Mining Discriminative Co-occurrence Patterns for Visual Recognition

Junsong Yuan
School of EEE
Nanyang Technological University
Singapore 639798
jsyuan@ntu.edu.sg

Ming Yang
Dept. of Media Analytics
NEC Laboratories America
Cupertino, CA, 95014 USA
myang@sv.nec-labs.com

Ying Wu
EECS Dept.
Northwestern University
Evanston, IL, 60208 USA
yingwu@eecs.northwestern.edu

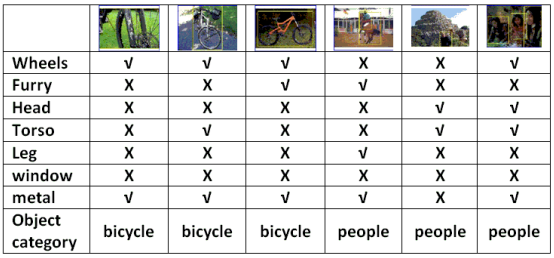
Abstract

The co-occurrence pattern, a combination of binary or local features, is more discriminative than individual features and has shown its advantages in object, scene, and action recognition. We discuss two types of co-occurrence patterns that are complementary to each other, the conjunction (AND) and disjunction (OR) of binary features. The necessary condition of identifying discriminative co-occurrence patterns is firstly provided. Then we propose a novel data mining method to efficiently discover the optimal co-occurrence pattern with minimum empirical error, despite the noisy training dataset. This mining procedure of AND and OR patterns is readily integrated to boosting, which improves the generalization ability over the conventional boosting decision trees and boosting decision stumps. Our versatile experiments on object, scene, and action categorization validate the advantages of the discovered discriminative co-occurrence patterns.

1. Introduction

Due to the compositional property of visual objects, scenes, and actions, the discovery of discriminative co-occurrence pattern is of fundamental importance in recognizing them. Although the extracted features, such as color, texture, shape, or motion features, can be quite weak individually, an appropriate combination of them will bring a strong feature which is much more discriminative [31] [29] [3] [16] [2] [9]. There has been a recent trend in mining co-occurrence patterns for visual recognition. For example, every real-world object is associated with numerous visual attributes in terms of its material, structure, shape, etc, [11] [4] [10], although it is difficult to differentiate them using a single visual attribute, they can be well distinguished by the co-occurrence of specific attributes, as illustrated in Fig. 1.

In a binary classification problem, given a collection of N binary features, the problem of co-occurrence pattern mining is to select a subset from these N features, such that the co-occurrence of them can best discriminate the two classes. In spite of many previous works in mining and integrating co-occurrence patterns [22] [20] [3] [29] [33] [16] [24] [9] [27], none of these methods is targeted at finding the most discriminative co-occurrence pattern with the smallest classification error. Given N binary features, because the co-occurrence pattern can contain an arbitrary number of features (up to N), the total number of candidates of co-occurrence patterns is exponentially large (*e.g.* 3^N if considering the negative value or 2^N if only considering the positive value.) As a result, it is computationally intractable to perform an exhaustive search. Even worse, unlike conventional feature selection task, the monotonic property of feature subset does not hold in searching co-occurrence patterns. Namely, a $(K+1)$-order binary feature is not necessarily better than a K-order one. Therefore, the branch-and-bound search cannot be applied directly [18]. Existing approaches for co-occurrence pattern search, such as sequential forward selection [16] [24], or recent data mining-driven approaches [22] [20] [3] [29] [33], do not guarantee the optimality of the selected co-occurrence patterns. In general, when the training data is noisy, it is still an open problem to find the co-occurrence pattern of the best performance, *e.g.*, minimum classification error [17].



Wheels	√	√	√	X	X	√
Furry	X	X	√	√	X	X
Head	X	X	X	X	√	√
Torso	X	√	X	X	√	√
Leg	X	X	X	√	X	X
window	X	X	X	X	X	X
metal	√	√	√	√	X	√
Object category	bicycle	bicycle	bicycle	people	people	people

Figure 1. By inferring binary visual attributes from raw image features, such as wheels, furry [6] [4], we can distinguish bikes from people by a co-occurrence of certain attributes, for example a bike has *metal* and *wheels*, but does not have *head*. Given two classes of objects, described by (possibly quite noisy) attributes, can we efficiently discover the co-occurrence of attributes that can best discriminate them ?

Besides the computational issue, it is difficult to combine co-occurrence patterns appropriately. When the target class exhibits a multi-mode distribution in the feature space, *i.e.* the intra-class variation is large, a single co-occurrence pattern is not enough to cover the positive training samples. Thus multiple co-occurrence patterns must be considered. Most previous works integrate co-occurrence patterns through a boosting procedure: boosting high-order features rather than individual ones [29] [13] [33]. However, all of these works only consider one type of co-occurrence pattern, namely the conjunction form (AND), while the disjunction form (OR) is neglected. As these two types of classifiers are complementary to each other [34], the OR pattern should also be considered.

To address the above issues, we propose an efficient data mining-based approach to discovering discriminative co-occurrence patterns and integrating them to a boosting classifier. Our contributions are two-fold: (1) in terms of mining co-occurrence patterns, the necessary conditions of discriminative patterns are obtained and the *optimal* co-occurrence pattern of minimum empirical error can be discovered efficiently from the noisy training data; (2) in terms of boosting co-occurrence patterns, we expand the pool of weak learners by considering both AND and OR patterns and incorporate them through a multi-class Adaboost. It improves conventional boosting decision stumps and boosting decision trees. The versatile experiments on the PASCAL VOC 2008 dataset, 15-scene dataset, and KTH action dataset validate the effectiveness and efficiency of our method.

