Pattern Recognition ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Training inter-related classifiers for automatic image classification and annotation

Peixiang Dong [a], Kuizhi Mei [a,*], Nanning Zheng [a], Hao Lei [a], Jianping Fan [b]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, PR China
[b] School of Information Science and Technology, Northwest University, Xi'an 710069, PR China

ARTICLE INFO

Article history:
Received 18 June 2012
Received in revised form
26 October 2012
Accepted 30 October 2012

Keywords:
Inter-related classifier training
Large-scale image classification
Structural learning
Visual concept network

ABSTRACT

A structural learning algorithm is developed in this paper to achieve more effective training of large numbers of inter-related classifiers for supporting large-scale image classification and annotation. A visual concept network is constructed for characterizing the inter-concept visual correlations intuitively and determining the inter-related learning tasks automatically in the visual feature space rather than in the label space. By partitioning large numbers of object classes and image concepts into a set of groups according to their inter-concept visual correlations, the object classes and image concepts in the same group will share similar visual properties and their classifiers are strongly inter-related while the object classes and image concepts in different groups will contain various visual properties and their classifiers can be trained independently. By leveraging the inter-concept visual correlations for inter-related classifier training, our structural learning algorithm can train the inter-related classifiers jointly rather than independently, which can enhance their discrimination power significantly. Our experiments have also provided very positive results on large-scale image classification and annotation.

1. Introduction

As digital cameras become more affordable and widespread, digital images are growing exponentially on the Internet. In the last three decades, many content-based image retrieval (CBIR) systems have been developed [1–3], in which low-level visual features are usually extracted for image indexing and retrieval. Unfortunately, most naive users may not be familiar with the low-level visual features and hence keywords would be more suitable for them to specify their queries intuitively. To support keyword-based image retrieval, it is very attractive to develop new algorithms for supporting automatic image classification and annotation.

To achieve large-scale image classification (i.e., categorizing large-scale images into large numbers of object classes and image concepts), it is very important to train a large number of classifiers for mapping the low-level visual features (computer interpretations of visual content of images) onto the high-level image concepts (human interpretations of visual content of images). It is well accepted that the performance of image classifiers largely depends on two inter-related critical issues: (a) *quality of visual features*; (b) *tools for classifier design and training*. The most popular visual features include color [4], texture [5], shape [6–8] and salient points [9]. Each type of these visual features is used to describe one particular type of visual properties of the images. In this paper, we focus on dealing with the second issue for automatic image classification and annotation, i.e., classifier design and training, and SIFT (scale invariant feature transform [9]) features are used for image representation in our experiments.

It is also worth noting that there are strong visual correlations among the object classes and image concepts, e.g., the relevant images for some object classes and image concepts may share some common or similar visual properties. Thus it is not a good idea to isolate such inter-related object classes and image concepts and train their inter-related classifiers independently. Training the classifiers for such inter-related object classes and image concepts independently may result in low accuracy rates for image classification. As a result, new algorithms, which can leverage the inter-concept visual correlations for inter-related classifier training, are strongly expected. For the tasks of inter-related classifier training, the following two issues are equally important: *classifier design* and *inter-class correlation measurement*.

There are two well-known approaches for multi-class classifier design: (a) *traditional approach* (i.e., learning the classifiers for all the object classes and image concepts independently) [10–12] and (b) *inter-related approach* (i.e., the inter-concept relationships among the object classes and image concepts are integrated for inter-related classifier training) [13–17]. Two representative

* Corresponding author. Fax: +86 29 82668672.
E-mail address: meikuizhi@mail.xjtu.edu.cn (K. Mei).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.10.029

solutions for the traditional approach are the *one-against-all* (or *one-against-rest*) method and *pairwise* (or *one-against-one*) method. When the object classes and image concepts are inter-related, the tasks for training their inter-related classifiers are strongly dependent and their inter-related classifiers should be trained jointly rather than independently. *Multi-task learning* is accepted as one potential solution for leveraging such inter-concept correlations to train the inter-related classifiers jointly [16–19].

For a large number of object classes and image concepts, how to measure their inter-concept correlations is another critical issue. Most existing classifier training approaches aim at exploiting the inter-concept semantic similarity contexts [20–23]. The most popular method is to extract the inter-concept semantic similarity contexts from some existing ontologies such as WordNet [24,25] (WordNet is a semantic lexical database which provides multiple types of relations among English words). Unfortunately, large amounts of text terms on the WordNet may not directly relate to the text terms for interpreting the object classes and image concepts on the Internet. The Google similarity distance [26] is another method to measure the inter-concept similarity relations among the object classes and image concepts, while it prefers the inter-concept contextual relations rather than the inter-concept semantic relations. In short, both the Google similarity distance and WordNet can characterize the inter-concept relatedness effectively at the label space. However, both classifier training and automatic image classification indeed happen in the visual feature space rather than in the label space, thus all these methods for inter-concept similarity measurement cannot be directly extended for determining the inter-related learning tasks accurately.

When a large number of object classes and image concepts come into view, most existing algorithms for classifier training may further suffer from some other challenging problems. Firstly, one challenging issue is how to excavate the inter-concept relatedness automatically. Secondly, the classification error may be transmitted among the classifiers for the inter-related object classes and image concepts when the inter-concept correlations are leveraged for classifier training. To support large-scale image classification and annotation, it is very attractive to design a low-computation-cost algorithm for training a large number of inter-related classifiers with high discrimination power but less error transmission.

In this paper, a structural learning algorithm is developed for training a large number of inter-related classifiers jointly. The followings highlight some main aspects of our proposed algorithm for large-scale image classification and annotation.

(a) Our structural learning algorithm can determine the inter-related learning tasks directly in the visual feature space rather than in the label space. It is worth noting that the visual feature space is the common space for classifier training and image classification.
(b) Our structural learning algorithm can significantly enhance the discrimination power of the inter-related classifiers while restraining their error transmission effectively.
(c) Our structural learning algorithm can lessen the computational cost dramatically for large-scale classifier training.

The rest of this paper is organized as follows. Section 2 briefly reviews some most related work. In Section 3, we introduce our algorithm for visual concept network construction. Our structural learning algorithm is presented in Section 4, where the visual concept network is used to determine the inter-related learning tasks and leverage the inter-concept visual correlations for inter-related classifier training. Section 5 describes our work on algorithm evaluation and we conclude this paper in Section 6.

2. Related work

In this section, we review some most relevant work on multi-class classifier training briefly, which can be summarized as two categories: *traditional approach* and *inter-related approach*.

Comprehensive studies of the traditional approaches can be found in many literatures, e.g., [27,28]. The one-against-all method is probably the earliest approach for multi-class classifier training [29] even it still deserves to be widespread concerned. The one-against-all method is also known as one-against-others or one-against-rest where one class is separated from the remaining classes. Many current state of arts methods adopt this binary strategy to deal with multi-class image classification tasks [11,12,30]. The pairwise method is also known as one-against-one method, which learns a set of pairwise binary classifiers to distinguish each pair of classes. The classification decision is made by aggregating the outputs of all the pairwise classifiers. A typical work of this method can be found in literature [10].

Many previous work have demonstrated that the pairwise method has better performance than the one-against-all method [27,29]. The unbalance of training samples may be the most serious problem that significantly decrease the performance of the one-against-all method. For the one-against-all approach, a correct prediction would require the true classifier to be more confident than other classifiers. It will be difficult to achieve such requirement especially when we deal with a large number of classes because the chance of false alarms from many other classifiers may dramatically increase. For a k-class classification problem, the pairwise method needs to construct $k(k-1)/2$ pairwise classifiers, thus the computation cost will grow quadratically as the number of classes increases.

The traditional approach for classifier training mainly focus on small-scale problems, typically from several classes to a few tens of classes. The inter-class relatedness would be more crucial for large-scale image classification, where we may usually deal with hundreds of object classes and image concepts simultaneously. Some pioneer work have been done to leverage the inter-concept similarity contexts for inter-related classifier training. There are two well-known approaches for inter-related classifier training: multi-task learning and hierarchical learning. We focus on giving a brief overview of those work that are most relevant to our proposed algorithm.

A potential solution for inter-related classifier training is multi-task learning [16–19]. Torralba et al. [18] have proposed a JointBoost algorithm to leverage the inter-task correlations for improving object detection, where the inter-task correlations are explicitly characterized by using pairwise object combinations. Fan et al. [17,19] have also integrated multi-task learning and concept ontology to leverage the hierarchical inter-concept similarity contexts for training multiple inter-related classifiers jointly.

To support hierarchical image classification, Barnard et al. [31] and Vasconcelos et al. [32] have incorporated hierarchical mixture models and concept ontology to leverage the hierarchical inter-concept semantic similarity contexts for training multiple inter-related classifiers jointly. Li et al. [33] have presented a linguistic structure for image database indexing, classification, and annotation. Fei-Fei et al. [13] have also incorporated prior knowledge of object parts and their locations to improve hierarchical classification. Sudderth et al. [14] have proposed a statistical approach to exploit the inter-object contexts to improve object detection. Luo et al. [15] have integrated Bayesian

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

network for automatic image classification. Marszałek et al. [34] have proposed a top-down approach for constructing category hierarchies which postpone the decisions when the uncertainty appears. The major problem for hierarchical learning approaches is the inter-concept error transmission [35], e.g., the classification errors will be propagated among the classifiers for the inter-related object classes and image concepts at different levels [17].

