Human confusion costs for object classification

Yang Wu
Yuanliu Liu
Zejian Yuan
Nanning Zheng
Xi'an Jiaotong University
Institute of Artificial Intelligence and Robotics
28 West Xianning Road
Xi'an, Shaanxi 710049, China
E-mail: wuyang0321@gmail.com

Abstract. Most of the traditional evaluation criteria of object classification are based on the error rate, assuming that the costs of different errors are equal. However, the subjective evaluation of the human vision system on such misclassification errors may be unequal. How do we design proper performance evaluation critera taking into account such inequalities is a kernel issue for mimicking the human vision on object classification. We propose the human confusion costs, which are derived from the statistical human confusions on the training and test data sets, for the model learning and the performance evaluation of the generic cost-sensitive object classification problem, respectively. Unlike the manually designed costs, the proposed ones can better represent the properties of human vision. Experimental results on a new data set with annotations from 20 subjects demonstrate the superiority of the proposed costs against other applicable costs.
[DOI: 10.1117/1.3533729]

Subject terms: object classification; performance evaluation; cost-sensitive classification; human confusion costs.

Paper 100577RR received Jul. 15, 2010; revised manuscript received Dec. 9, 2010; accepted for publication Dec. 10, 2010; published online Feb. 1, 2011.

1 Introduction

Object classification is one of the fundamental research topics in computer vision that has attracted much attention from the community. It is a general problem that may be referred to as different tasks in different scenarios, such as object categorization and within-category object classification (e.g., face recognition and object pose classification) according to the semantic level of its outputs, or binary classification and multiclass classification based on the volume of its output space. In this paper, we are concerned with the general problem itself and especially focusing on its performance evaluation, which is critical for comparing the existing algorithms and promoting the research on it.

To the best of our knowledge, most of the research on object classification attempts to minimize the number of classification errors on the test set, and the evaluation criterion is also based on the error rate. For example, the well-known PASCAL VOC (Visual Object Classes) Challenge[1] evaluates its two main competitions on classification (object presence/absence classification and object detection) based on Precision-Recall curves and more condensedly average precision values, while algorithms performed on the multiclass categorization data sets, such as Caltech-101,[2] Caltech-256,[3] and ImageNet,[4] are quantitatively compared by average recognition rate, receiver operating characteristic (ROC) curves, or the area under ROC curve (AUC). All these criteria are conditioned on the assumption that the costs of different errors are equal.

However, in many cases the classification errors are perceptually unequal to human beings and, in certain real applications, they may even result in quite different costs. For example, the confusion among different kinds of cats is more reasonable than mistaking a cat for a car. Practically, for a door-locker system based on face recognition, misclassifying a stranger as a family member should be much worse than treating a family member as a stranger.[5] As far as we are aware, the generic cost-sensitive object classification beyond the specific face recognition problem[5] is far from being well studied in computer vision.

In the fields of machine learning and data mining, however, cost-sensitive learning (which aims at minimizing total cost rather than total error) has been studied for years,[6] in which many kinds of misclassification costs have been researched.[7] Even though, most such works are on designing learning algorithms conditioned on given costs, but not finding proper cost matrices. Even in the work of real face recognition,[5] the cost matrix is given experientially or selected from several predefined candidates via cross-validation. Such subjective settings can hardly coincide with the visual feelings of humans.

The major contribution of this paper is proposing a novel cost named the "human confusion cost," directly derived from the statistical confusion of human beings, which can generate significantly closer classification results to those of the humans than other applicable costs. Though we can directly measure the similarity of the output confusion matrix and the confusion matrix of humans, the measurement cannot be directly used for model learning. Therefore, we have to somehow map the human confusion matrix to the confusion cost matrix so that cost-sensitive learning methods can be adopted to train a classification model. Details on choosing a proper mapping function is given in Sec 2.

To incorporate a cost matrix in cost-sensitive learning, a popular approach is to do some rescaling according to the misclassification cost matrix.[6] One typical process for such rescaling is to weight the training examples differently before learning, for which Zhou and Liu have found that, for multiclass classification, the weighting strategy is only effective when the consistency of the costs is ensured; otherwise, the problem has to be solved by a combinatorial number of pairwise binary-class classification problems followed by voting.[6] Though the human confusion cost can be used directly in existing cost-sensitive learning approaches, it is

0091-3286/2011/$25.00 © 2011 SPIE

likely that the weighting strategy is unapplicable because the human confusion cost is unconstrained, which may result in a slow learning process. Therefore, we propose to use the maximum-margin–based structured prediction model[8] instead, which has no constraint on the cost matrix and can be optimized efficiently.

The rest of the paper starts with the discussion on the importance of human confusion and the mapping from the confusion matrix to the cost matrix. Then, the structure prediction model is introduced for learning with such costs followed by the experiments and results. Finally, the conclusion and possible future work are given.