2. Related Work

Because co-occurrence patterns are more discriminative than individual features, they have been extensively applied in classification tasks, such as the feature co-occurrence [16], multi-local feature [2], compositional feature [29] [30], high-order feature [13], and visual grouplet [27]. In [24] [2] [16], co-occurrence local features are applied for object categorization and detection. Because of the complexity in searching co-occurrence features, only the second-order feature is considered in [13]. To handle the huge search space, frequent itemset mining is applied [31] [32] [29] [22] [33] for mining co-occurrence patterns. Despite many previous work, however, few of them carefully studied the optimality of the co-occurrence patterns from a theoretical perspective, but ad-hoc methods were usually applied to find co-occurrence patterns to avoid the exponential cost of mining. Thus, these methods cannot guarantee the optimality of the mined co-occurrence patterns.

As each co-occurrence pattern serves as a classification rule, co-occurrence pattern mining is also related to rule induction in machine learning and data mining. Some conventional methods, such as the version space approach and the candidate-elimination algorithm, normally require noise-free training data for efficient rule induction [17]. When a perfect rule (*i.e.* a co-occurrence pattern) with zero training error does not exist, these approaches cannot work well. It remains an open problem to efficiently find the discriminative co-occurrence pattern from noisy training data [19].

3. Discriminative Co-occurrence Patterns

3.1. Basic definitions and formulation

We consider a 2-class problem for discriminative analysis. The training dataset contains N samples of two classes: $\mathcal{D}_N = \{\mathbf{x}_t, c_t\}_{t=1}^N$, where $\mathbf{x}_t \in \mathbb{R}^P$ denotes the feature vector and $c_t \in \{0, 1\}$ is the label of $\mathbf{x}_t$. We define an *attribute* as a Boolean-valued function, $\mathbf{f}(\cdot) : \mathbf{x} \rightarrow \{0, 1\}$.

Such a binary feature of $\mathbf{x}$ can be semantic if defined as the visual properties of objects. For example, some recent works introduced such attributes as color, texture, shape to describe visual objects [11] [4] [6] and faces [10], where $\mathbf{x}$ represents an object and $\mathbf{f}_i(\mathbf{x}) \in \{0, 1\}$ indicates whether the i_{th} attribute is active or not, *e.g.*, the object is furry or not. Different object categories can share the same vocabulary of the pre-defined attributes.

In addition, a binary feature $\mathbf{f}$ can also be non-semantic if it is induced from $\mathbf{x}$ by a simple classifier. Taking the *decision stump* for example:

$$\mathbf{f}_j(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}(j) \geq \theta_j \\ 0 & \text{if } \mathbf{x}(j) < \theta_j \end{cases}, \tag{1}$$

where $\mathbf{x}(j)$ is the j_{th} element in $\mathbf{x}$ and $\theta_j \in \mathbb{R}$ is the threshold to determine the response of $\mathbf{f}$. The induced features do not necessarily have semantic meanings, yet they are still informative to describe $\mathbf{x}$.

Considering a set of Boolean-valued features $\mathcal{A} = \{\mathbf{f}_i(\mathbf{x})\}_{i=1}^{|\mathcal{A}|}$, a *co-occurrence pattern* corresponds to a subset $\mathcal{B} \subseteq \mathcal{A}$ of features:

$$\mathcal{F}(\mathbf{x}) = \begin{cases} \bigwedge_{i \in \mathcal{B}} \mathbf{f}_i(\mathbf{x}) & \text{conjunction} \\ \bigvee_{i \in \mathcal{B}} \mathbf{f}_i(\mathbf{x}) & \text{disjunction} \end{cases}. \tag{2}$$

Now the co-occurrence pattern $\mathcal{F}(\cdot) : \mathbf{x} \rightarrow \{0, 1\}$ serves as a classifier to distinguish the two classes. It can contain an arbitrary number (up to $|\mathcal{A}|$) of features $\mathbf{f}$ (or $\bar{\mathbf{f}}$).

For simplicity, in Eq. (2), we call the conjunction as the *AND* pattern, denoted by $\mathcal{F}_A$, and the disjunction as the *OR* pattern, denoted by $\mathcal{F}_O$. Our target co-occurrence pattern is the one with the minimum empirical error:

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \epsilon_{\mathcal{F}}, \tag{3}$$

where $\mathcal{F}$ is either an AND or OR pattern, and $\epsilon_{\mathcal{F}}$ is the empirical error to measure the discriminative ability of $\mathcal{F}$ on the 2-class problem:

$$\epsilon_{\mathcal{F}} = P(\mathcal{F}(\mathbf{x}) \neq c(\mathbf{x}) | \mathcal{D}_N),$$

where $\mathcal{F}(\mathbf{x})$ is the binary prediction of $\mathbf{x}$ and $c(\mathbf{x})$ is the ground truth; and $\mathcal{D}_N$ is the training dataset.

Because each co-occurrence pattern $\mathcal{F}$ is uniquely determined by the selected subset $\mathcal{B} \subseteq \mathcal{A}$, all of the candidates $\mathcal{B}$ form a powerset of size $2^{|\mathcal{A}|}$ or $3^{|\mathcal{A}|}$, depending on whether the negative values are considered. For example, if the negative value $\bar{\mathbf{f}}_i$ is considered in Eq. (2), then each attribute $\mathbf{f}_i \in \mathcal{A}$ has 3 possible status in $\mathcal{F}$: $\bar{\mathbf{f}}_i$, $\mathbf{f}_i$ and null. Thus the total number of candidates is $3^{|\mathcal{A}|}$. As a result, it is intractable to search for $\mathcal{F}^*$ exhaustively if $|\mathcal{A}|$ is large.

Before discussing how to find the optimal $\mathcal{F}^*$ through data mining in Sec. 4, we first explain the duality between the AND and OR patterns in the next subsection.