In order to construct the hierarchies for measuring the inter-task correlations, many pioneer work have been proposed. Marszałek et al. [22] and Fei-Fei et al. [36] have used WordNet to find the semantic relationships between the labels and combined discriminative classifiers through the semantic hierarchies. Wang et al. [23] have proposed a Conditional Random Field framework to incorporate the contextual relations between the words for multi-label image annotation, where the normalized Google distance (NGD) [26] is utilized as the inter-word contextual potential. Bengio et al. [37] have learned a tree structure through a confusion matrix which is obtained by the one-against-all classifiers for all the classes. Tousch et al. [38] have proposed a comprehensive study of the semantic hierarchies that are used in the field of image annotation. However, all these similarity measurements are not able to exactly reflect the visual correlations among the object classes and image concepts, so it is not quite reasonable to apply such semantic similarity contexts for the tasks of multi-class image classification. For the purpose of large-scale image classification, a visual concept network is first proposed in [39] and the inter-concept visual correlations are leveraged to jointly train the classifiers for the inter-related object classes and image concepts in the same group (i.e., intra-group separation). However, the literature [39] has not provided good solutions for separating the object classes and image concepts in different groups (i.e., inter-group separation).

Different from all these previous work, two inter-related issues are equally taken into account in this paper: (a) the construction of the structure (i.e., visual concept network construction for advising how to cluster the object classes and image concepts into a set of groups); and (b) the learning from the structure (i.e., machine learning algorithm for advising how to design group-based classifier training).

A flowchart in Fig. 1 illustrates our structural learning algorithm for inter-related classifier training. From this flowchart one can observe that our structural learning algorithm consists of three key components.

(1) Computing the inter-concept visual similarity contexts among the object classes and image concepts, where a visual concept network is constructed for characterizing such inter-concept visual similarity contexts intuitively.
(2) Partitioning the object classes and image concepts into a set of small groups, where the object classes and image concepts in the same group will share some common or similar visual properties while the object classes and image concepts in different groups will contain various visual properties.
(3) Designing a group-based structural learning algorithm for training multi-class inter-related classifiers with less error transmission but high discrimination power.

3. Task relatedness quantification

In this section, a visual concept network is constructed to quantify the inter-task relatedness directly in the visual feature space. The visual concept network could achieve more effective organization of a large number of object classes and image concepts according to their inter-concept visual similarity contexts in the visual feature space. In addition, the visual concept network can provide a good environment to identify the inter-related learning tasks for training multiple inter-related classifiers jointly.

We take ImageNet [40] as our image set because a large number of object classes and image concepts and their relevant images are available. ImageNet image set has collected more than 9,353,897 Internet images with sufficient visual diversity and it contains more than 14,791 object classes and image concepts at different semantic levels. ImageNet currently provides densely sampled SIFT features. The k-means clustering algorithm is performed on a random subset of 10 million SIFT descriptors to generate a visual vocabulary with 1000 visual words.

In this paper, we use Bag of Words (BoW [41]) representations which are provided by ImageNet, so that we can focus on the second issue for large-scale image classification, i.e., classifier design and training. Three hundred object classes and image concepts at different semantic levels are used for assessing the effectiveness of our structural learning algorithm on large-scale classifier training, where a 1000-bin codeword histogram is used for image content representation. Parts of these 300 most popular object classes and image concepts are given in Table 1.

Table 1
Part of 300 object classes and image concepts in ImageNet [40] for algorithm evaluation.

fox	cat	bear	bee	fish
hare	abacus	aircraft	ballon	tiger
desk	mirror	boat	iPod	coat
suit	tank	cherry	banana	flower
sandbar	beach	valley	daisy	rail
cliff	dam	clock	computer	···

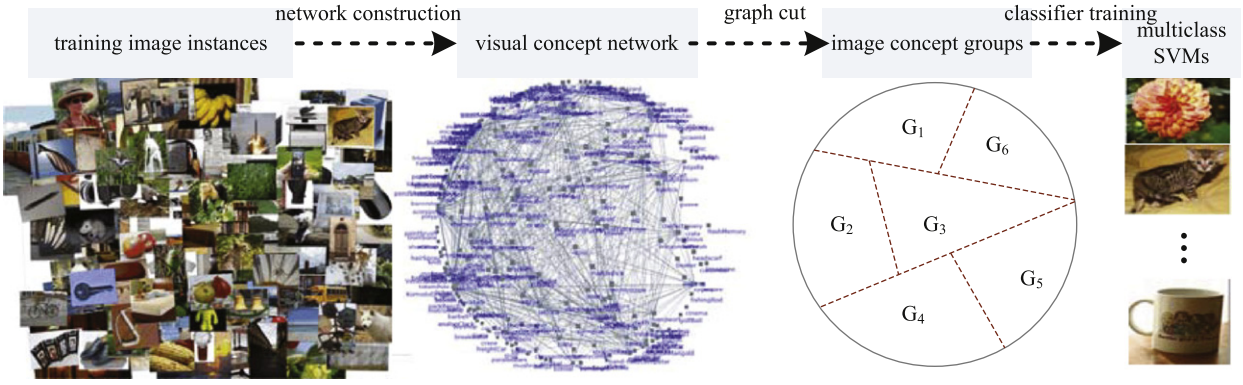


Fig. 1. Flowchart of our structural learning algorithm for inter-related classifier training.

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

The visual concept network consists of two key components: *object classes or image concepts* and *their inter-concept visual correlations*. For two given object classes or image concepts C_i and C_j, their inter-concept visual similarity context $\gamma(C_i,C_j)$ is defined as

$$\gamma(C_i,C_j)=\frac{1}{N_i\cdot N_j}\sum_{h\in C_i}\sum_{k\in C_j}\rho(h,k) \tag{1}$$

where N_i and N_j are the total numbers of image instances for the object classes or image concepts C_i and C_j, respectively, $\rho(h,k)$ is the kernel function for characterizing the visual similarity context between the image instances h and k for C_i and C_j:

$$\rho(h,k)=\exp\left(-\frac{\chi^2(h,k)}{\sigma}\right). \tag{2}$$

As mentioned above, each image instance is represented by using a 1000-bin histogram (BoW), namely we can denote h and k as: $h=(h_1,h_2,\ldots,h_{1000})$ and $k=(k_1,k_2,\ldots,k_{1000})$. A χ^2 distance-based kernel function is used to measure the visual similarity context between two image instances. So the kernel function $\rho(h,k)$ is reformulated as

$$\rho(h,k)=\exp\left(-\sum_{l=1}^{1000}\frac{\chi_l^2(h,k)}{\sigma_l}\right) \tag{3}$$

where the $\chi_l^2(h,k)$ is the χ^2 distance between the lth entry of h and k:

$$\chi_l^2(h,k)=\frac{1}{2}\cdot\frac{(h_l-k_l)^2}{h_l+k_l} \tag{4}$$

Normally, $\sigma=(\sigma_1,\sigma_2,\ldots,\sigma_{1000})$ is setting to the mean values of the χ^2 distances. We have found, however, if the size of visual vocabulary is very large, σ would be very small, so all the visual similarity contexts will be near to zero. Thus a logarithmic transformation is used:

$$\sigma_l=1\Big/\left|\log\left[\frac{1}{N_iN_j}\sum_{h\in C_i}\sum_{k\in C_j}\chi_l^2(h,k)\right]\right| \tag{5}$$

The inter-concept visual similarity contexts $\gamma(\cdot,\cdot)$ are further normalized to the range [0,1].

Some experimental results on the inter-concept visual correlations $\gamma(\cdot,\cdot)$ are given in Table 2, and larger values of $\gamma(\cdot,\cdot)$ mean that there are stronger visual correlations among the corresponding object classes and image concepts.

The visual concept network for our test image set is illustrated in Fig. 2 (left) and some examples are given in Fig. 2 (right), where each object class or image concept is linked with multiple inter-related object classes and image concepts with larger values of the inter-concept visual similarity contexts $\gamma(\cdot,\cdot)$. A full inter-concept visual correlation map for 100 object classes and image concepts is illustrated in Fig. 3(a), where the luminance of the color bar denotes the strength of the inter-concept visual correlations. A representative part and its inter-concept visual similarity matrix are given in Fig. 3(b) and (c), respectively. It is worth noting that different object classes and image concepts have different numbers of inter-related objects classes and image concepts on the visual concept network.

Our visual concept network can: (a) characterize the inter-concept visual similarity contexts intuitively and identify the inter-related learning tasks directly in the visual feature space; (b) provide a good environment to leverage the training instances for multiple inter-related object classes and image concepts to train their inter-related classifiers jointly, i.e., integrating their training instances to learn their common prediction components (which are shared among their inter-related classifiers) jointly. Training the inter-related classifiers jointly can enhance their discrimination power significantly.

Table 2
Part of inter-concept visual similarity contexts.