2 Human Confusion Costs

2.1 *Confusion Matrix*

The most widely used misclassification costs in cost-sensitive learning belong to the group of class-dependent cost,[6] which tries to model the relationships between any pair of two different classes. Instead of trying to manually design a cost matrix for such relationships, such as what is usually done in machine learning and text classification,[9] we propose to step back and think about what a good cost matrix should be for the problem of object classification.

The ground-truth labels for object classification are provided by humans, which means that the correctness of predicted results is actually verified by we humans. Since humans are the judgers, why not have the degree of misclassification errors also be evaluated by humans? The confusion matrix is a good statistic for representing the classification errors of an algorithm; thus, an intuitive way for performance evaluation is to compare such a confusion matrix with the one of human beings. Using such a criterion, it is expected that the object classification models will be forced to be more like humans than before.

Suppose $\mathbf{M} \in \mathbb{N}^{\mathbf{N}\times\mathbf{N}}$ and $\bar{\mathbf{M}} \in \mathbb{N}^{\mathbf{N}\times\mathbf{N}}$ are the confusion matrices, of humans and the predicted results of the machine, respectively, in which $\mathbb{N}$ is the set of natural numbers while N is the number of semantic classes. We think that a reasonable criterion for measuring the dissimilarity of $\mathbf{M}$ and $\bar{\mathbf{M}}$ is to demand all the correct and incorrect predictions of these two matrices have the same distribution, but to prefer a lower error rate of the predictions while at the same time making the classification mistakes look similar to those of humans. Because minimizing the error rate is a basic demand for all classification algorithms, it is more important to measure how much the distribution of misclassifications is like the confusion distribution in human vision. Let $\mathbf{M^0}$ and $\bar{\mathbf{M}}^{\mathbf{0}}$ be two matrices that have the same elements as $\mathbf{M}$ and $\bar{\mathbf{M}}$, respectively, except that their diagonal elements are all zeros, then we think the dissimilarity of $\mathbf{M}$ and $\bar{\mathbf{M}}$ is just equal to that of $\mathbf{M^0}$ and $\bar{\mathbf{M}}^{\mathbf{0}}$, which can be computed as follows:

$$d_{\mathrm{M}}(\bar{\mathbf{M}}, \mathbf{M}) = d_{\mathrm{M}}(\bar{\mathbf{M}}^{\mathbf{0}}, \mathbf{M^0}) = \sum_{i=1}^{N} \bar{n}_i d_{\mathrm{P}}\left(\frac{1}{\bar{n}_i}\bar{\mathbf{M}}^{\mathbf{0}}_{i\cdot}, \frac{1}{n_i}\mathbf{M^0}_{i\cdot}\right) \Big/ \sum_{i=1}^{N} \bar{n}_i, \quad (1)$$

where

$$\forall i \in \{1, \cdots, N\}, n_i = \sum_j \mathbf{M}^{\mathbf{0}}_{\mathbf{ij}}, \quad \bar{n}_i = \sum_j \bar{\mathbf{M}}^{\mathbf{0}}_{\mathbf{ij}}$$

In such a criterion, $\bar{\mathbf{M}}_{i\cdot}/\bar{n}_i$ is a vector representing the probabilities of confusing examples of the ith class to any other classes by the machine, while $\mathbf{M}_{i\cdot}/n_i$ denotes those probabilities of human vision. d_{P} is a distance measurement for comparing these two probability distributions, and the actual number of mistakes $\bar{n}_i$ weights the distance for the integrated measure $d_{\mathrm{M}}\left(\bar{\mathbf{M}}, \mathbf{M}\right)$. In general, d_{P} can be any probability distance measurement. In our experiments to be presented later, we chose the widely used χ^2 distance for the final performance evaluation, though other distances (such as earth mover's distance, Kullback–Leibler divergence, the histogram intersection, etc.) are also applicable and may generate similar results.

2.2 *Cost Matrix*

Though the dissimilarity measurement expressed by Eq. (1) is an objective and fair evaluation of the object classification algorithm compared to the human vision system, it seems to be infeasible to use it as a direct minimization criterion for optimizing a learning model, because the measurement depends on the actual classification results, which cannot be exactly predicted in advance. As mentioned earlier, in cost-sensitive learning the relationships between different classes are measured by the costs, which can be encoded in the learning model. Therefore, a feasible solution for learning the way humans recognize objects is to properly translate the human confusion matrix into a cost matrix.

It is expected that a higher confusion probability should correspond to a lower misclassification cost. Therefore, any function that increases as the confusion probability decreases is applicable because it can somehow represent the property of human vision on object recognition. However, it is difficult to say which function is the best when evaluated by the objective criterion of Eq. (1) because it is unclear how the minimization of the total misclassification cost and that of the dissimilarity of $\mathbf{M}$ and $\bar{\mathbf{M}}$ can be analytically connected.

