Discovery of Collocation Patterns: from Visual Words to Visual Phrases

Junsong Yuan, Ying Wu, Ming Yang
EECS Department, Northwestern University
2145 Sheridan Road, Evanston, IL, USA 60208
{j-yuan, yingwu, m-yang4}@northwestern.edu

Abstract

A visual word lexicon can be constructed by clustering primitive visual features, and a visual object can be described by a set of visual words. Such a "bag-of-words" representation has led to many significant results in various vision tasks including object recognition and categorization. However, in practice, the clustering of primitive visual features tends to result in synonymous visual words that over-represent visual patterns, as well as polysemous visual words that bring large uncertainties and ambiguities in the representation. This paper aims at generating a higher-level lexicon, i.e.visual phrase lexicon*, where a visual phrase is a meaningful spatially co-occurrent pattern of visual words. This higher-level lexicon is much less ambiguous than the lower-level one. The contributions of this paper include: (1) a fast and principled solution to the discovery of significant spatial co-occurrent patterns using frequent itemset mining; (2) a pattern summarization method that deals with the compositional uncertainties in visual phrases; and (3) a top-down refinement scheme of the visual word lexicon by feeding back discovered phrases to tune the similarity measure through metric learning.*

1. Introduction

The success of data mining and information retrieval techniques in structured data (*e.g.*, transaction data) and semi-structured data (*e.g.*, text) has recently aroused our curiosity in applying them to many computer vision tasks including object retrieval [19], discovery [17, 16, 21], categorization [22] and recognition [3]. Once we can extract some visual primitives such as interest points [13] or regions [15] that highlight the local image invariants, and treat their labels (*e.g.*, the codewords for quantization) as " visual words", an image can be represented by a "bag of words". This *visual lexical representation*, as an analogical treatment of texts, may allow the leverage of the research of text data mining in vision.

Although such ideas appear to be quite exciting, the leap from text data that are semi-structured to images that are non-structured is not trivial, because text data are discrete and have much less ambiguities of semantical meanings, while visual data are continuous and generally exhibit much larger variabilities and uncertainties. The same visual pattern, no matter how local it is, is likely to exhibit quite different visual appearances under different lighting conditions, views, scales, not to mention partial occlusion. Thus, it is very difficult, if not impossible, to find invariant visual features that are insensitive to these variations to uniquely characterize visual patterns. Although a discrete *visual word lexicon* (VWL) of a finite collection of visual words may be forcefully obtained by clustering those primitive visual features (*e.g.*, by vector quantization or K-means clustering), such visual words tend to be much more ambiguous than texts. Specifically, the ambiguity lies in two aspects: *synonymy* and *polysemy*. A synonymous visual word shares the same semantic meanings with other visual words. Because the corresponding underlying semantics is split and represented by multiple visual words, synonymy leads to over-representations. On the other hand, a polysemous visual word may mean different things under different contexts. Thus polysemy leads to under-representations. Both phenomena have impeded the promising leap, and the root of this impedient is the large uncertainties within non-structured visual data. Therefore, it is crucial to address the uncertainty issues.

One possible solution to resolve the ambiguity of polysemous visual words may be to put them into a spatial context. In other words, the *collocation* (or *co-occurrence*) of several visual words is likely to be much less ambiguous. Therefore, it is of great interest to automatically discover these collocation visual patterns. Some recent work studied this issue by simply finding frequent (or repetitive) co-occurrent visual words [19, 10, 12]. Although this is a good starting point of image data mining, there are many challenges that need to be overcome:

- **Spatial dependency in image data**. To discover collocation patterns, a first step is to construct a database where each record is a *word group* located in a local

spatial neighborhood. However, these records are not independent as they have spatial overlaps in images. This phenomenon largely complicates the data mining process, because simply counting the occurrence frequencies is doubtable and a frequent pattern is not necessarily a meaningful pattern. Thus special care needs to be taken;

- **Synonymy in collocation patterns**. The collocation patterns inherit synonymy in the visual word lexicon. This is largely reflected by the compositional variations in the collocation patterns. In other words, a semantically-coherent collocation pattern may be split into different patterns, which can be caused by partial occlusion of the pattern, miss detection of the visual primitives *etc*. This creates a big obstacle when moving to higher-level patterns.
- **Ambiguities in visual word lexicon**. If the generation of the collocation patterns is a purely bottom-up process, then the imperfectness in the visual word lexicon will never be reduced, and the quantization error in VWL will never be corrected. This issue has never been addressed before.