Object pair	γ	Object pair	γ	Object pair	γ
tailed frog-start fish	0.87	shaver-watch	0.41	CD player-wok	0.26
lion-kit fox	0.81	jeep-stone wall	0.51	speed boat-aircraft carrier	0.44
abacus-mug	0.17	scabious-bee	0.64	ballon-airline	0.21
desk-bonsai	0.26	snake-flash	0.12	wattle-pheasant	0.74
sandbar-seashore	0.57	church-stone wall	0.37	suit-kimono	0.33
life boat-seashore	0.48	kit fox-cheetah	0.90	sun flower-chrysanthemum	0.49

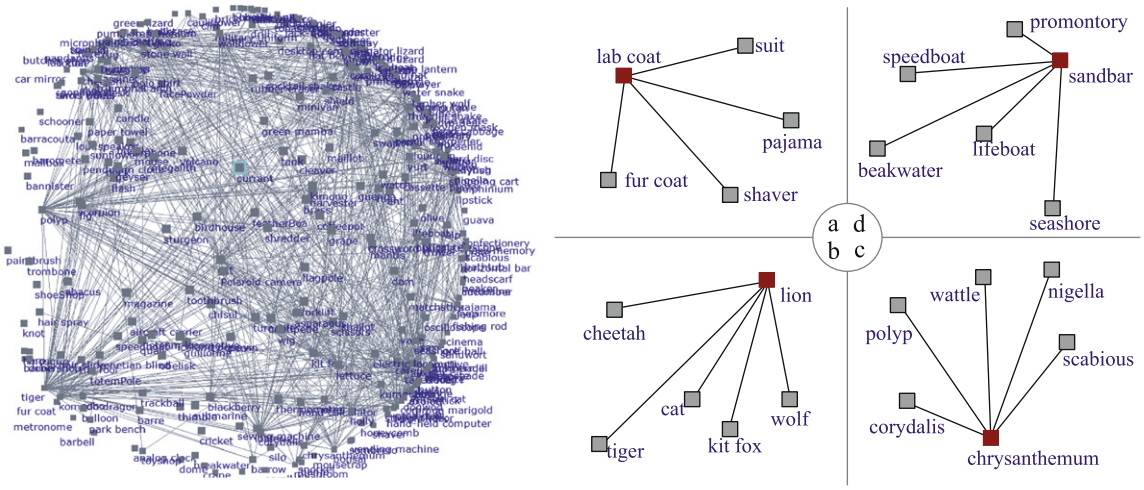
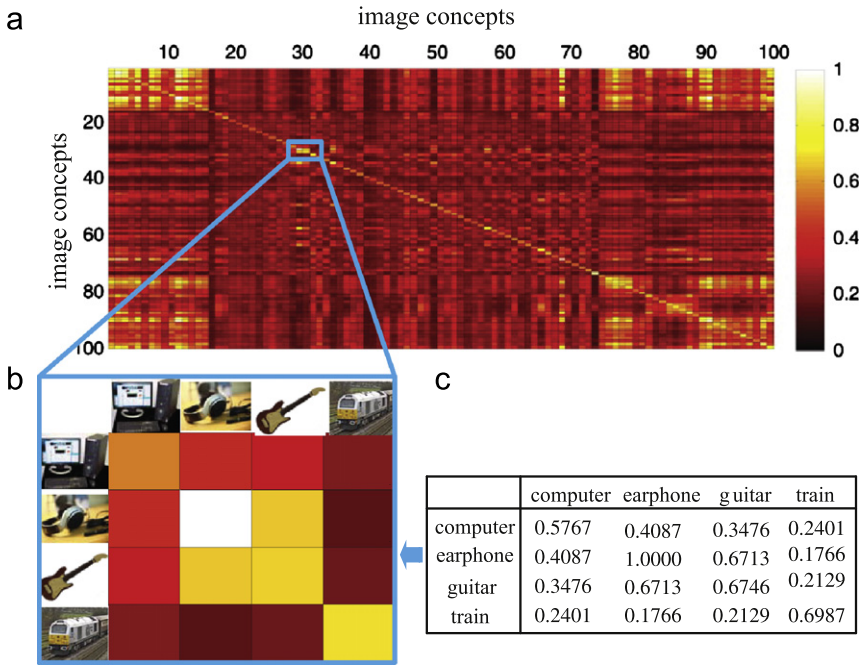


Fig. 2. Our visual concept network with 300 object classes and image concepts (left) and some examples for the inter-related object classes and image concepts (right).

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029



	computer	earphone	guitar	train
computer	0.5767	0.4087	0.3476	0.2401
earphone	0.4087	1.0000	0.6713	0.1766
guitar	0.3476	0.6713	0.6746	0.2129
train	0.2401	0.1766	0.2129	0.6987

Fig. 3. The correlation map for 100 object classes and image concepts (a). A representative part with four classes (i.e., "computer", "earphone", "guitar" and "train") is shown in (b). The corresponding inter-concept visual similarity contexts are given in (c). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

4. Classifier design and training

To support keyword-based image retrieval, it is very attractive to develop new algorithms for learning more accurate mapping functions (i.e., classifiers) between the low-level visual features and the high-level object classes and image concepts. From our visual concept network, the inter-concept visual correlations can be represented precisely by the strengths of $\gamma(\cdot,\cdot)$. To design our multi-class structural classifiers, support vector machine (SVM [42]) is employed as the basis learner since it has achieved superior performance in domains such as image retrieval and classification [43–45].

As shown in Fig. 2, a given object class or image concept is strongly related to multiple object classes and image concepts on the visual concept network and their image instances share some common visual properties. For example, the concept "sandbar" is strongly related to "promontory", "speedboat", "beakwater", "lifeboat" and "seashore". Thus it is not appropriate to completely ignore their inter-concept visual correlations and train their inter-related classifiers independently. In this paper, a structural learning scheme is developed to leverage such inter-concept visual correlations for training a large number of inter-related classifiers jointly, which can significantly enhance the discrimination power of the inter-related classifiers while restraining their error transmission effectively.

In our visual concept network, the object classes and image concepts, which have larger values of $\gamma(\cdot,\cdot)$, are linked together. For some object classes and image concepts, their inter-concept visual correlations could be very weak (i.e., having smaller values of $\gamma(\cdot,\cdot)$), thus it is not necessary for each object class and image concept to be linked with all the others on the visual concept network. To reduce the computation complexity for inter-related classifier training, the object classes and image concepts on our visual concept network are partitioned into a set of groups according to the strength of their inter-concept visual similarity contexts. The object classes and image concepts in the same group will share some common or similar visual properties and their classifiers should be trained jointly since they are strongly inter-related. On the other hand, the object classes and image concepts in different groups will contain various visual properties and the inter-group visual correlations are much weaker than the intra-group visual correlations, thus the classifiers for the object classes and image concepts in different groups can be trained independently. Theoretically, any unsupervised clustering method can be used to cluster the object classes and image concepts on our visual concept network into a set of groups. In this paper, the *normalized cut* (N-cut [46]) algorithm is employed because our visual concept network is essentially a graph.

Table 3 illustrates the results for clustering 100 object classes and image concepts into 20 groups. One can observe that most of

Table 3
The grouping results for 100 object classes and image concepts of Fig. 3.

Group	Items
1	kit fox, lion, brown bear, ice bear, hare, orangutan, cat, guenon
2	earphone, guitar, optical telescope, sax, scissors
3	abacus
4	feather boa, fur coat, headscarf, kimono, pajama, polo shirt, pullover, stole, suit
5	star fish, water snake, shrimp, tailed frog, komodo dragon, otter
6	Chinese lantern, jack-o-lantern, flash, balloon, candle
7	web site
8	hand-held computer, hand calculator, computer
9	aircraft carrier, fire boat, life boat, speed boat, submarine
10	loudspeaker
11	mailbox
12	bonsai, blackberry, cherry, cymbid, fig, head cabbage, cauliflower, spinach, lettuce
13	daisy, chrysanthemum, sunflower, scabious, bee
14	car mirror, CD player, flash memory, golf ball, hard disc, iPod, mouse
15	geyser, volcano, alp, valley
16	coffeepot, mug, wok
17	stone wall, church, desk, tank, mosque, obelisk
18	promontory, sandbar, seashore, dune
19	ginko, wattle, banana, pandanus, palm
20	electric locomotive, steam locomotive, model T, racer, shopping cart, ambulance, bobsled, gondola, jeep, limousine, minivan, train

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

the object classes and image concepts in the same group are indeed visually similar, which has good consistence with human perception and cognition. However, there are some exceptions, for example, "tank" *vs.* "desk" (in group 17). Human would not think they are visually similar, but they are clustered into the same group. One explanation for the appearance of such phenomenon is that feature-based interpretations of images (i.e., SIFT features in this paper) are more or less inconformity with human perceptions at semantic level.

For the inter-related object classes and image concepts in the same group, a structural leaning algorithm is performed to learn their inter-related classifiers jointly by sharing a common prediction component [39]. In order to describe our proposed structural leaning algorithm clearly, we predefine some notations shown in Table 4.

Our group-based multi-class structural learning architecture is illustrated in Fig. 4. We denote the object classes and image concepts as C_1, C_2, …, C_N, and the groups as G_1, G_2, …, G_M. For a given group G_i, it consists of N_i inter-related object classes and image concepts. Obviously, $\sum_{i=1}^{M} N_i = N$, where N is the total number of object classes and image concepts on our visual concept network and M is the total number of groups. For example, in Table 3, we have $N=100$ and $M=20$.

All these groups are determined automatically by performing N-cut clustering algorithm over our visual concept network. It is worth noting that the object classes and image concepts in the same group have stronger inter-concept visual correlations than those in different groups, thus the classifiers for the inter-related object classes and image concepts in the same group are strongly inter-related and should be trained jointly to enhance their discrimination power. It is easy for us to design multiple discrimination functions $h_{G_i}(x)$ to distinguish the object classes and image concepts in different groups, because the inter-group visual correlations are much weaker. On the other hand, it could be more difficult to distinguish the inter-related object classes and image concepts in the same group because their inter-concept visual correlations are stronger. Thus multiple intra-group classifiers $h_{C_i}(x)$ should be designed and learned for distinguishing the inter-related object classes and image concepts in the same group.