Among all possible mappings, we choose the exponential function family with a free parameter, which is defined as follows:

$$\Delta_{\mathrm{Conf}}(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} e^{-\rho \cdot \mathbf{M}_{c(\mathbf{y_i}),c(\mathbf{y_j})}}, & \mathbf{y}_i \neq \mathbf{y}_j, \\ 0, & \mathbf{y}_i = \mathbf{y}_j, \end{cases} \quad (2)$$

where $c(\mathbf{y}_i)$ is the class index of the output label $\mathbf{y}_i$ in the output space, and ρ is a parameter controlling the variation of the cost. Because such a cost function is computed from the human confusions, it is named the "confusion cost". To make the control of ρ visualized and operable, we use another parameter η that directly measures the min-max ratio of all possible confusion costs,

$$\eta = \frac{\min\limits_{\mathbf{y}_i,\mathbf{y}_j,\mathbf{y}_i\neq\mathbf{y}_j} \Delta_{\mathrm{Conf}}(\mathbf{y}_i, \mathbf{y}_j)}{\max\limits_{\mathbf{y}_i,\mathbf{y}_j,\mathbf{y}_i\neq\mathbf{y}_j} \Delta_{\mathrm{Conf}}(\mathbf{y}_i, \mathbf{y}_j)}. \quad (3)$$

Note that correct predictions have zero costs and are not participating for computing η. Therefore, we can just adjust η to get the best parameter for the confusion cost function. Once η is fixed, ρ can be derived from

$$\rho = \frac{\ln \eta}{\min\limits_{\mathbf{y}_i,\mathbf{y}_j,\mathbf{y}_i\neq\mathbf{y}_j} \left\{\mathbf{M}_{c(\mathbf{y_i}),c(\mathbf{y_j})}\right\} - \max\limits_{\mathbf{y}_i,\mathbf{y}_j,\mathbf{y}_i\neq\mathbf{y}_j} \left\{\mathbf{M}_{c(\mathbf{y_i}),c(\mathbf{y_j})}\right\}}. \quad (4)$$

3 Cost-Sensitive Classification via Structured Prediction

In this paper, we propose to use the maximum-margin based structured prediction model[8] for cost-sensitive learning because it has no assumptions on the training cost matrix and can be easily optimized. Concretely, given a set of training samples $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where $\mathbf{x}_i \in \mathbf{X}, i \in \{1, \ldots, n\}$ are input patterns and $\mathbf{y}_i \in \mathbf{Y}, i \in \{1, \ldots, n\}$ are corresponding output vectors. Maximum-margin–based discriminative structured learning is trying to maximize the margin between the compatibility $E(\mathbf{x}_i, \mathbf{y}_i) = \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) \rangle$ of the correct pattern-output pair and the compatibility $E(\mathbf{x}_i, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}) \rangle, \mathbf{y} \neq \mathbf{y}_i$ of any other incorrect pattern-output pairs, where $\mathbf{w}$ is the weight vector for the joint input-output feature vector $\Phi(\mathbf{x}_i, \mathbf{y})$ and $\langle \cdot, \cdot \rangle$ is the inner product. By introducing the cost matrix $\Delta(\mathbf{y}_i, \mathbf{y}_j)$, the learning problem can be formulated as

$$\min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_i \xi_i \right\}$$
$$s.t. \begin{cases} \langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \ , \ \forall i, \mathbf{y} \neq \mathbf{y}_i, \\ \xi_i \geq 0 \ , \ \forall i \end{cases} \tag{5}$$

where $\Delta\Phi(\mathbf{x}_i, \mathbf{y})$ stands for $\Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})$ and ξ_i is the slack that stands for the tolerance of the training sample $\mathbf{x}_i$ across the margin border in the maximum-margin–based model. In the constraints, the cost matrix directly rescales the margin between the correct class and any other incorrect ones. When the model parameter vector $\mathbf{w}$ (i.e., the feature weight vector) is trained, the prediction of an arbitrary input pattern $\mathbf{x} \in \mathbf{X}$ is just as simple as;

$$f(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathbf{Y}} \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle . \tag{6}$$

For a deep understanding of the above formulation, the reader is referred to the rich literature on maximum-margin–based structured prediction models.[8,11,12] The proposed model has many advantages compared to the others. First, it explicitly utilizes the cost matrix for model learning. Second, it does not need to weight the examples or change their labels;[10] thus no consistency examination is needed. Third, it is suitable for both binary classification and multiclass classification. Fourth, it can be solved using off-the-shelf tools such as SVMstruct,[11] or SOnline,[12] which is an approximate yet more efficient learning algorithm.