3.2. Duality between AND and OR patterns

If the negative responses are considered, the total number of OR combinations is also $3^{|\mathcal{A}|}$. According to the De Morgan's law, we have the duality between the AND and OR:

$$\overline{\mathcal{F}_O} = \overline{\vee_{i\in\mathcal{B}}\, \mathbf{f}_i(\mathbf{x})} = \wedge_{i\in\mathcal{B}}\, \overline{\mathbf{f}_i(\mathbf{x})} = \mathcal{F}_A. \tag{4}$$

Therefore, an OR pattern, $\mathcal{F}_O$, can be transformed to an AND pattern, $\mathcal{F}_O$, by inversing its prediction and the attribute values. This duality leads to a uniformed way to find the discriminative co-occurrence patterns.

Remark 1 Duality between AND and OR Patterns
An OR pattern that predicts for the positive class is equivalent to the AND pattern that predicts for the negative class.

4. Efficient Mining of Co-occurrence Patterns

4.1. Necessary conditions of optimal $\mathcal{F}^*$

Due to the duality between AND and OR patterns, we only discuss how to find an optimal AND $\mathcal{F}_A^*$. The search of $\mathcal{F}_O^*$ follows the same strategy. To explain how to perform an efficient search, we first discuss the requirements for discriminative patterns. We denote the *frequency* of a pattern $\mathcal{F}$ by:

$$P(\mathcal{F}) = \frac{frq(\mathcal{F})}{N} = \frac{|\{t : \mathcal{F}(\mathbf{x}_t) = 1\}|}{N}, \tag{5}$$

where N is the total number of samples. $P(c = 1|\mathcal{F})$ is the precision rate and $P(\mathcal{F}|c = 1)$ is the recall rate. For a perfect pattern $\mathcal{F}^*$, we have $P(c = 1|\mathcal{F}) = P(\mathcal{F}|c = 1) = 1$, thus the empirical error $\epsilon_{\mathcal{F}} = 0$.

In the case of noisy training data, a perfect pattern may not exist. To find the optimal $\mathcal{F}^*$ with the smallest $\epsilon_{\mathcal{F}}$, we establish the necessary conditions of a discriminative $\mathcal{F}$. Lemma 1 states the mild-frequency requirement. A pattern of high frequency is likely to appear in both positive and negative samples, thus leads to a low precision rate. On the other hand, a pattern of low frequency cannot cover the whole positive class and thus leads to a low recall rate. Both of them are not discriminative. To complement Lemma 1, Lemma 2 states the recall requirement. Clearly, a classifier of a low recall cannot be discriminative.

Lemma 1 the mild-frequency requirement for a discriminative $\mathcal{F}$
For any $\mathcal{F}$ of small error $\epsilon_{\mathcal{F}} \leq \hat{\epsilon}$, where $0 \leq \hat{\epsilon} \leq r^+$, and $r^+ = P(c = 1)$, it must satisfy the following mild frequency requirement:

$$r^+ - \hat{\epsilon} \leq P(\mathcal{F}) \leq r^+ + \hat{\epsilon}.$$

Proof: suppose $P(\mathcal{F}) > r^+ + \hat{\epsilon}$, then because $P(\mathcal{F} = 1, c = 1) \leq P(c = 1) = r^+$, we have $P(\mathcal{F} = 0, c = 1) = P(\mathcal{F}) - P(\mathcal{F} = 1, c = 1) > r^+ + \hat{\epsilon} - r^+ = \hat{\epsilon}$. Therefore the error is at least $P(\mathcal{F} \neq c) = P(\mathcal{F} = 0, c = 1) + P(\mathcal{F} = 1, c = 0) \geq P(\mathcal{F} = 0, c = 1) > \hat{\epsilon}$. On the other hand, if $P(\mathcal{F}) < r^+ - \hat{\epsilon}$, then it is easy to show that $P(\mathcal{F} \neq c) \geq P(\mathcal{F} = 1, c = 0) > \hat{\epsilon}$.

Lemma 2 the recall requirement for a discriminative $\mathcal{F}$
For any co-occurrence pattern $\mathcal{F}$ of small error $\epsilon_{\mathcal{F}} \leq \hat{\epsilon}$, where $0 \leq \hat{\epsilon} \leq r^+$, and $r^+ = P(c = 1)$, it must satisfy the following recall requirement:

$$P(\mathcal{F}|c = 1) \quad \geq \quad 1 - \frac{\hat{\epsilon}}{r^+}$$

Proof: suppose $P(\mathcal{F} = 1|c = 1) < 1 - \frac{\hat{\epsilon}}{r^+}$, then $P(\mathcal{F} = 0|c = 1) > \frac{\hat{\epsilon}}{r^+}$ and $P(\mathcal{F} = 0, c = 1) = P(\mathcal{F} = 0|c = 1)P(c = 1) > \hat{\epsilon}$. Thus we have $P(\mathcal{F} \neq c) \geq P(\mathcal{F} = 0, c = 1) > \hat{\epsilon}$.

Combining Lemma 1 and Lemma 2, if the optimal $\mathcal{F}^*$ satisfies $\epsilon_{\mathcal{F}^*} \leq \hat{\epsilon}$ in the training data $\mathcal{D}_N$, where $0 \leq \hat{\epsilon} \leq r^+$, then $\mathcal{F}^*$ must meet both requirements and be included in the following candidate set:

$$\mathcal{F}^* \quad \in \quad \{\mathcal{F} : r^+ - \hat{\epsilon} \leq P(\mathcal{F}) \leq r^+ + \hat{\epsilon}\} \bigcap \{\mathcal{F} : P(\mathcal{F}|c = 1) \geq 1 - \frac{\hat{\epsilon}}{r^+}\}.$$

As a result, we have the following theorem.