From the architecture of our visual concept network shown in Fig. 4, one can observe that the inter-related classifiers consist of at least two components: (1) the inter-group discrimination functions $h_{G_i}(x)$ (i.e., prediction components for separating one certain group from other groups on the visual concept network); and (2) the intra-group discrimination functions $h_{C_i}(x)$ (i.e., prediction components for separating one particular object class or image concept from its inter-related object classes and image concepts in the same group). Our inter-related classifiers are defined as

$$H_{C_i}(x) = \Lambda(h_{C_i}(x), h_{G(C_i)}(x)) \tag{6}$$

where $i = 1,2,3, \ldots, N$, N and $G(C_i)$ are defined in Table 4. The function $h(x)$ is a combination of some basis learners, and the function $\Lambda(\cdot,\cdot)$ denotes the combination mode of $h_{C_i}(x)$ and $h_{G(C_i)}(x)$. Experimentally, the combination mode is hardly unique. Obviously, different combination modes will result in different architectures for supporting structural learning. In this paper, the *weighted sum method* is adopted for constructing our structural learning algorithm.

Table 4
Some tokens used in this paper.

Notation	Description
C_i	a given object class
G_i	a given group contains some objects
G_g	$G_g = G(C_i)$, the group which C_i belongs
Ω_{G_i}	$\Omega_{G_i} = \{C_j \mid C_j$belongs to group$G_i\}$
T	a set of all groups $T = \{G_1, G_2, \ldots, G_M\}$
M	the total number of groups
N	the total number of classes

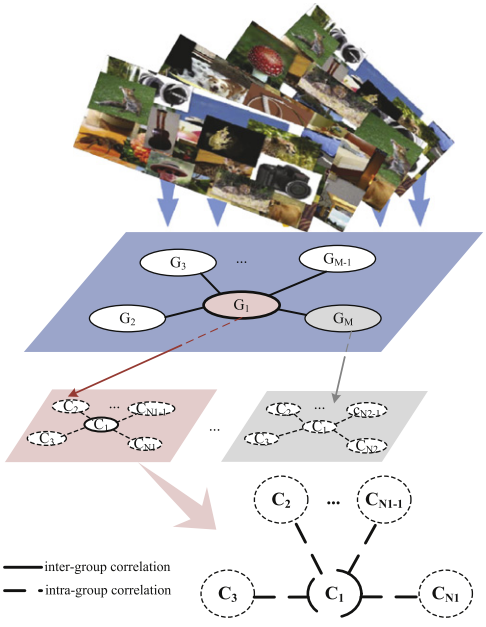


Fig. 4. Our group-based structural learning architecture.

4.1. Our hybrid design approach

In order to restrain the issue of *error transmission*, we use a weighted sum to instantiate (6) as

$$H_C(x) = \alpha_C \cdot h_C(x) + \beta_C \cdot h_{G(C)}(x) \tag{7}$$

where C is one particular object class or image concept and $G(C)$ is the corresponding group for the given object class or image concept C. α_C and β_C are the weighted coefficients, which are subjected to $\alpha_C + \beta_C = 1$.

The item $\alpha_C \cdot h_C(x)$ indicates the intra-group classifier while the item $\beta_C \cdot h_{G(C)}(x)$ indicates the inter-group classifier. However, Eq. (7) may be *weak* to make a final decision. A simple example is illustrated in Fig. 5. Assuming there are nine object classes and image concepts $C_1 \sim C_9$, which are clustered into three groups $G_1 \sim G_3$. Without loss of generality, let us take the object class or image concept C_5 into consideration. For the intra-group item (the dash line in Fig. 5) in (7), it would be able to distinguish C_5 from C_4 and C_6. And the inter-group item $\beta_C \cdot h_{G(C)}(x)$ (solid line in Fig. 5) can also be able to distinguish G_2 from G_1 and G_3. Theoretically, for all the object classes and image concepts C_4, C_5 and C_6, the item $\beta_C \cdot h_{G(C)}(x)$ may have same value. Because the inter-concept visual similarity contexts for the object classes and image concepts in the same group are much stronger than those of the object classes and image concepts in different groups, it is still hard to discriminate C_5 from C_4 and C_6.

In order to enhance the discrimination power of the classifiers (7) for the object classes and image concepts in the same group (e.g., C_4, C_5 and C_6), a *hybrid level* item (dash and dot line in Fig. 5) is added to the classifiers (7), then the discrimination function for

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

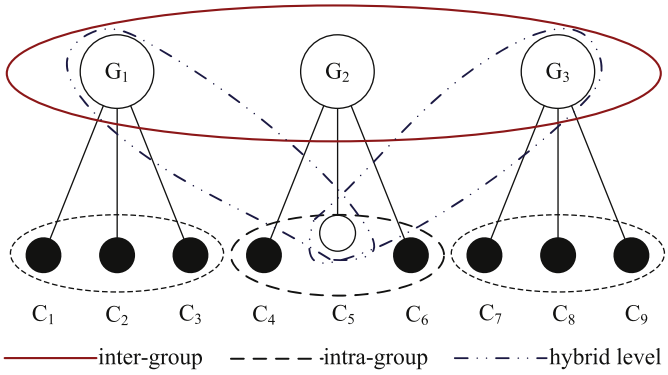


Fig. 5. The design of our structural learning algorithm: an example for nine image concepts and object classes which are partitioned into three groups.

the object class or image concept C becomes

$$H_C(x) = \alpha_C \cdot h_C(x) + \beta_C \cdot h_{G(C)}(x) + \gamma_C \cdot h_{R(C)}(x) \tag{8}$$

similar with (7), α_C, β_C and γ_C are subjected to

$$\alpha_C + \beta_C + \gamma_C = 1$$

The hybrid level item $h_{R(C)}(x)$ is used to discriminate the object class or image concept C from other groups (here *the other groups* mean all the groups except the one which the object class or image concept C belongs). So we can divide our inter-related classifiers into three levels: *the intra-group level*, *the inter-group level* and *the hybrid level* corresponding to three items in (8), respectively.

Intuitively, our structural learning algorithm relies on the fact that there are natural groupings for large numbers of object classes and image concepts, i.e., common "attributes". According to the description above, the overall structural learning algorithm for training multiple multi-class structural SVM classifiers is summarized as follows.

(1) *Selecting the basis learner*. In this paper, SVM is employed as our basic binary classifier. The basic binary discrimination function is denoted as $f(x, C_+, C_-)$, where x is the input test instance, C_+ denotes the positive training concept(s), and C_- denotes the negative training concept(s). Namely, all the training instances in C_+ are the positive samples, while all the training instances in C_- are the negative samples.

(2) *Designing and training the intra-group classifiers*. After the concept clustering process, the object classes and image concepts are partitioned into a set of groups. For each group, the number of object classes and image concepts is small, so we can employ the pairwise method and one-against-all method cohesively. Assume that there are N object classes and image concepts, C_1, C_2, …, C_N, then we should train N classifiers $h_{C_1}, h_{C_2}, \ldots, h_{C_N}$:

$$h_{C_i}(x) = \sum_{C_j \in \{\Omega_{C_g} \setminus C_i\}} \alpha_{ij} f(x, C_i, C_j) + \alpha_{ii} f(x, C_i, \{\Omega_{C_g} \setminus C_i\}) \tag{9}$$

where $G_g = G(C_i)$ is the group which C_i belongs ($g \in \{1,2,\ldots,M\}$), M is the total number of groups and $\alpha_{i\cdot}$ are the weighted coefficients. The first term in (9) denotes the pairwise method while the second one denotes the one-against-all method. These two items are combined by a weighted sum. It is noteworthy that the subscript j is discontinuous. The weighted coefficients are subjected to $\sum_{j \in \{j | C_j \in \Omega_{C_g}\}} \alpha_{ij} = 1$.

(3) *Designing and training the inter-group classifiers*. Assuming that there are M groups totally, which are denoted as G_1, G_2, …, G_M. Each group is considered as an independent entity. Similar with (2), both the pairwise method and the one-against-all method are employed cohesively. Then we should train M classifiers $h_{G_1}, h_{G_2}, \ldots, h_{G_M}$:

$$h_{G_g}(x) = \sum_{G_j \in \{T \setminus G_g\}} \beta_{gj} f(x, \Omega_{G_g}, \Omega_{G_j}) + \beta_{gg} f(x, \Omega_{G_g}, \Omega_{\{T \setminus G_i\}}) \tag{10}$$

where $\{T \setminus G_g\}$ contains all the $(M-1)$ groups except G_g, and $\beta_{g\cdot}$ are the weighted coefficients. Similar with (9), the first term in (10) denotes the pairwise method while the second one denotes the one-against-all method. The weighted coefficients are subjected to $\sum_{j=1}^{M} \beta_{gj} = 1$.

(4) *Designing and training the hybrid classifiers*. This process is different from (2) or (3). These hybrid classifiers are used to discriminate one particular object class or image concept from those object classes and image concepts in other $(M-1)$ groups. For example as shown in Fig. 5, a hybrid classifier is used to distinguish C_5 from G_1 and G_3. One choice is to design a "one-against-all" classifier. That is, for the given object class or image concept C, all the training samples in the other $M-1$ groups (on our visual concept network) are treated as negative training samples. This approach may be feasible, but we do not use it due to both the excessive unbalance and the high computational cost of the one-against-all method. Rather than designing a single unbalance classifier, we design $M-1$ classifiers for separating the given object class or image concept C from all the other groups individually, that is:

$$h_{R(C_i)}(x) = \sum_{G_j \in \{T \setminus G_g\}} \gamma_{ij} f(x, C_i, \Omega_{G_j}) \tag{11}$$

where $G_g = G(C_i)$ is the group which C_i belongs. For all the weight coefficients we have $\sum_{j \in \{j | G_j \in \{T \setminus G_g\}\}} \gamma_{ij} = 1$.