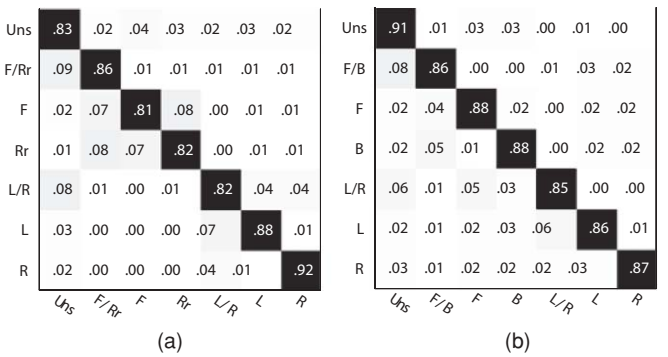

Fig. 1 Confusion matrices of humans on object classification in IAIR-CarPed data set through rapid recognition: (a) car and (b) pedestrian. These two matrices are computed from the whole data set just for demonstration. In practice, they should be computed on the training set and the test set for model learning and performance evaluation, respectively, which should look quite similar to the ones presented, so they are omitted here.

4 Experiments and Results

As far as we are aware, to date there is no publicly available object classification data set equipped with human confusion information between different semantics. Usually, each object in the data sets is annotated by just one single person and then the label is served as the ground truth. ImageNet[4] is an exception that has more than one person label a single image; however, the labels from multiple users are only used for cleaning the images for the predefined categories but not recorded for analyzing human confusion. Therefore, we introduce a new data set "IAIR-CarPed," with each object annotated by as many as 20 subjects, and experimentally show that the human confusion costs computed from these labels are better than traditional zero-one costs and other applicable hand-designed costs for both performance evaluation and model learning.

4.1 *The IAIR-CarPed Data Set*

The IAIR-CarPed data set focuses on two representative categories: "car" and "pedestrian." It involves 3132 images collected from pictures taken under various conditions and 8567 objects carefully annotated by 20 subjects. Unlike other data sets, IAIR-CarPed provides nonflat within-category semantics on the orientation of objects. The annotation of such objects is done by performing a psychophysical experiment of rapid recognition as presented in Ref. 13, which was proved to be effective for representing the confusion of humans on such semantics in their daily lives where rapid recognition has been widely adopted.

The data set provides a final semantic label (i.e., the ground truth) for each object instance by integrating the annotations from all 20 subjects. Therefore, by treating the votes to other semantic labels as recognition mistakes, we can compute the human confusion matrices for these two categories as shown in Fig. 1.

The data set has two lists of images within it for training and testing, respectively, so that different algorithms can be compared on them. These two sets include 1566 images, and the numbers of objects within them are roughly the same for both car and pedestrian. The experiments presented below are also trained and evaluated on the two predefined subsets.

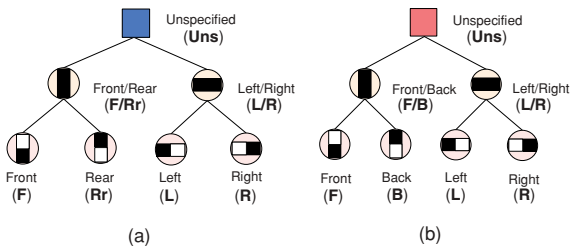

Fig. 2 Tree-structured semantics for the two categories: (a) car and (b) pedestrian. Going down along the trees, the semantics get more and more certain and specific. The shapes of the nodes are designed to illustrate the semantic meanings associated to them. Proper abbreviations are provided beside the nodes for easy references.

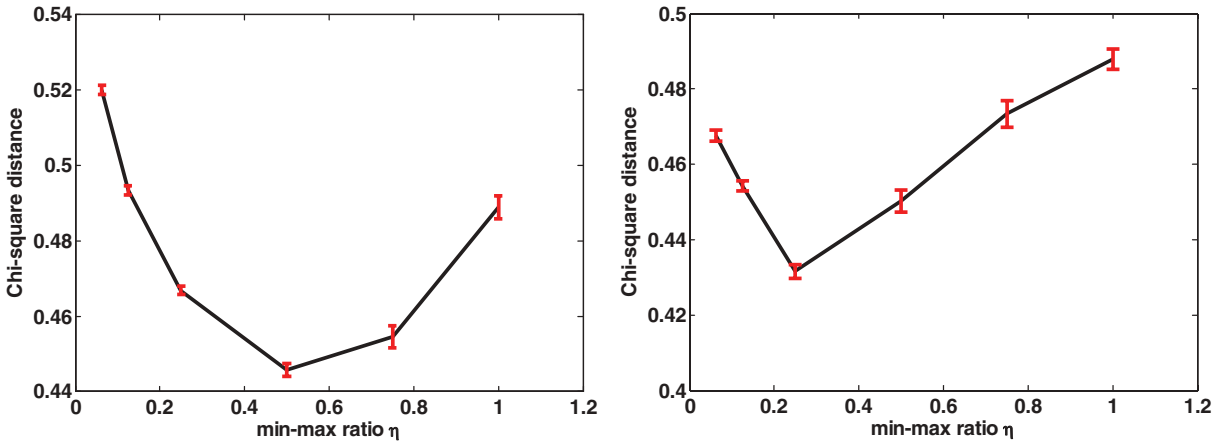

Fig. 3 Performance influence of the free parameter in the human confusion cost Δ_{Conf} (the min-max ratio $\eta = 1$ actually denotes the zero-one cost $\Delta_{0/1}$). The two subgraphs are for car and pedestrian, respectively. The bounded vertical lines stand for the standard errors of 10 trials.