Theorem 1 a necessary condition of optimal $\mathcal{F}^*$
For a co-occurrence pattern $\mathcal{F}$ to predict the positive class, suppose its empirical error satisfies $\epsilon_{\mathcal{F}} \leq \hat{\epsilon}$, where $0 \leq \hat{\epsilon} \leq r^+$ and $r^+ = P(c = 1)$. Let $\mathbf{\Psi}_1 = \{\mathcal{F} : P(\mathcal{F}) \geq r^+ + \hat{\epsilon}\}$, $\mathbf{\Psi}_2 = \{\mathcal{F} : P(\mathcal{F}) \geq r^+ - \hat{\epsilon}\}$, $\mathbf{\Psi}_3 = \{\mathcal{F} : P(\mathcal{F}|c = 1) \geq 1 - \frac{\hat{\epsilon}}{r^+}\}$, then the optimal $\mathcal{F}^*$ must reside in the candidate set:

$$\mathcal{F}^* \in (\mathbf{\Psi}_2 \backslash \mathbf{\Psi}_1) \cap \mathbf{\Psi}_3, \tag{6}$$

where $\mathbf{\Psi}_2 \backslash \mathbf{\Psi}_1 = \{\mathcal{F} : \mathcal{F} \in \mathbf{\Psi}_2, \mathcal{F} \notin \mathbf{\Psi}_1\}$.

According to Theorem 1, since $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$ are two sets of frequent patterns for the entire training samples, while $\mathbf{\Psi}_3$ is the set of frequent patterns for positive training samples, all of them have a small size and can be efficiently obtained through frequent pattern mining. As a result, although the full search space of $\mathcal{F}^*$ is exponentially large, the above candidate set of $\mathcal{F}^*$ is much smaller and can be exhaustively checked.

Algorithm 1: Mining Optimal AND pattern

input : Training dataset $\mathcal{D} = \{\mathcal{D}^+, \mathcal{D}^-\}$, minimum error $\hat{\epsilon}$
output : $\mathcal{F}_A^* = \arg\min_{\mathcal{F} \in \mathcal{F}_A} \epsilon_{\mathcal{F}}$

Mining frequent patterns from $\mathcal{D}$:
$\mathbf{\Psi}_1 = \{\mathcal{F} : P(\mathcal{F}) \geq r^+ + \hat{\epsilon}\}$,
$\mathbf{\Psi}_2 = \{\mathcal{F} : P(\mathcal{F}) \geq r^+ - \hat{\epsilon}\}$.
Mining frequent patterns from $\mathcal{D}^+$:
$\mathbf{\Psi}_3 = \{\mathcal{F} : P(\mathcal{F}|c=1) \geq 1 - \frac{\hat{\epsilon}}{r^+}\}$
let $\mathbf{\Psi}_A = (\mathbf{\Psi}_2 \backslash \mathbf{\Psi}_1) \cap \mathbf{\Psi}_3$
return $\mathcal{F}_A^* = \arg\min_{\mathcal{F} \in \mathbf{\Psi}_A} \epsilon_{\mathcal{F}}$

4.2. Algorithm implementation

Our algorithm is designed based on Eq. (6). We present the search of discriminative AND patterns for a positive class in Alg. 1. First of all, frequent patterns are discovered from the whole training dataset ($\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$). Then frequent patterns of the positive class are discovered ($\mathbf{\Psi}_3$). Finally, we perform an exhaustive check of the candidate set $(\mathbf{\Psi}_2 \backslash \mathbf{\Psi}_1) \cap \mathbf{\Psi}_3$ to find $\mathcal{F}^*$. It is worth noting that for multi-class problems, different classes can share the same $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ as they are discovered from the whole dataset. Therefore, we only need to search for the frequent pattern $\mathbf{\Psi}_3$ for each individual class.

To avoid the exhaustive search of all possible combinatorial patterns, classic frequent pattern mining methods apply a branch-and-bound search. By using the bounds, they either apply a breath first search (Apriori algorithm) or a depth first search (FP-growth algorithm) to overcome the exponential complexity in the search [8]. Although the worst case complexity can still be exponential, its average complexity is mild if it is properly designed. In order to obtain $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$ and $\mathbf{\Psi}_3$, we apply the FP-growth algorithm in [7] for *closed* frequent itemset mining. Closed frequent itemsets are compact representations of frequent patterns. They have been recently applied in computer vision literature for visual pattern and feature mining [29] [20] [22] [33].

For searching the OR pattern in the *positive* class, according to the duality between AND and OR, we can target on the AND pattern in the *negative* class instead. Due to the space limit, we omit the mining procedure of the best OR pattern, which follows the same procedure as mining the AND pattern. The only differences are that $\mathbf{\Psi}_3$ is a frequent pattern set from the negative class and we need to replace r^+ with r^- in Alg. 1.

5. Integration of AND and OR Patterns

The single $\mathcal{F}^*$ discovered via data mining yields the minimum empirical error, nevertheless, it may not be a good classifier individually. For example, if the target class has a multi-mode distribution, rather than relying on a single $\mathcal{F}^*$ (*e.g.* a decision node), it is desirable to have a set of co-occurrence patterns (*e.g.* a decision list) to cover all of the training samples. Moreover, AND and OR patterns are complementary to each other: AND patterns generally bring high precisions but with low recall rates, while OR patterns bring high recalls by scarifying the precision. Therefore, incorporating both of them in learning a classifier is expected to result in a balanced precision and recall.

