

From the results in Fig. 10, we can see that comparing to the baseline algorithm, part-based deformable model generalizes better to objects with infrequent shape and occlusions, proving the benefit of using the latent part-based deformable model. Since both of these two algorithms are based on HOG features, they all suffer from the low contrast difficulty which weakens the gradients. Though these two algorithms are not completely comparable as they were trained on different data, by comparing their results we can still get the inspiration that a *divide-and-conqueror* strategy can promote the performance. The superiority of the part-based deformable model is more significant on cars than on pedestrians because the two components of the car model trained on PASCAL are all applicable to the IAIR-CarPed dataset while only one component of the two-component person model can be used to detect the full-body pedestrians. Therefore, if a detection model can take the advantage of the fine-grained and layered labels provided by the IAIR-CarPed dataset, it may significantly boost the performance.

As presented in (Dollar et al., 2009), the miss rate of the part-based deformable model at 1 FPPI on the INRIA person dataset is 0.21, while for subset "S" and subset "SD-D2" of the IAIR-CarPed dataset, it has a value of 0.29 and 0.34, respectively. Therefore,

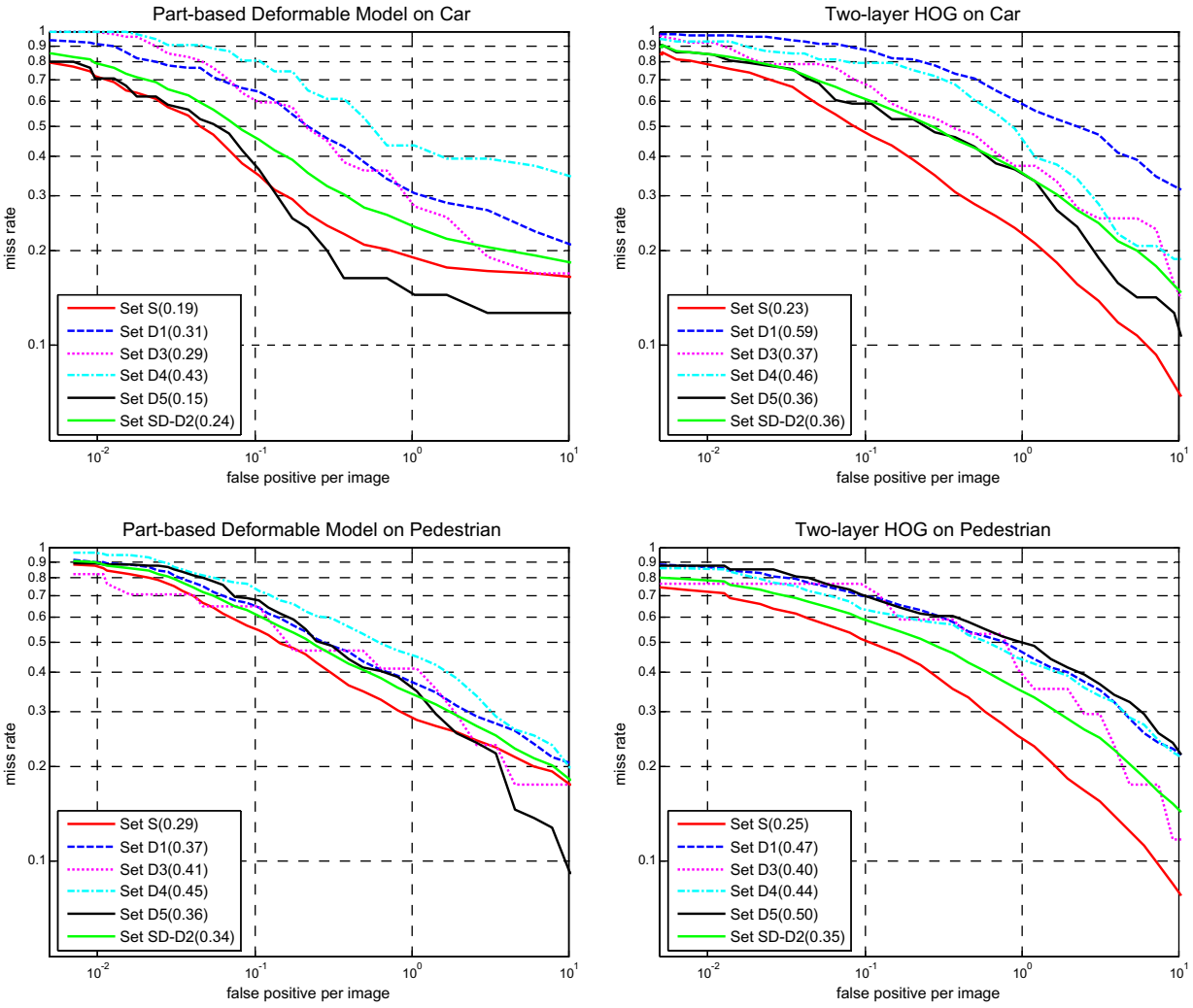


Fig. 10. Object detection results measured by per-image DET curves.

IAIR-CarPed is more challenging than its closest competitor on pedestrian detection.