4.2 *Feature Representation*

For feature representation, we choose the well-known histogram of gradients (HOG) features,[14] which have been proved to be effective for describing cars and pedestrians, especially for the task of detection. Unlike the detection problem, the within-category classification problem we are trying to solve demands the exploration of shape subtleties between pairs of the predefined semantics. Therefore, we double the density of the cells to be 12×16 for cars and 24×8 for pedestrians. Margins around the objects are excluded in our feature representation as we experimentally find that they are unhelpful for the within-category classification problem. We used PCA-HOG[15] instead of the original HOG for both efficiency and compactness.

4.3 *Classification Costs*

To verify the superiority of the human confusion cost (Δ_{Conf}), we compared it to the flat zero-one cost (denoted as $\Delta_{0/1}$) and other nonflat costs used in other cost-sensitive learning works. $\Delta_{0/1}$ treats different classes equally so that all possible mistakes have the same cost of 1, and the correct prediction has zero cost. $\Delta_{0/1}$ can be viewed as a special case of Δ_{Conf} when η approaches 1. As mentioned earlier, in cost-sensitive learning, the cost matrix is usually hand-designed according to certain priors. About the orientational semantics annotated in the IAIR-CarPed data set, it is most natural to organize them in tree structures as shown in Fig. 2. Therefore, applicable costs are those taking into account of the structure of the trees.

An extensive study of different tree-structure costs is presented in the work of Rousu et al. on hierarchical multilabel text classification,[9] in which four typical costs have been compared. These four costs belong to the following two different types:

4.3.1 *Type I*

Type I has only one cost originally named as "symmetric difference loss." It treats each node on the tree as a binary element of the output vector, and once a node is selected as the output semantic, both its corresponding element and those representing its ancestors are labeled as 1. Then, the cost of confusing two semantics is measured by the number of elements with different values (similar to the Hamming distance metric). To respect its original name, we call it "symmetric difference cost" and denote it as Δ_{Symm}.

4.3.2 *Type II*

Type II only penalizes the first mistake along a path from the root to the two different semantics. By weighting the nodes differently, three different costs have been proposed:

1. the cost with uniform weights (all equal 1), which is denoted by Δ_{Unif};
2. the cost with divided weights by the number of siblings of a referring node, which is denoted by Δ_{Sibl}; more concretely, the weight of the root is set as 1 and the weight for any other node is computed from dividing its parent's weight by the number of its siblings (including itself);
3. the cost with divided weights by the proportion of the number of nodes within the subtree rooted at this node compared to all the nodes in the subtree rooted at its parent node, and such a cost is denoted by Δ_{Subt}.

4.4 *Experimental Results*

All the costs mentioned above can be used to learn a maximum-margin–based structured prediction model

Table 1 Classification performance comparison on the IAIR-CarPed data set using different cost matrices for training. The results are the objective χ^2 distances between the actual confusion matrices and the human confusion matrix on the test set. The parameter ρ in Δ_{Conf} is set by forcing η to be 0.5 and 0.25 for car and pedestrian, respectively.

Cost	$\Delta_{0/1}$	Δ_{Symm}	Δ_{Unif}	Δ_{Sibl}	Δ_{Subt}	Δ_{Conf}
Car	0.4920	0.4776	0.5926	0.5139	0.5127	**0.4457**
Pedestrian	0.4808	0.5399	0.6361	0.5892	0.5764	**0.4316**