Algorithm 2: Mining AND/OR patterns for boosting

input : training dataset $\mathcal{D} = \{\mathcal{D}^+, \mathcal{D}^-\}$, a pool of weak learners $\mathbf{\Omega} = \{\mathbf{f}_i\}$, # of iterations M
output : a binary classifier, $\mathbf{g}(\cdot) : \mathbf{x} \rightarrow \{0, 1\}$

Init: set the training sample weights $w_i = 1/N$, $i = 1, ..., N$.
for $m = 1, 2, ..., M$ **do**
  $\mathcal{F}^m = \arg\min_{\mathbf{f} \in \mathbf{\Omega}} \sum_{i=1}^N w_i \mathbb{I}(c_i \neq \mathbf{f}(\mathbf{x}_i))$,
  if *training error decreases slowly* **then**
    mining AND and OR candidates:
    $\mathbf{\Psi} = \mathbf{\Psi}_A \cup \mathbf{\Psi}_O$ (using Alg. 1 and its variant)
    $\mathcal{F}^m = \arg\min_{\mathcal{F} \in \mathbf{\Psi}} \sum_{i=1}^N w_i \mathbb{I}(c_i \neq \mathcal{F}^m(\mathbf{x}_i))$
  Compute weighted training error:
    $err^m = \frac{\sum_{i=1}^N w_i \mathbb{I}(c_i \neq \mathcal{F}^m(\mathbf{x}_i))}{\sum_{i=1}^N w_i}$.
  Compute: $\alpha^m = \log \frac{1 - err^m}{err^m}$.
  Update weight: $w_i \leftarrow w_i \cdot \exp[\alpha^m \mathbb{I}(c_i \neq \mathcal{F}^m(\mathbf{x}_i))]$.
  Re-normalize w_i.
Return $\mathbf{g}(\mathbf{x}) = \arg\max_k \sum_{m=1}^M \alpha^m \cdot \mathbb{I}(\mathcal{F}^m(\mathbf{x}) = k)$

Both AND and OR patterns can be naturally integrated through the boosting procedure. The new algorithm is presented in Alg. 2, where we follow the standard Adaboost algorithm for binary classification. The threshold of a single decision stump is determined by the conventional boosting procedure, *i.e.*, given the training samples and their weights at current step, we search the optimal threshold for a decision stump to minimize the weighted training error. The only difference is when a single decision stump cannot effectively decrease the training error. In such a case, it implies that a decision stump is too weak to distinguish the hard samples and over-fitting may occur. Thus, we apply the proposed data mining method to discover more discriminative high-order AND or OR patterns to help. By applying such a strategy, we can achieve the demanded performance faster with fewer weak learners. Moreover, as less complex decision stumps are used in the initial rounds of boosting, practically it does not tend to over-fit the dataset. Compared with previous method [13] of boosting first-order and second-order features, we do not constrain the order of the AND/OR. It relies on data mining method to find good co-occurrences.

6. Experiments

To demonstrate the advantages of discriminative co-occurrence patterns, we apply the proposed method to 3

Table 1. Selected results of the pair-wise discriminative analysis. The top 5 examples are the pairs that are difficult to discriminate based on the attributes, *i.e.*, with largest testing errors. The next 10 examples are the pairs that can be easily distinguished. '+' denotes that attribute appears; '-' denotes it does not. We highlight the testing error if it is not larger than training error.

objects	train/ test err	discriminative AND
bottle v.s. pottedplant	0.490 / **0.473**	-Occluded
cat v.s. dog	0.418 / 0.467	-Occluded+Tail+Head +Ear+Foot/Shoe
dog v.s. cat	0.427 / 0.439	+Leg
chair v.s. pottedplant	0.308 / 0.431	+Occluded
chair v.s. bottle	0.320 / 0.420	+Occluded
train v.s. person	0.000 / 0.003	-Occluded-Head-Ear -Eye-Torso-Leg -Foot/Shoe+Metal
bicycle v.s. person	0.003 / **0.003**	-Head-Ear-Eye-Torso -Leg-Foot/Shoe+Metal
horse v.s. person	0.004 / **0.004**	+Furry
cow v.s. person	0.003 / 0.004	+Furry
aeroplane v.s. person	0.002 / 0.005	-Head-Ear-Eye-Torso -Leg-Foot/Shoe+Metal
motorbike v.s. person	0.002 / 0.006	-Head-Ear-Eye-Torso-Leg-Foot/Shoe+Metal
cat v.s. person	0.002 / 0.006	+Furry
dog v.s. person	0.005 / 0.007	+Furry
pottedplant v.s. sheep	0.029 / **0.009**	-Tail-Head-Ear -Snout-Eye-Torso -Leg-Foot/Shoe
pottedplant v.s. train	0.000 / 0.013	-3D Boxy-Window -Wheel-Door-Headlight -Taillight-Exhaust -Metal-Shiny

different tasks. Specifically, we conduct discriminative object category analysis on the PASCAL VOC 2008 dataset, scene recognition on the 15-scene dataset, and action recognition on the KTH dataset. All of these experiments validate the effectiveness and efficiency of mining discriminative AND/OR patterns.

6.1. Discriminative analysis between object categories

To evaluate the discriminative ability of the discovered AND/OR classifier, we test our method on an attribute dataset provided in [4]. With the purpose to describe, compare, and categorize objects, this dataset provides 64 attribute labels for the PASCAL VOC 2008 trainval set of roughly 12,000 images in 20 categories. Learned from the raw image features, each image is described by 64 binary visual attributes. Following its experimental setting, in total of 6340 images are used for training and another 6355 images are used for testing.

Given two categories of objects, our task is to discover discriminative co-occurrence attributes for classification. For each of the 20 categories, we compare it to the rest 19 categories for discriminative analysis. In total we have 380 individual pairs of object categories. For each pair of objects, we treat one object as the positive class and the other as the negative class.The minimum error is set to $\hat{\epsilon} = 0.3$. If the error of the optimal AND/OR is lower than 0.3, it ensures to find the optimal AND/OR combination. Given a pair of objects, we discard the attributes that never appear in both classes and perform data mining on the rest of attributes. An optimal AND classifier from the training samples is discovered for each pair, and the corresponding testing error is presented in Fig. 2. Each element $e(i, j)$ of the matrix in Fig. 2 is the testing error, which evaluates the AND pattern that predicts for the i_{th} category (the positive class). Similarly, $e(j, i)$ is the testing error of the AND for the j_{th} category. Due to the duality between AND and OR, $e(j, i)$ is also the testing error of the OR pattern predicting for the i_{th} category. As a result, the pair-wise analysis is not symmetric since $e(i, j) \neq e(j, i)$.