6.3. Fine-grained and layered object recognition

In this application, an ideal system should be able to localize all the interested objects in a given image, and find the fittest label for each of them. To do this, one has to take into account both the category-level object/non-object classification problem and the fine-grained and layered within-category classification problem. Therefore, the feature representation has to be powerful enough to distinguish the objects from their backgrounds while at the same time differentiate within-category subsets with nonidentical semantic labels. A possible solution is to extend the structured learning framework for fine-grained and layered object classification by treating the background as one more class, and incorporate the data-mining strategy in object detection to grasp representative non-object examples for learning the recognition model. Note that the relationship between objects and their key parts should be properly modeled and the features for representing them should be somehow combined. Since this application is a novel problem which has no off-the-peg solutions, we leave it for further research.

7. Discussion

In this paper we experimentally show that humans tend to have fine-grained within-category interpretations of objects instead of a simple categorization and such interpretations are likely to be layered due to different visual appearances of objects. This property of human vision may be important for in-depth recognition of objects and also helpful for category-level recognition. To mimic it in computer vision, we introduce the IAIR-CarPed dataset which is so far the first fine-grained and layered object recognition dataset with carefully collected human annotations from 20 subjects on two representative categories: car and pedestrian.

Three typical applications have been discussed in this paper with primary experimental results on the first two of them. The first one can be viewed as a simplified version of the third one. Even though, it poses the new problem of differentiating the fine-grained and layered labels which is one of the kernel issues towards solving the ultimate fine-grained and layered object recognition problem. The second application on object detection is a byproduct of the dataset, but it is contributive to researchers working on detecting cars and pedestrians with some new properties that other datasets do not have. Though we have not presented experimental results on the third, hints on designing algorithms for solving it have been given. Exploring effective and efficient algorithms for such an application will be our future work.

Due to the labels we have chosen, the annotation based on human rapid recognition is informative but also expensive. IAIR-CarPed took about 200 person hours for the annotation. For generalizing the idea to many other object categories, we may be able to design new less demanding semantic labels for them and put the annotation tool on the Internet for collecting the labels from the public, e.g. using the service of Amazon Mechanical Turk (AMT) (Deng et al., 2009). Anyway, for both research purpose and real applications, we believe that IAIR-CarPed is a good starting point, and we hope that it will advance the research on fine-grained and layered object recognition.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 90820017, and the National Basic Research Program of China under Grant Nos. 2007CB311005 and 2012CB316400. It was also supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

Agarwal, S., Awan, A., Roth, D., 2004. Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Anal. Machine Intell. 26, 1475–1490.

Aghajanian, J., Warrell, J., Prince, S.J., Li, P., Rohn, J.L., Baum, B., 2009. Patch-based within-object classification. In: Proc. Internat. Conf. on Computer Vision (ICCV), pp. 1125–1132.

Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Lett. 30 (2), 88–97.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. I: 886–893.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Dickinson, S., 2009. Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press, Ch. The Evolution of Object Categorization and the Challenge of Image Abstraction.

Dollar, P., Wojek, C., Schiele, B., Perona, P. 2009. Pedestrian detection: A benchmark. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 304–311.

Enzweiler, M., Gavrila, D., 2009. Monocular pedestrian detection: Survey and experiments. IEEE Trans. Pattern Anal. Machine Intell. 31 (12), 2179–2195.

Ess, A., Leibe, B., van Gool, L., 2007. Depth and appearance for mobile scene analysis. In: Proc. Internat. Conf. on Computer Vision (ICCV), pp. 1–8.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Internat. J. Comput. Vision 88 (2), 303–338.

Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing objects by their attributes. In: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1778–1785.

Fei-Fei, L., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE Trans. Pattern Anal. Machine Intell. 28 (4), 594–611.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Machine Intell. 32 (9), 1627–1645.

Huang, C., Ai, H., Li, Y., Lao, S., 2007. High-performance rotation invariant multiview face detection. IEEE Trans. Pattern Anal. Machine Intell. 29 (4), 671–686.

Joachims, T., Finley, T., Yu, C.N.J., 2009. Cutting-plane training of structural svms. Machine Learn. 77 (1), 27–59.

Lampert, C., Nickisch, H., Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 951–958.

Nene, S., Nayar, S.K., Murase, H., 1996. Columbia object image library (coil-100). Tech. Rep. CUCS-006-96, Columbia University.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. LabelMe: A database and web-based tool for image annotation. Internat. J. Comput. Vision 77 (1–3), 157–173.

Schneiderman, H., Kanade, T., 2000. A statistical model for 3d object detection applied to faces and cars. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T., 2005. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Tech. Rep. AI Memo 2005-036/CBCL Memo 259, Massachusetts Inst. of Technology.

Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proc. National Acad. Sci. (PNAS) 104 (15), 6424–6429.

Wu, B., Nevatia, R., 2009. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. Internat. J. Comput. Vision 82 (2), 185–204.

Wu, Y., Yuan, Z., Liu, Y., Zheng, N., 2009. Discriminative structured outputs prediction model and its efficient online learning algorithm. In: IEEE Internat. Workshop on Emergent Issues in Large Amounts of Visual Data, pp. 2087–2094.

Wu, Y., Liu, Y., Yuan, Z., Zheng, N., 2011. Human confusion costs for object classification. Optical Eng. 50 (2), 1–7.

Yao, B., Yang, X., Zhu, S.-C., 2007. Introduction to a large-scale general purpose ground truth dataset: Methodology, annotation tool and benchmarks. In: Proc. EMMCVPR, pp. 169–183.

Yao, B., Khosla, A., Fei-Fei, L., 2011. Combining randomization and discrimination for fine-grained image categorization. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Yuan, Q., Thangali, A., Ablavsky, V., Sclaroff, S., 2008. Multiplicative kernels: Object detection, segmentation and pose estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.