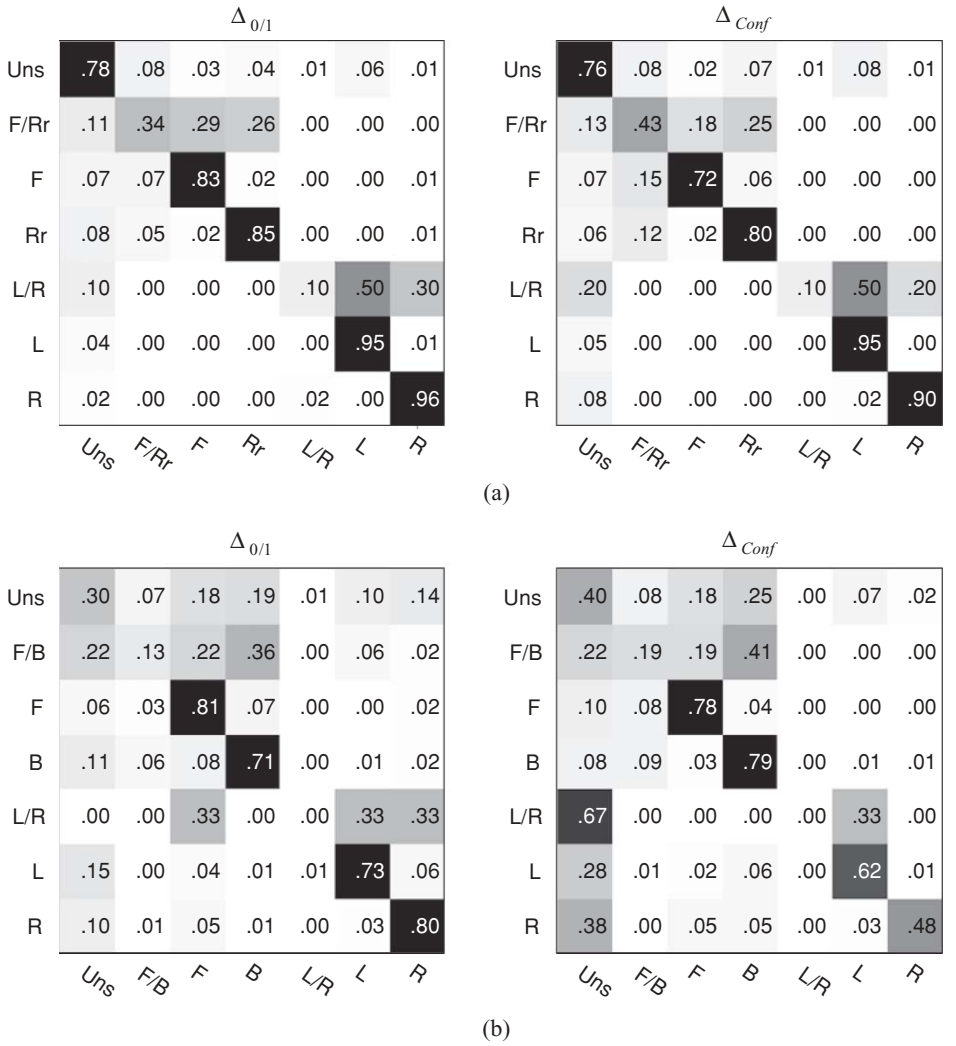

Fig. 4 Confusion matrices on the test set using the models trained with the two different cost matrices ($\Delta_{0/1}$ and Δ_{Conf}). Comparing to the flat multiclass classification cost ($\Delta_{0/1}$), learning with the relatively more conservative confusion cost (Δ_{Conf}) can lead to more conservative predictions. (a, b) graphs for car and pedestrian, respectively.

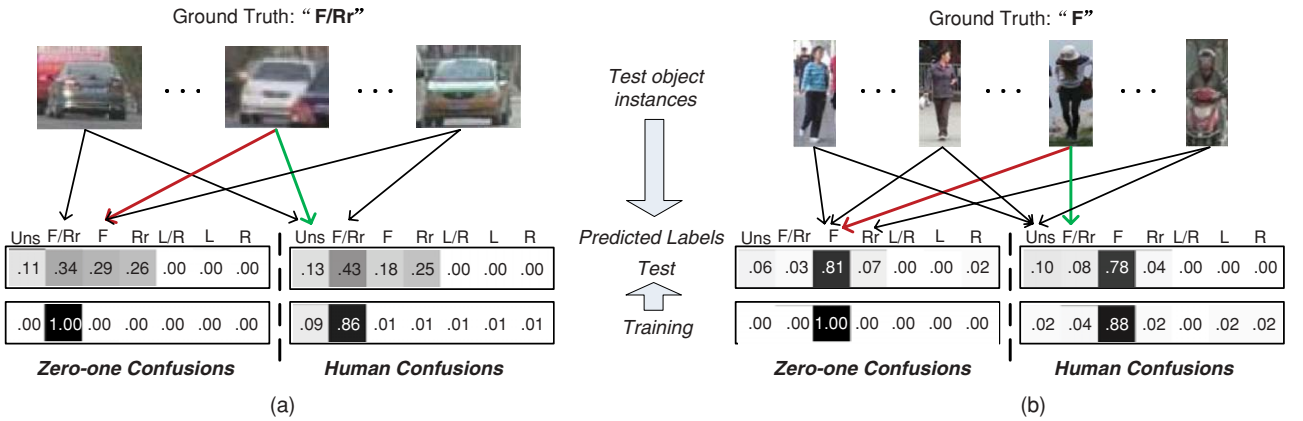

Fig. 5 Some typical object classification results using $\Delta_{0/1}$ and Δ_{Conf} for training and testing. (a, b) graphs for car and pedestrian, respectively, in which a group of object instances belonging to a particular class are presented on the top, and the corresponding training and test confusions between this class and the others are listed at the bottom. The figure shows how the objects are classified into two different classes when trained with $\Delta_{0/1}$ and Δ_{Conf}, respectively. Clearly, the predicted results using Δ_{Conf} are more conservative than those using $\Delta_{0/1}$, which is closer to the way humans make mistakes.