Among the 380 pair-wise analysis, only 5 pairs can find a perfect discriminative co-occurrence pattern with zero training error. It validates that the training data is noisy. This is not surprising because the learned attributes are not perfect. The mean training error among the 380 pairs is 0.0814, while the mean testing error is 0.1126. This result validates the good generalization ability of the AND/OR classifiers.

In Table 1, we list the top 5 most difficult pairs, as well as the top 10 easiest pairs, ranked by their testing errors. We notice that unreliable attributes (based on the measurement in [4]), such as 'occluded', appear more often in the difficult pairs. On the other hand, reliable attributes appear more often in easy pairs, such as 'metal' and 'furry'. Although we do not provide the classification results of the 20 categories, the pair-wise discriminative analysis provides a guidance of objects that can be easily confused. For example, to better distinguish cat and dog, bottle and potted plant, more reliable attributes should be introduced.

6.2. Boosting co-occurrence patterns for scene recognition

We also evaluate the effectiveness of the AND and OR co-occurrence patterns for scene recognition on the 15-scene category dataset and improve the state-of-the-art results. The 15 scene category dataset was collected gradually by several research groups [21, 5, 12] and it consists of a variety of indoor and outdoor scenes: *bedroom*, *livingroom*, *suburb*, *industrial*, *kitchen*, *coast*, *forest*, *highway*,

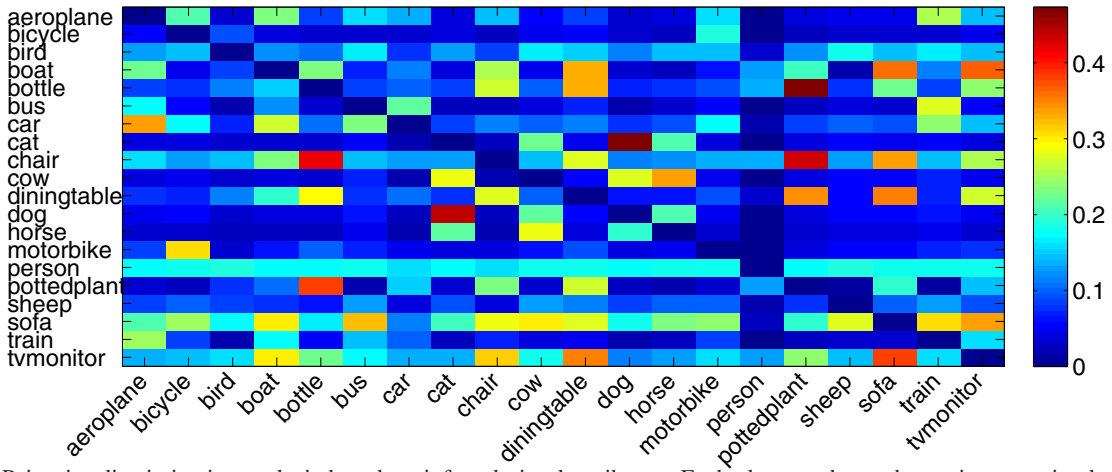


Figure 2. Pair-wise discriminative analysis based on inferred visual attributes. Each element shows the testing error in classifying two object categories: $e(i,j)$ and $e(j,i)$ are the testing errors of the OR and AND patterns, respectively, for the j_{th} category.

inside city, *mountain*, *open country*, *street*, *tall building*, *office*, and *store*. Each scene category includes 216 to 410 images with resolution around 300×250.

We investigate two recent features for scene recognition: (1) a holistic feature–CENTRIST [25], and (2) a local descriptor based feature–linear coordinate coding (LCC) [28]. For the CENTRIST feature, we follow the standard setting in [25]. Each image is partitioned into 25 blocks (level 2), 5 blocks (level 1), and 1 block (level 0), respectively. After using the principal component analysis to reduce the dimensionality of CENTRIST to 40, each scene image results in a 1240-dimensional feature. For the LCC feature, we calculate the dense SIFT [15] features every 8 pixels with 4 patch sizes, *i.e.*, 7×7, 16×16, 25×25, and 31×31, to learn a 4096-dimensional codebook using clustering, which is used in the linear coordinate coding to encode the textural characteristics of images. Following the idea of the spatial pyramid matching (SPM) [12], we partition an image to 10 cells, *i.e.*, 1×1 and 3×3, to delineate the spatial layout, where the LCC codes of each cell are concatenated. Thus, each scene image is represented by a 40960-dimensional feature. Combining both CENTRIST and LCC features, each image has in total 40960 + 1240 = 42200 features.

In the first experiment, we discover the AND/OR features only from the LCC codes to train the boosting classifiers (Alg. 2), and compare it with the linear SVM classifiers and the boosting classifiers with single features. Following the same test protocol in [12, 25], we use 100 images in each category for training and the rest of the images for testing. The proposed algorithm achieves the average recognition accuracy 83.7%, which improves over the boosting with single features by about 1.7%. The improvement boils down to the capacity of the compositional patterns to delineate more sophisticated decision boundaries than merely using the single features. As shown in Table 1, the boosting with co-occurrence AND/OR patterns from LCC codes achieves comparable performance to the state-of-the-art methods.