According to Eqs. (8)–(11), the final discrimination function can be formulated as

$$\begin{aligned} H_{C_i}(x) &= \alpha_{C_i} \cdot h_{C_i}(x) + \beta_{C_i} \cdot h_{G_i}(x) + \gamma_{C_i} \cdot h_{R(C_i)}(x) \\ &= \sum_{C_j \in \{\Omega_{C_g} \setminus C_i\}} \alpha_{C_i} \alpha_{ij} f(x, C_i, C_j) + \alpha_{C_i} \alpha_{ii} f(x, C_i, \{\Omega_{C_g} \setminus C_i\}) \\ &\quad + \sum_{G_j \in \{T \setminus G_g\}} \beta_{C_i} \beta_{gj} f(x, \Omega_{G_i}, \Omega_{G_j}) \\ &\quad + \beta_{C_i} \beta_{gg} f(x, \Omega_{G_i}, \Omega_{\{T \setminus G_i\}}) \\ &\quad + \gamma_{C_i} \sum_{G_j \in \{T \setminus G_g\}} \gamma_{ij} f(x, C_i, \Omega_{G_j}) \end{aligned} \tag{12}$$

where $G_g = G(C_i)$ is the group which the object class or image concept C_i belongs.

In order to simplify the problem, we use the average weights for classifier combination in our structural learning algorithm. Specifically, we use the average weight strategy to determine the weighted parameters in (8)–(12), respectively. We take β_{gj} in (10) as an example, all the weight parameters are specified an equal value and the sum of β_{gj} is equal to 1. Hence, all the weights are subjected to $\sum_{j \in \{j | C_j \in \Omega_{C_g}\}} \alpha_{ij} = 1$, $\sum_{j \in \{1,2,\ldots,M\}} \beta_{gj} = 1$, $\sum_{j \in \{j | G_j \in \{T \setminus G_g\}\}} \gamma_{ij} = 1$, and $\alpha_{C_i} = \beta_{C_i} = \gamma_{C_i} = 1/3$.

In the classifier training stage, we need to train N discrimination functions: $H_{C_1}(x)$, $H_{C_2}(x)$, …, $H_{C_N}(x)$. In the test stage, for any given image instance x, we test it among N classifiers, which are combined to make a final decision. The classifier, which generates the highest confidence value, is selected as the winner:

$$c = \arg\max_{C_i} P(C_i | x) = \arg\max_{C_i} H_{C_i}(x) \tag{13}$$

In case that two classes have identical decision value, we use the same strategy as in [47] that we simply choose the class appearing first in the array of storing class names.

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

4.2. Computational complexity

Most existing algorithms for multi-class object detection and scene recognition are often handled by combining multiple binary classifiers, thus they may have square complexity with the number of object classes and image concepts N, i.e., the complexity is $O(N^2)$. On the other hand, our structural learning algorithm just needs to combine N_{struct} binary classifiers, and N_{struct} is determined as

$$N_{struct} = \underbrace{\sum_{i=1}^{M}\left(\frac{N_i(N_i-1)}{2}+N_i\right)}_{intra\text{-}group} + \underbrace{\frac{M(M-1)}{2}+M}_{inter\text{-}group} + \underbrace{N(M-1)}_{hybrid\text{-}level}$$
$$= \frac{1}{2}\sum_{i=1}^{M} N_i^2 + \frac{1}{2}M^2 + NM + \frac{1}{2}M - \frac{1}{2}N \quad (14)$$

where M is the number of groups and N_i is the number of object classes and image concepts for the ith group.

Typically, $N_i \ll N$ and $M \ll N$, thus our structural learning algorithm can achieve sub-quadratic complexity with the number of object classes and image concepts N, i.e., $O(\frac{1}{2}\sum_{i=1}^{M} N_i^2 + \frac{1}{2}M^2 + NM + \frac{1}{2}M - \frac{1}{2}N)$. As a result, our structural learning algorithm is very attractive for training a large number of inter-related classifiers for large-scale image classification.

By using a group-based approach to model the task relatedness explicitly, the issue of huge inter-concept visual similarity can be addressed more effectively by leveraging the task relatedness for training the inter-related classifiers jointly. Thus the discrimination power of the inter-related classifiers can be enhanced significantly by learning from the image instances for the inter-related object classes and image concepts. Incorporating the image instances from other inter-related object classes and image concepts for inter-related classifier training can significantly enhance the generalization ability of the classifiers, especially when the image instances for the given object class or image concept are not representative for large amounts of unseen test images.

5. Algorithm evaluation

5.1. Basic setup

To evaluate our structural learning algorithm, extensive experiments are carried out on two datasets: (1) a subset from ImageNet [40] dataset with 300 most popular object classes and image concepts and (2) Caltech256 dataset [48]. We consider the LIBSVM [47] with non-linear RBF kernel as our basic learner for each binary discrimination function $f(x,C_i,C_j)$. For each problem (basic learner), we use a fivefold cross-validation to determine the kernel parameters γ and the cost parameters C among the range $\gamma=[2^{-10},2^{-9},2^{-8},\ldots,2^4]$ and $C=[2^0,2^1,2^2,\ldots,2^{14}]$. Therefore, for each problem we try 225 combinations to complete the cross-validation process. We perform this cross-validation procedure on every basic learner (i.e., every binary discrimination function $f(x,C_i,C_j)$ shown in (12) in our paper) on its own separate training set. For example, we perform cross-validation on a basic learner $f(x,C_i,C_j)$ on the validation examples (image instances) from class C_i and C_j.

The goal for algorithm evaluation is to estimate the average classification accuracy rates and average F scores of our structural learning algorithm, where the experiments are carried out on the following aspects:

(a) we compare our algorithm with the traditional approaches, i.e., the *one-against-all* approach and the *pairwise* approach;

(b) we focus on comparing our algorithm with two state-of-arts algorithms for inter-related classifier training, including the *JointBoost algorithm* [18] and the *Hierarchical Method* [34];

(c) we compare the performance of our structural learning algorithm by leveraging different structures for inter-related classifier training, i.e., our hybrid combination *vs.* the tree-like hierarchy. For the tree-like hierarchy, we decompose the inter-related object classes and image concepts into two levels: inter-group and intra-group. We train the inter-group classifiers h_{G_i} and intra-group classifiers h_{C_i} independently. For a given test image instance x, firstly we should determine which group it belongs. We test it over all the M inter-group classifiers $h_{G_1} \sim h_{G_M}$, and the group-based classifier which has the highest confidence value is selected as the winner and the corresponding group index G_g is assigned to x. And then, we repeat the same procedure over all the intra-group classifiers h_{C_j}, where h_{C_j} is the intra-group classifier of the object class or image concept C_j, which belongs to the group G_g.

(d) we analyze the generalizability of our structural learning algorithm and other algorithms (both the traditional approaches and inter-related approaches) when they are used to deal with different numbers of object classes and image concepts (i.e., different category sizes);

(e) we also compare our visual concept network with other similarity measurements, including the ontologies from WordNet [25] and NGD [26]. WordNet [25] is one of the most popular semantic networks for a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet database contains 155,287 words organized in 117,659 synsets for 206,941 word–sense pairs. In our experiments, the similarity computation software module from [49] is used to calculate the WordNetsemantic similarities between different object classes and image concepts. Google distance is proposed to calculate the contextual relationship between two concepts by using their correlation in their search results from Google search engine when two concepts are used as query terms. For two given concepts x and y, the normalized Google distance (NGD) [26] is defined as

$$NGD(x,y) = \frac{\max\{\log f(x),\log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x),\log f(y)\}} \quad (15)$$

where $f(x)$ denotes the number of pages containing x, $f(y)$ denotes the number of pages containing y and $f(x,y)$ denotes the number of pages containing both x and y. N is the total number of web pages indexed by Google search engine. In this experiment, we use a approximate value 10^{12} for N.

5.2. Algorithm evaluation on imagenet

We directly use 1000-bin BoW histograms (which are provided by ImageNet) for image content representation. For each image concept or object class, 200 image instances are used as the training samples, and another 200 image instances (which are different from the 200 training samples) are used as the test samples. Besides that, 100 image instances are used as cross-validation samples to select proper SVM parameters. Totally about 150k image instances are employed in our experiments.

Experiment 1: *our method vs. traditional approaches*. In this experiment, we aim at comparing our method with the *one-against-all approach* and the *pairwise approach*. Three hundred most popular object classes and image concepts from ImageNet are used in our experiments. The classification accuracy rates for

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

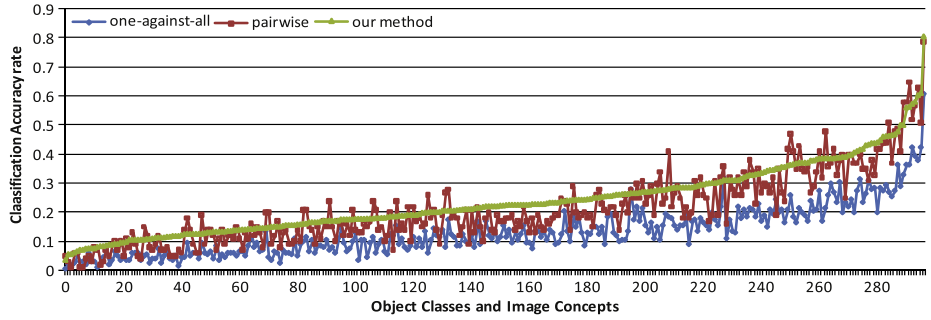


Fig. 6. Performance comparison on ImageNet with 300 object classes and image concepts, where the object classes and image concepts are illustrated in the orders of their accuracy rates for our structural learning algorithm.