Table 2 Classification results on the IAIR-CarPed data set using different cost matrices for training and measured by different total costs on the test set. The rows are different costs for training, while the costs in the columns are for performance evaluation. The lowest total test cost in each column is marked with bold.

	Car						Pedestrian					
Cost	$\Delta_{0/1}$	Δ_{Symm}	Δ_{Unif}	Δ_{Sibl}	Δ_{Subt}	Δ_{Conf}	$\Delta_{0/1}$	Δ_{Symm}	Δ_{Unif}	Δ_{Sibl}	Δ_{Subt}	Δ_{Conf}
$\Delta_{0/1}$	**0.2251**	0.1543	0.1725	0.1467	0.1363	0.1870	0.3628	0.3292	0.3088	0.3006	0.2973	0.2934
Δ_{Symm}	0.2423	0.1481	0.1758	0.1493	0.1385	0.2064	0.3970	0.3162	0.3110	0.3019	0.2983	0.3370
Δ_{Unif}	0.2236	**0.1422**	**0.1576**	**0.1393**	**0.1319**	0.1936	0.3810	0.3207	**0.2931**	0.2977	0.2998	0.3180
Δ_{Sibl}	0.2318	0.1512	0.1699	0.1450	0.1350	0.1961	0.3704	**0.3132**	0.2948	**0.2898**	0.2879	0.3077
Δ_{Subt}	0.2360	0.1561	0.1761	0.1480	0.1366	0.1955	0.3717	0.3159	0.2983	0.2904	**0.2872**	0.3015
Δ_{Conf}	0.2302	0.1624	0.1805	0.1507	0.1387	**0.1844**	**0.3496**	0.3365	0.3086	0.2997	0.2961	**0.2571**

(Sec. 3) for cost-sensitive classification, and an objective performance evaluation criterion is the χ^2 distance between the confusion matrix of the predicted results and the human confusion matrix on the test set. All the models are learned using the SOnline algorithm[12] for its efficiency. To reduce the influence of the probabilistic sampling within the optimization, all experiments were done 10 times and the average results are adopted for the final comparisons.

Because our proposed human confusion cost Δ_{Conf} has a free parameter in it, we first discuss how it influences the final performance. We tried five different values of η (0.0625, 0.125, 0.25, 0.5, and 0.75) and got the results as shown in Fig. 3. Note that the zero-one cost $\Delta_{0/1}$ stood by $\eta = 1$ is also evolved in Fig. 3. Clearly, for a good performance, η should be neither too small (overexaggerating the differences between the classes) nor too large (downplaying the differences). As the results show, the optimal η for car should be around 0.5, whereas the one for pedestrian is close to 0.25. $\Delta_{0/1}$ (i.e., $\eta \to 1$) is clearly not as good as the well-tuned human confusion cost. The standard error of the 10 trials for each parameter value is also shown in the Fig. 3, which indicates that the probabilistic randomness is rather small and it does not influence the relative ranking of the results. Therefore, we only use the average results for the following comparisons.

The overall classification performances of the six different costs on the IAIR-CarPed data set are listed in Table 1, where Δ_{Conf} uses the best parameters we have tried for car and pedestrian respectively. Training with the human confusion cost can generate classification results closer to those of the humans than using any other costs.

In cost-sensitive learning, the total test costs are widely adopted for performance evaluation. The classification biases (such as the human confusion properties) can somehow be represented by the different costs for different misclassifications. Therefore, we also present the classification results using the six costs for training and evaluated by each of these six costs, as shown in Table 2. It seems to be plausible that training with a particular cost function should result in a lower total test cost compared to the others when measured by the same cost function. However, the experimental results show that this is not necessarily the case. For example, the well-tuned human confusion cost Δ_{Conf} generates not only the lowest confusion cost but also the best zero-one cost on pedestrians.

Figure 4 presents the detailed classification confusions on the test set using the models trained with $\Delta_{0/1}$ and Δ_{Conf} respectively for comparison. The comparison to the results trained with other costs is similar, so it is omitted here. Because the human confusion cost matrix biases on conservative prediction, the model trained with Δ_{Conf} generates fewer aggressive mistakes but more conservative mistakes than the one trained using $\Delta_{0/1}$. Figure 5 gives some concrete examples of the classification results. It can be seen that for some orientation ambiguous cases, such as "F/Rr," the model trained with $\Delta_{0/1}$ tends to output more specific labels, whereas the one learned with Δ_{Conf} does not. For those orientation-specific cases like "F", when their visual appearances are not very distinguishable, the model learned with Δ_{Conf} may bias on ambiguous labels, whereas the one guided by $\Delta_{0/1}$ prefers specific predictions which might be risky. Therefore, from the humans' point of view, Δ_{Conf} is better for performance evaluation and model learning.