We further explore the AND/OR patterns from both the CENTRIST and LCC codes, and train the boosting classifiers using Alg. 2. Now an AND/OR feature can be a combination of both CENTRIST and LCC features, thus is likely to be more discriminative than individual features. The combination of these two complimentary features shows excellent description power for the scene images. As shown in Table 2, the boosting with co-occurrence AND/OR patterns further improves the state-of-the-art results, from 83.9% to 87.8%. Comparing to the boosting of individual features from the CENTRIST+LCC pool, we also observe a 1.9% improvement when boositng AND/OR features, from 85.9% to 87.8%. This further validates the effectiveness of boosting higher order features. The confusion matrix of the proposed method is presented in Fig. 3.

The training time of our method is determined by the number of training samples, the feature dimensionality, and the mining step of the AND and OR patterns. As the feature dimensionality is 42200 and the number of training samples is 1500, it costs around 4-5 seconds to mine one co-occurrence pattern, on a laptop with a CPU Core 2 Duo 2.6GHz. Considering both AND and OR patterns, it takes about 9-10 seconds to select one composite feature. However, the classification is very efficient just like the conventional boosting classifiers.

6.3. Boosting co-occurrence patterns for action recognition

We apply the proposed boosting algorithm with co-occurrence patterns (Alg. 2) to the action recognition task

Method	Avg. accuracy
SPM + SIFT with 400 clusters [12]	81.4%
SPM + SIFT with 400 concepts [14]	83.3%
SP-pLSA + SIFT with 1200 topics [1]	83.7%
CENTRIST+ RBF SVM [25]	83.9%
LCC+Linear SVM	80.7%
LCC+Boosting	82.0%
LCC+Boosting (AND/OR)	83.7%
CENTRIST+LCC+Boosting	**85.9%**
CENTRIST+LCC+Boosting (AND/OR)	**87.8%**

Table 2. Comparison of the average recognition accuracy on the 15-scene category dataset.

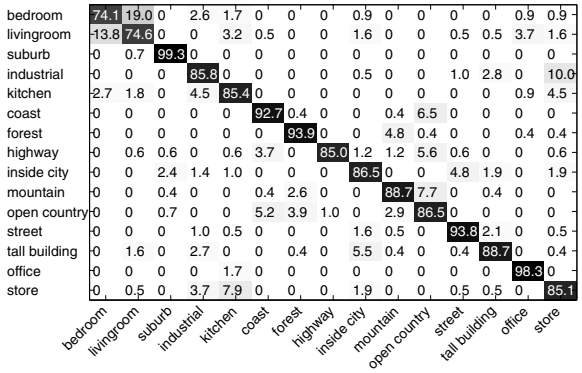


Figure 3. The confusion matrix of 15-scene recognition.

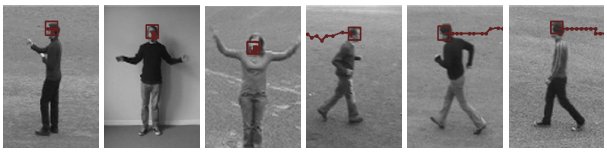


Figure 4. Sample frames of 6 action categories in the KTH dataset.

on the benchmark KTH dataset. The KTH dataset was first recorded for [23] and includes 6 types of actions: *box*, *clap*, *wave*, *jog*, *run*, and *walk*. There are 25 subjects performing these actions under 4 different scenes: outdoors (s1), outdoor with scale variations (s2), outdoors with different clothes (s3), and indoors (s4). In total, there are 2391 sequences with image resolution 160×120. Sample frames are illustrated in Fig. 4.

We apply the same features in [26] to recognize the actions. The candidate regions are first located by human detection and tracking. For each detected human, an enlarged region around the tracked head is cropped in the so called motion edge history images (MEHI) [26]. Then, a large number of 2D Haar features are extracted to train classifiers for each action category. We perform a 5-fold cross-validation to evaluate the performance, where the sequences of 20 persons are used in training and those of the other 5 persons for testing.

Our approach is compared against two methods: (1) a

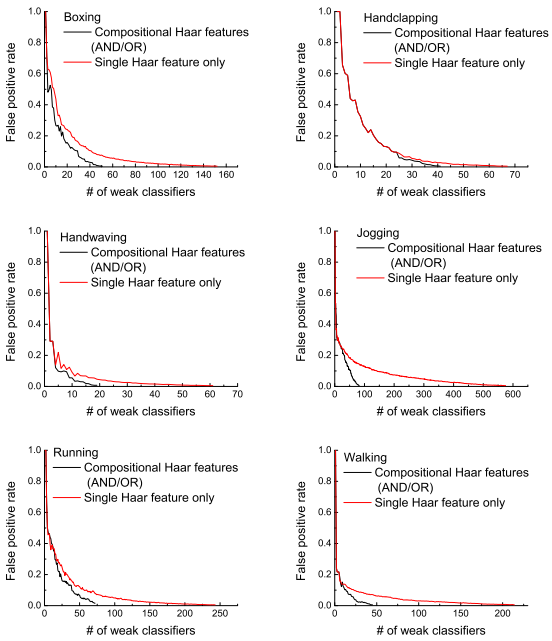


Figure 5. For the s1 scenario, false positive rates in training the boosting classifiers using single Haar features only *v.s.* using the compositional Haar features as well.

boosting classifier using decision stumps of single Haar features (denoted by *B-stumps*); and (2) a boosted 4-node decision tree classifier based on Haar features (denoted by *B-trees*). In our method, we treat the Haar features as weak learners and discover the AND and OR patterns from binary Haar features. During the boosting procedure, when the false positive rate decreases less than $1 \times e^{-3}$ by adding a single Haar feature, we switch to add AND/OR patterns. We specify the desired detection rate to be 0.99 and false positive rate 0.005 in training, which can derive the minimum error ε in Alg. 1. The size of $\mathbf{\Psi}$ in Alg. 2 is around several thousands. We observe that the false positive rates drop faster when compositional features are employed, as shown in Fig. 5. For an example, for *jog* if only using single Haar features, the training requires a selection of 575 weak classifiers to reach a false positive rate of 0.005. In contrast, when using the co-occurrence features, it only needs 82 Haar features. Even though there may be multiple Haar features in an ANR/OR pattern, the number of total Haar features is actually fewer, as shown in Table 3. Therefore, it brings less computation at the testing stage. In the boosted decision tree classifier, we enforce it to have the same total number of Haar features as our method for fair comparison. To test the generalization ability, we perform a 5-fold cross validation for each scene and the results are listed in Table 4. By incorporating the compositional AND and OR features, the average testing accuracy is improved by about 3% over the method in [26].