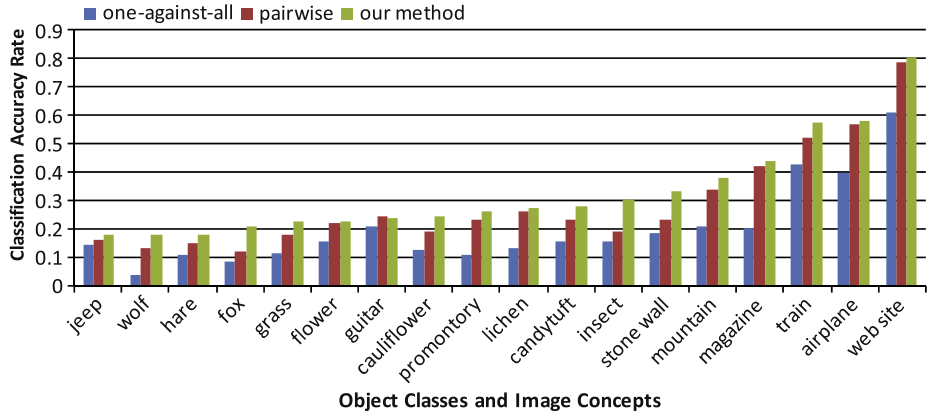


Fig. 7. Performance comparison on classification accuracy rates: our algorithm *vs.* traditional approaches.

all these 300 object classes and image concepts are shown in Fig. 6. All the object classes and image concepts in Fig. 6 are illustrated according to their classification accuracy rates from low to high for our structural learning algorithm, that is why our approach is very smooth while the other algorithms show random behaviors. Fig. 7 illustrates the classification accuracy rates for some object classes and image concepts. From these experimental results one can observe that our structural learning algorithm has obtained very competitive results compared with the traditional algorithms. The significant improvement on the classification accuracy rates benefits from: the inter-related classifiers for the inter-related object classes and image concepts are trained jointly by leveraging their inter-concept visual correlations for inter-related classifier training.

In addition, one can observe that the classification accuracy rates are bad for a small number of object classes and image concepts. This phenomenon is mainly caused by the issue of huge intra-concept diversity for some object classes and image concepts (which are called "hard" object classes and image concepts). These "hard" object classes and image concepts are usually difficult to be differentiated from others because the images from these "hard" object classes and image concepts have huge diversity on their visual properties, which may result in higher chances to be overlapped with other object classes and image concepts. The classification accuracy rates for them are usually very low especially when a large number of object classes and image concepts come into view. In our experiments, some examples of such "hard" object classes and image concepts are "moth" (indexed n02283201 in ImageNet dataset), "syringe" (n04376876 in ImageNet) and "banjo" (n02787622 in ImageNet).

Experiment 2: *our method vs. other inter-related approaches*. Two state-of-arts inter-related approaches, the JointBoost [18] algorithm and the Hierarchical algorithm [34], are compared with our structural learning mechanism in this experiment. For the JointBoost [18] algorithm, the basic learner is the boosted decision stump, and the classifier is generated by using 5000 rounds of boosting. For the Hierarchical algorithm, we use 100 image instances from each of these 300 object classes and image concepts to construct the hierarchies. The classification accuracy rates for these 300 object classes and image concepts are shown in Fig. 8. Fig. 9 illustrates the classification accuracy rates for some object classes and image concepts. We can see that our structural learning algorithm makes more efficient use of the inter-concept visual correlations for inter-related classifier training.

It is worth noting that among these three approaches (which are compared in our experiments), the JointBoost method performs worst on the classification accuracy rates for all these 300 object classes and image concepts, as shown in Figs. 8 and 9. These results may be surprising. Firstly, we recall the overall algorithm of JointBoost in Table 5 concisely. As shown in Table 5, in order to address the issue of searching all possible 2^C-1 subsets of classes, it uses the *best-first search* and a forward selection procedure, the overall complexity of which is $O(C^2)$ rather than $O(2^C)$. $S(i)$ is denoted as a subset of classes that are shared in the ith boosting round. At the decision stage, for each image class c, we find all the subsets $S(\cdot)$ that contain c, and sum up their additive models to give the final form of the classifiers:

$$H(v,c) = \sum_{s \in \{s | s \in S(\cdot), c \in s\}} G^s(v) \tag{16}$$

where v is a vector of features and c is the class. However, this JointBoost procedure would not be able to guarantee that every class would be shared through the boosting procedure.

In the example shown in Table 6, there are five classes $c_1 \sim c_5$ and the number of booting rounds is 4. The shared subsets in each boosting round is $S(1)=\{c_2,c_4,c_5\}$, $S(2)=\{c_1,c_5\}$, $S(3)=\{c_2,c_4\}$,

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

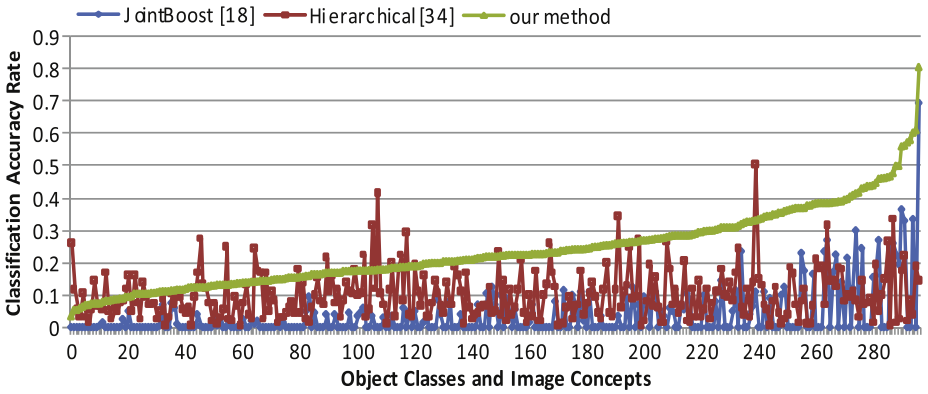


Fig. 8. Performance comparison on ImageNet with 300 object classes and image concepts, where the object classes and image concepts are illustrated in the orders of their accuracy rates for our structural learning algorithm.

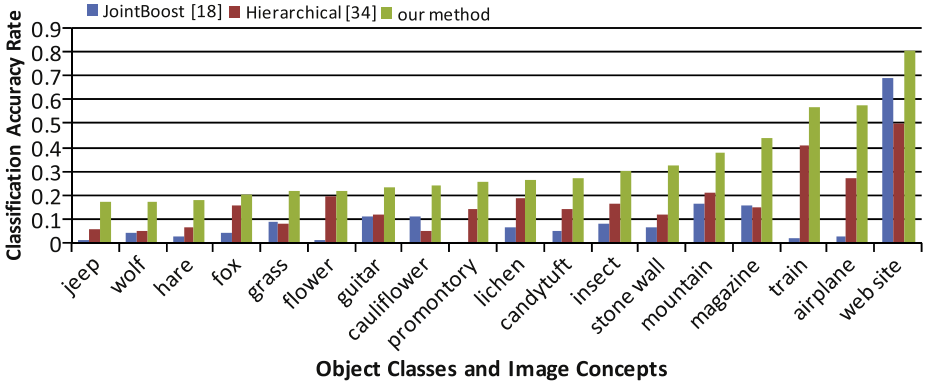


Fig. 9. Performance comparison on classification accuracy rates: our algorithm *vs.* other inter-related approaches.

Table 5
The overall algorithm of JointBoost [18].

v_i is the sample vector, N is the number of training samples, and C is the number of classes and M is the boosting rounds.

(1) Initialize the weights ω_i^c and set $H(v_i,c)=0$, $i=1,2,\ldots,N$, $c=1,2,\ldots,C$
(2) Repeat for $m=1,2,\ldots,M$
  Initialize the subset that shares features $S(m)=\emptyset$.
    repeat for $n=1,\ldots,C$
      (a) repeat for p in $\{\{1,2,\ldots,C\}\backslash S(m)\}$
        (i) $s=S(m)\bigcup p$
        (ii) fit shared stump
        (iii) evaluate error $J(s)$;
      (b) find best sub set $s^*=\text{argmin}_s J(s)$
        if $J(s^*)<J(S(m))$
          update the subset $S(m)=s^*$
  Updates the weights.

$S(4)=\{c_5\}$, respectively. One can observe that none of the subsets contains c_3. In this case, the test samples which come from c_3 would not be classified correctly by the JointBoost classifier, because $H(v,c_3)$ is always equal to zero, i.e., the initial value. When the number of object classes and image concepts becomes large, this phenomenon becomes more common. In Fig. 10, we compare the shared classes in the first 30 boosting rounds for 3 object classes and image concepts (left) and 48 object classes and image concepts (right). All the object classes and image concepts are selected from ImageNet image set. The indexes of the boosting rounds are illustrated as the vertical axis in Fig. 10. The white grid(s) located in the same line in Fig. 10 denotes the classes that are shared in one boosting round. For small-scale (about tens of classes/concepts/categories) image classification tasks, we can make a conclusion that the JointBoost algorithm will perform very well. However, for a large-scale or medium-scale (thousands or hundreds of classes) image classification tasks, the JointBoost algorithm will perform significantly worse than our inter-related learning algorithm.

Experiment 3: *performance comparison with the tree-like hierarchy*. It is worth noting that our structural learning algorithm is different from the tree-like hierarchy which is commonly employed in the previous work [34,37]. In a tree-like architecture, a leaf node is independent with its non-parent nodes (i.e., the high-level nodes at the parent level which are not the parent node for the given leaf node). However, in our architecture, the correlations between the nodes of object classes and the nodes for other groups (which the given object class does not belong to) are also taken into consideration for inter-related classifier training.