5 Conclusions and Future work

This paper introduces the human confusion costs, which are derived from human confusion statistics on the training data, for cost-sensitive object classification. To the best of our knowledge, this is the first time the performance evaluation and model learning of object classification are designed according to the real human confusion, but not predefined subjective values. To learn a classification model with these unconstrained costs, we propose to use maximum-margin–based structured prediction model that is more straightforward and less demanding than traditional cost-sensitive learning methods. Experimental results on the newly built IAIR-CarPed data set show that the proposed human confusion costs are more suitable than other costs for both performance evaluation and model training.

Currently, the different types of human confusion in the data set are carefully collected from subjects via a strict psychophysical experiment, which is good for doing research, but too expensive for scaling to large amounts of object classes. In contrast, there are many noisy image tags,

annotations, and descriptions on the Internet that can be got for free; therefore, the research on using them for constructing human confusion costs for large-scale cost-sensitive object classification is an interesting and promising future work.

Acknowledgments

This work was supported by the National Basic Research Program of China under Grant No. 2007CB311005 and the National Natural Science Foundation of China under Grant No. 90820017.

References

1. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.* **88**, 303–338 (2010).
2. L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 594–611 (2006).
3. G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category data set," CalTech Tech. Report No. 7694 (2007).
4. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1**, 248–255 (2009).
5. Y. Zhang and Z. Zhou, "Cost-sensitive face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1758–1769 (2010).
6. Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," in *Proc. of 21st nat. conf. on Artificial Intelligence*, pp. 567–572 (2006).
7. P. Turney, "Types of cost in inductive concept learning," in *Proc. of ICML'2000 Workshop on Cost-Sensitive Learning*, pp. 15–21 (2000).
8. G. H. BakIr, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. Vishwanathan, *Predicting Structured Data*, MIT Press, Cambridge, MA, p. 360 (2007).
9. J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "Kernel-based learning of hierarchical multilabel classification models," *J. Mach. Learning Res.*, **7**, 1601–1626 (2006).
10. P. Domingos, "Metacost: a general method for making classifiers cost-sensitive," in *Proc. of Fifth Int. Conf. on Knowledge Discovery and Data Mining*, pp. 155–164, ACM, New York (1999).
11. T. Joachims, T. Finley, and C. N. J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learning* **77**, 27–59 (2009).
12. Y. Wu, Z. Yuan, Y. Liu, and N. Zheng, "Discriminative structured outputs prediction model and its efficient online learning algorithm," in *Proc. IEEE International Workshop on Emergent Issues in Large Amounts of Visual Data* **1**, 2087–2094 (2009).
13. T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proc. Nat. Acad. Sci. USA* **104**, 6424–6429 (2007).
14. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition* **1**, 886–893 (2005).
15. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645 (2009).

Yang Wu received his BEng in information engineering from Xi'an Jiaotong University, China, in 2004, where he then began pursuing his PhD in pattern recognition and intelligent system at Institute of Artificial Intelligence and Robotics, China. Sponsored by the China Scholarship Council in 2007, he has been a visiting student of the University of Pennsylvania for one year. His research focuses on object and scene recognition with interests in machine learning and vision science.

Yuanliu Liu received his BEng in control theory and engineering at Xi'an Jiaotong University, China, in 2008, where he has since been pursuing his PhD at the institute of Artificial Intelligence and Robotics. His research mainly focuses on object recognition, especially pedestrian detection.

Zejian Yuan received his MS in electronic engineering in 1999 and PhD in 2003 in pattern recognition and intelligent system, from Xi'an Jiaotong University, China. He was a visiting scholar in the Advanced Robotics Lab of Chinese University of Hong Kong, from May 2008 to December 2009. He is currently an associate professor at the Department of Automatic Engineering, Xi'an Jiaotong University, and a member of Chinese Association of Robotics. He is the author of more than 30 technical publications. His research interests include image processing, pattern recognition, as well as machine learning methods in computer vision.

Nanning Zheng graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975, received his MS in information and control engineering there in 1981. He received his PhD in electrical engineering from Keio University, Yokohama, Japan, in 1985. He joined Xi'an Jiaotong University in 1975 and is currently a professor and the director of the Institute of Artificial Intelligence and Robotics. His research interests include computer vision, pattern recognition, machine vision and image processing, neural networks, and hardware implementation of intelligent systems. He became a member of the Chinese Academy of Engineering in 1999. He is also a member of the Board of Governors of the IEEE ITS Society and the Chinese Representative on the Governing Board of the International Association for Pattern Recognition. He is also a fellow of IEEE.