# of Haar features	box	clap	wave	jog	run	walk
# of Haar (AND/OR)	71	46	24	113	105	76
# of Haar (dec. stump)	152	67	61	575	243	214

Table 3. The comparison of the number of Haar features: boosting decision stumps + AND/OR v.s. boosting decision stumps only.

Scene	B-stumps	B-trees	Ours
s1	73.9%	73.5%	*77.03%*
s2	71.0%	70.3%	*73.98%*
s3	73.6%	73.1%	*77.48%*
s4	78.9%	79.3%	*80.80%*
Scene	B-stumps	B-trees	Ours
s1	83.7%	85.3%	*87.83%*
s2	84.4%	85.5%	*87.05%*
s3	82.6%	84.5%	*86.89%*
s4	92.4%	93.6%	*94.36%*

Table 4. The average recognition accuracy of a 5-fold cross-validation on the KTH dataset: per-frame results (up); per-video segment results (bottom).

7. Conclusions

We present a data mining approach to discovering discriminative co-occurrence patterns for visual recognition. The complementary AND and OR patterns are elaborated, as well as the derivation of the necessary condition in identifying discriminative patterns (both AND and OR). Based on the necessary condition, the proposed data mining based method is capable to efficiently find the optimal co-occurrence pattern with the minimum empirical error, despite the exponentially large search space and the noisy training data. The versatile experiments on object, scene, and action recognition validate the advantages of the discovered AND and OR patterns.

Acknowledgment

This work was supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan, National Science Foundation grant IIS-0347877, IIS-0916607, and US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE PAMI*, 30(4):712–727, 2008. 2783

[2] O. Danielsson, S. Carlsson, and J. Sullivan. Automatic learning and extraction of multi-local features. In *Proc. IEEE Intl. Conf. on Computer Vision*, pages 917 – 924, 2009. 2777, 2778

[3] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2777

[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009. 2777, 2778, 2781

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, San Diego, CA, June21 - 23 2005. 2781

[6] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proc. of Neural Information Processing Systems*, 2007. 2777, 2778

[7] G. Grahne and J. Zhu. Fast algorithms for frequent itemset mining using FP-trees. *IEEE Transaction on Knowledge and Data Engineering*, 2005. 2780

[8] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. In *Data Mining and Knowledge Discovery*, 2007. 2780

[9] S. Ito and S. Kubota. Object classification using heterogeneous co-occurrence features. In *Proc. European Conf. on Computer Vision*, 2010. 2777

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE Intl. Conf. on Computer Vision*, pages 365 – 372, 2009. 2777, 2778

[11] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 2777, 2778

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2169 – 2178, 2006. 2781, 2782, 2783

[13] D. Liu, G. Hua, , P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2778, 2780

[14] J. Liu and M. Shah. Scene modeling using co-clustering. In *Proc. IEEE Intl. Conf. on Computer Vision*, Rio de Janeiro, Oct.14 - 21 2007. 2783

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004. 2782

[16] T. Mita, T. Kaneko, B. Stenger, and O. Hori. Discriminative feature co-occurrence selection for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(7):1257–1269, 2008. 2777, 2778

[17] T. Mitchell. *Machine Learning*. McGraw Hill, 1997. 2777, 2778

[18] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computer*, 26(9):917–922, 1977. 2777

[19] P. K. Novak, N. Lavrac, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. In *Journal of Machine Learning Research*, volume 10, pages 377–403, 2009. 2778

[20] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. Bakir. Weighted substructure mining for image analysis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2777, 2780

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int'l Journal of Computer Vision*, 42(3):145–175, 2001. 2781

[22] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool. Efficient mining of frequent and distinctive feature configurations. In *Proc. IEEE Intl. Conf. on Computer Vision*, pages 1–8, 2007. 2777, 2778, 2780

[23] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR'04*, volume 3, pages 32–36, Cambridge, UK, Aug. 23-26, 2004. 2783

[24] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):854–869, 2007. 2777, 2778

[25] J. Wu and J. M. Rehg. CENTRIST: a visual descriptor for scene categorization. *IEEE PAMI*, 2010. 2782, 2783

[26] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong. Human action detection by boosting efficient motion features. In *IEEE Workshop on Video-oriented Object and Event Classification in Conjunction with ICCV*, pages 522 – 529, Kyoto, Japan, Sept.29 - Oct.2, 2009. 2783

[27] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. 2777, 2778

[28] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS'09*, 2009. 2782

[29] J. Yuan, J. Luo, and Y. Wu. Mining compositional features for boosting. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2777, 2778, 2780

[30] J. Yuan, J. Luo, and Y. Wu. Mining compositional features from GPS and visual cues for event recognition in photo collections. *IEEE Trans. on Multimedia*, 12(7):705–716, 2010. 2778

[31] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2777, 2778

[32] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proc. ACM SIGKDD*, pages 864–873, 2007. 2778

[33] B. Zhang, G. Ye, Y. Wang, J. Xu, and G. Herman. Finding shareable informative patterns and optimal coding matrix for multiclass boosting. In *Proc. IEEE Intl. Conf. on Computer Vision*, pages 56 – 63, 2009. 2777, 2778, 2780

[34] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. 2778