A comparison result on the average classification accuracy rates for 100 object classes and image concepts between our structural learning algorithm and the tree-like approach described above is shown in Table 7. Low computational cost is the most significant advantage of the tree-like hierarchical method, since in the test stage it only needs to test among the inter-group classifiers and the corresponding intra-group classifiers. However, the classification error in the inter-group level may be completely transmitted to the intra-group level. By suppressing the error transmission from the inter-group level classifiers to the intra-group classifiers, one can observe that our structural learning approach can enhance the discrimination power of the inter-related classifiers significantly.

Experiment 4: *performance comparison on different category sizes*. In order to verify the efficiency of our structural learning algorithm on different numbers of object classes and image

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

concepts (i.e., different category sizes), we have also compared the average classification accuracy rates of the four algorithms on different size of categories. The experimental results on 100, 200 and 300 image categories for different algorithms are shown in Fig. 11. The number of boosting rounds for the JointBoost algorithm is 3000, 4000 and 5000, respectively. From these results, one can observe that the classification accuracy rates reduce significantly when the number of object classes and image concepts increases. Even so, our structural learning algorithm still outperforms others and it drops slower than others. From all these results, one can observe that our structural learning algorithm can obtain reasonable accuracy rates for large-scale image classification.

Experiment 5: *performance comparison on different similarity measures*. As what we have emphasized, the visual feature space is the common space for both classifier training and image classification, thus the inter-concept visual similarity is adopted to construct our visual concept network. In this experiment, we aim to compare the visual similarity measurement with the most popular semantic similarity measurements from the label space. The experiments are carried out on 100, 200, and 300 categories independently, and the results of the average classification accuracy rates are illustrated in Fig. 12. From the results, one can observe that our visual similarity measurement can reflect the inter-concept correlations among the object classes and image concepts more accurately. Thus our visual similarity measurement is more suitable for determining the inter-related learning tasks.

In order to cover the false-positive rate, the average F scores for different algorithms are also calculated in this paper, where

$$Fscore = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

From Tables 8 and 9, one can observe that our structural learning algorithm can achieve better performance on the average F scores when the visual similarity measurement is used.

Table 6
An example of shared classes (+) and unshared classes (−) for the JointBoost [18] algorithm.

Boosting round	c_1	c_2	c_3	c_4	c_5
1	−	+	−	+	+
2	+	−	−	−	+
3	−	+	−	+	−
4	−	−	−	−	+

5.3. Algorithm evaluation on Caltech256 dataset

The Caltech256 dataset [48] is another standard multi-class object recognition dataset which contains 30 607 images. We evaluate our structural learning algorithm over 256 classes of the Caltech256 dataset excluding its clutter category. We randomly sample 80 images for each class, and split them into two parts: 40 images per class for training and the remaining for test. In order to describe an image, we first construct a visual vocabulary of 1000 visual words and then extract the Bag-of-words SIFT descriptors on the gray-scale images. Specifically, each image is resized to have a max side length of no more than 300 pixels and raw SIFT [9] features are extracted from it. We perform k-means clustering algorithm on 500,000 raw SIFT features randomly sampled from all Caltech256 images to form a visual vocabulary of 1000 visual words. Subsequently, the SIFT descriptors are quantized using the vocabulary by performing vector quantization

Table 7
Average classification accuracy rates on 100 categories of ImageNet dataset for different classifier structures.

Structures	Accuracy rate
Tree-like Hierarchy Method	0.3480
Our Hybrid Method	**0.3696**

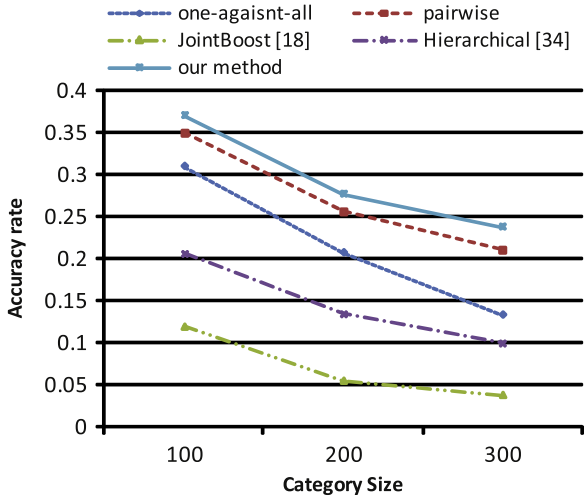


Fig. 11. Average classification accuracy on 100, 200 and 300 categories of ImageNet dataset: our algorithm *vs.* others.

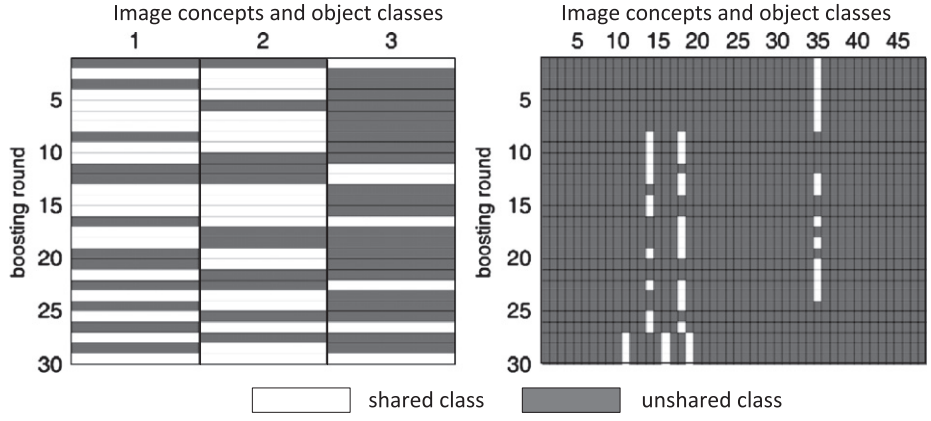


Fig. 10. Schematic diagram of shared and unshared classes in the first 30 boosting rounds for 3 classes (left) and 48 classes (right).

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

(VQ) coding method. All these procedures are implemented by running the VLFeat [50] toolbox.

Experiment 1: *performance comparison with other approaches*. Table 10 shows the average classification accuracy and average *F* scores of our structural learning algorithm and other most related approaches on the Caltech256 dataset. For the JointBoost algorithm [18], the number of the boosting round is 5000. For the Hierarchical method [34], we use k-means method with 40 images per class for constructing the class hierarchy and set $\alpha = 0.2$. Our method can significantly improve the discrimination power of classifiers by leveraging the inter-concept correlations for classifier training.

Experiment 2: *performance comparison with the tree-like hierarchy*. We have also compared our hybrid design of classifiers with the tree-like combination method on Caltech256 dataset. The average classification accuracy rates are shown in Table 11.

Experiment 3: *performance comparison on different similarity measures*. Table 12 shows the experimental results on Caltech256 dataset for different similarity measurements. These results demonstrate that it is much more reasonable to use the visual similarity rather than the semantic or contextual measurements in the label space for the image classification tasks.

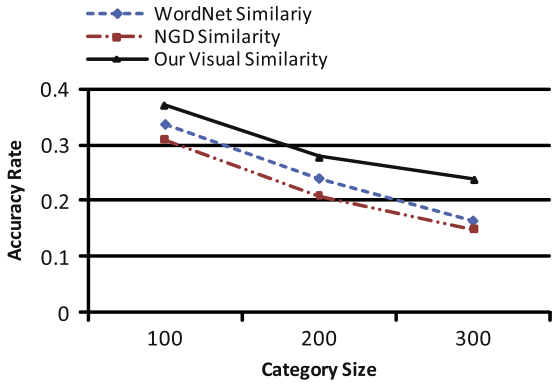


Fig. 12. Average classification accuracy on 100, 200 and 300 categories of ImageNet dataset: our visual similarity *vs.* other measures.

Table 8
Performance comparison of average *F* scores on some subsets of ImageNet dataset: our method *vs.* other most related approaches.

# concepts	Algorithm				
	One-against-all	Pairwise	JointBoost [18]	Hierarchical [34]	Our method
100	0.3000	0.3488	0.1640	0.2059	**0.3670**
200	0.1934	0.2532	0.0771	0.1362	**0.2593**
300	0.1338	0.2054	0.0543	0.1011	**0.2167**

Table 9
Performance comparison of average *F* scores on some subsets of ImageNet dataset: our visual similarity *vs.* other measures.

# concepts	Similarity measurement		
	WordNet similarity	NGD similarity	Our visual similarity
100	0.3310	0.3075	**0.3670**
200	0.2395	0.2076	**0.2593**
300	0.1654	0.1509	**0.2167**

6. Conclusions and future work

A structural learning algorithm is developed in this paper to train a large number of inter-related classifiers jointly for supporting large-scale image classification and annotation. A visual concept network is constructed for characterizing the inter-concept visual correlations intuitively and determining the inter-related learning tasks directly in the visual feature space rather than in the label space. For large-scale image classification, our experimental results have demonstrated that our structural learning algorithm can significantly outperform both the traditional approaches (i.e., the pairwise method and one-against-all method) and other inter-related approaches (i.e., JointBoost[18] and Hierarchical [34] methods).

Although our structural learning algorithm has provided very positive results on large-scale image classification tasks, we should clearly see that it still has some shortcomings. Firstly, in order to simplify the problem of weight determination for classifier combination, we simply use the average weights in our algorithm. However, developing new algorithms for estimating the optimal weights automatically will significantly enhance the performance of our structural learning algorithm and it will be one of future research directions. Secondly, the performance of our proposed learning algorithm is limited to the utilization of

Table 11
Performance comparison on Caltech256 dataset: our hybrid method *vs.* the tree-like combination.

Structure of classifiers	Average accuracy rate
Tree-like structure	0.1788
Our hybrid structural	**0.2120**

Table 12
Performance comparison on Caltech256 dataset: our visual similarity measurement *vs.* other semantic measurements.

Similarity measurement	Performance measure	
	Average accuracy rate	Average *F* score
WordNet similarity [25]	0.1803	0.1932
Google Distance similarity [26]	0.1743	0.1874
Our visual similarity	**0.2120**	**0.2094**

Table 10
Performance comparison on Caltech256 dataset: our method *vs.* other most related approaches.

Algorithms	Performance measure	
	Average accuracy rate	Average *F* score
One-against-all	0.1947	0.1821
Pairwise	0.1977	0.1933
JointBoost (Torralba et al. [18])	0.0559	0.0845
Hierarchical method (Marszalek et al. [34])	0.1446	0.1662
Our Structural learning algorithm	**0.2120**	**0.2094**

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

single SIFT features and the simple VQ coding method. With more comprehensive features and more accurate algorithms for feature coding, our structural learning algorithm will still have better performance than other approaches. In our future work, we will consider some comprehensive features and more accurate feature coding methods to improve the accuracy rates of our structural learning algorithm for large-scale image classification.

Acknowledgments

This work is supported by National Basic Research Program of China (973 Program, Grant Nos. 2012CB316400), National Natural Science Foundation of China (NSFC, Grant Nos. 60905007 and 61272285), and Doctoral Program of Higher Education of China (Grant Nos. 20126101110022).

References

[1] Y. Rui, T.S. Huang, S.F. Chang, Image retrieval: current techniques, promising directions, and open issues, Journal of Visual Communication and Image Representation 10 (1) (1999) 39–62.
[2] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[3] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 5:1–5:60.
[4] K. Bunte, M. Biehl, M. Jonkman, N. Petkov, Learning effective color features for content based image retrieval in dermatology, Pattern Recognition 44 (9) (2011) 1892–1902.
[5] J. Vargas, M. Ferrer, C. Travieso, J. Alonso, Off-line signature verification based on grey level information using texture features, Pattern Recognition 44 (2) (2011) 375–385.
[6] D. Vizireanu, Generalizations of binary morphological shape decomposition, Journal of Electronic Imaging 16 (1) (2007) 1–6.
[7] D. Vizireanu, Morphological shape decomposition interframe interpolation method, Journal of Electronic Imaging 17 (1) (2008) 1–5.
[8] D. Vizireanu, S. Halunga, G. Marghescu, Morphological skeleton decomposition interframe interpolation method, Journal of Electronic Imaging 19 (2) (2010) 1–3.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[10] L. Xu, P. Mordohai, Automatic facial expression recognition using bags of motion words, in: BMVC'10, 2010, pp. 1–13.
[11] K. Polat, S. Gune, A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Systems with Applications 36 (2) (2009) 1587–1592.
[12] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, T. Huang, Large-scale image classification: fast feature extraction and svm training, in: CVPR'11, IEEE, 2011, pp. 1689–1696.
[13] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: CVPR'05, IEEE, 2005, pp. 524–531.
[14] E. Sudderth, A. Torralba, W. Freeman, A. Willsky, Learning hierarchical models of scenes, objects, and parts, in: ICCV'05, organization IEEE, 2005, pp. 1331–1338.
[15] J. Luo, A.E. Savakis, A. Singhal, A Bayesian network-based framework for semantic image understanding, Pattern Recognition 38 (6) (2005) 919–934.
[16] T. Evgeniou, C.A. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, Journal of Machine Learning Research 6 (1) (2005) 615–637.
[17] J. Fan, Y. Gao, H. Luo, R. Jain, Mining multilevel image semantics via hierarchical classification, IEEE Transactions on Multimedia 10 (2) (2008) 167–187.
[18] A. Torralba, K. Murphy, W. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, in: CVPR'04, IEEE, 2004, pp. 762–769.
[19] J. Fan, Y. Gao, H. Luo, Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation, IEEE Transactions on Image Processing 17 (3) (2008) 407–426.
[20] R. Fergus, H. Bernal, Y. Weiss, A. Torralba, Semantic label sharing for learning with many categories, in: ECCV'10, 2010, pp. 762–775.
[21] Y. Jin, L. Khan, B. Prabhakaran, Knowledge based image annotation refinement, Journal of Signal Processing Systems 58 (3) (2010) 387–406.
[22] M. Marszałek, C. Schmid, Semantic hierarchies for visual object recognition, in: CVPR'07, IEEE, 2007, pp. 1–7.
[23] Y. Wang, S. Gong, Refining image annotation using contextual relations between words, in: CIVR'06, ACM, 2006, pp. 425–432.
[24] G. Miller, Wordnet: a lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.
[25] C. Fellbaum, Wordnet, in: Theory and Applications of Ontology: Computer Applications, 2010, pp. 231–243.
[26] R. Cilibrasi, P. Vitanyi, The google similarity distance, IEEE Transactions on Knowledge and Data Engineering 19 (3) (2007) 370–383.
[27] D. Tsujinishi, Y. Koshiba, S. Abe, Why pairwise is better than one-against-all or all-at-once, in: IJCNN'04, IEEE, 2004.
[28] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes, Pattern Recognition 44 (8) (2011) 1761–1776.
[29] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks 13 (2) (2002) 415–425.
[30] X. Yu, Y. Aloimonos, Attribute-based transfer learning for object categorization with zero/one training example, in: ECCV'10, 2010, pp. 127–140.
[31] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, in: ICCV'01, IEEE, 2001, pp. 408–415.
[32] N. Vasconcelos, Image indexing with mixture hierarchies, in: CVPR'01, IEEE, 2001, pp. 3–10.
[33] J. Li, J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1075–1088.
[34] M. Marszałek, C. Schmid, Constructing category hierarchies for visual recognition, in: ECCV'08, Springer, 2008, pp. 479–491.
[35] J. Fan, H. Luo, M. Hacid, Mining images on semantics via statistical learning, in: ACM SIGKDD'05, ACM, 2005, pp. 22–31.
[36] J. Deng, A. Berg, K. Li, L. Fei-Fei, What does classifying more than 10,000 image categories tell us?, in: ECCV'10, 2010, pp. 71–84.
[37] S. Bengio, J. Weston, D. Grangier, Label embedding trees for large multi-class tasks, in: NIPS'10, 2010, pp. 163–171.
[38] A. Tousch, S. Herbin, J. Audibert, Semantic hierarchies for image annotation: a survey, Pattern Recognition 45 (1) (2011) 333–345.
[39] J. Fan, Y. Shen, C. Yang, N. Zhou, Structured max-margin learning for inter-related classifier training and multilabel image annotation, IEEE Transactions on Image Processing 20 (3) (2011) 837–854.
[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: CVPR'09, IEEE, 2009, pp. 248–255.
[41] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: CVPR'06, IEEE, 2006, pp. 2169–2178.
[42] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, 2000.
[43] K. Goh, E. Chang, K. Cheng, SVM binary classifier ensembles for image classification, in: CIKM'01, ACM, 2001, pp. 395–402.
[44] G.D. Guo, A.K. Jain, W.Y. Ma, H.J. Zhang, Learning similarity measure for natural image retrieval with relevance feedback, IEEE Transactions on Neural Networks 13 (4) (2002) 811–820.
[45] J. Li, N. Allinson, D. Tao, X. Li, Multitraining support vector machine for image retrieval, IEEE Transactions on Image Processing 15 (11) (2006) 3597–3601.
[46] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.
[47] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
[48] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Technical Report 7694, California Institute of Technology, 2007. http://authors.library.caltech.edu/7694.
[49] T. Simpson, M. Crowe, Wordnet.net, 2005. http://opensource.ebswift.com/WordNet.Net.
[50] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, http://www.vlfeat.org/, 2008.

Peixiang Dong received the B.S. degree in Electronic Engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. He is currently working toward the Ph.D. degree in Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. His research interests include computer vision and large scale image classification.

Kuizhi Mei received the B.E. and M.S. degrees in electronics engineering and the Ph.D. degree in pattern recognition and intelligence systems in 1999, 2002, and 2006, respectively, Xi'an Jiaotong University, Xi'an, China. He is currently an Associate Professor of the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include image analysis and embedded vision computing.

Nanning Zheng graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975, and received the M.S. degree in information and control engineering from Xi'an Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985.

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029

He jointed Xi'an Jiaotong University in 1975, and he is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, image processing, neural networks, and hardware implementation of intelligent systems.

Dr. Zheng became a member of the Chinese Academy of Engineering in 1999. He is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition.

Hao Lei received the B.S. degree in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. He is currently pursuing the Ph.D degree in Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. His research interests include computer vision, image/video analysis and representation.

Jianping Fan received the M.S. degree in theory physics from Northwest University, Xi'an, China, in 1994, and the Ph.D degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997.

He was a postdoctoral researcher with Fudan University, Shanghai, China, during 1998. From 1998 to 1999, he was a researcher with the Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a postdoctoral researcher with the Department of Computer Science, Purdue University, West Lafayette, IN. He is currently a professor in the Department of Computer Science, University of North Carolina at Charlotte. From 2012, he is also a professor at Northwest University in China. His research interests include image/video analysis, semantic image/video classification, personalized image/video recommendation, surveillance videos, and statistical machine learning.

Please cite this article as: P. Dong, et al., Training inter-related classifiers for automatic image classification and annotation, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2012.10.029