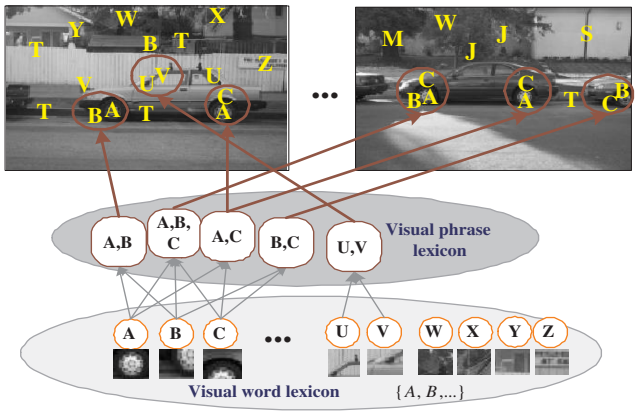


Figure 1. Discovery of collocation patterns: each visual phrase is composed of meaningful co-located visual words. For example, after discovering the car category database, we find $\{A, B\}$, $\{A, B, C\}$, $\{A, C\}$ and $\{B, C\}$ are four *synonymous* visual phrases all associated with wheels, while $\{U, V\}$ is the phrase associated with windows.

This paper presents a novel solution for discovering meaningful word-collocation patterns, *i.e.visual phrases*, and constructing a higher-level *visual phrase lexicon* (VPL) together with a set of semantically-coherent visual phrase patterns from the atomic level lexicon (*i.e*. VWL here). The concepts are illustrated in Fig. 1. By addressing the above three difficulties, our contributions are three-folds:

- *new criteria in repetitive pattern discovery*. The co-occurrence frequency is no longer a sufficient condition for discovering meaningful collocation patterns. Due to the spatial dependency of the visual data, a more plausible likelihood ratio test method is proposed to evaluate the significance of a visual word-set. In addition, to conquer the complexity in searching for the meaningful patterns (the total number of possible word-sets is exponential to the cardinality of VWL), we develop an improved frequent itemset mining (FIM) algorithm based on the new criteria so as to discover significant visual phrases in a very efficient way;
- *pattern summarization*. To handle the compositional uncertainties of visual phrases, a novel pattern summarization method is proposed to further cluster these "synonymous" visual phrases into semantically-coherent patterns;
- *top-down refinement of VWL*. To reduce the ambiguities in VWL, a top-down refinement is proposed by taking advantage of the discovered visual phrase patterns. They serve as supervision to tune the metric in the feature space of visual primitives for better visual word clustering.

2. Overview and Basic Concepts

We follow the notations in [19]. For each image in the database, we first detect the visual primitives a_i (*e.g*. SIFT-like features) and obtain a *visual word lexicon* (VWL) $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = M$) through clustering. We call every item W_k in the lexicon $\mathbf{\Omega} = \{W_1, ..., W_M\}$ as a *visual word* (or *word* for short). Then an image can be represented as a "bag of words": $\mathcal{I} = \{a_i\}$.

For each word $a_i \in \mathcal{I}$, we further define its local spatial neighborhood (*e.g*. K-nearest neighborhood) as a *group of words* $\mathcal{G}_i = \{a_i, a_{i_1}, a_{i_2}, \cdots, a_{i_K}\}$. The image database $\mathbf{D}_\mathcal{I} = \{\mathcal{I}_t\}_{t=1}^T$ generates a collection of such groups to form a *group database* $\mathbf{G} = \{\mathcal{G}_i\}_{i=1}^N$, which contains a collection of N groups with M attributes of visual words. Similar to the data mining scenario, this is a *transaction database* with N records, where each record $\mathcal{G}_i$ simply indicates which words are included. This database can also be represented by a sparse binary matrix $X_{N\times M}$, where the entry $x_{ij} = 1$ indicates the i_{th} group contains the j_{th} word in VWL and $x_{ij} = 0$ otherwise.

Once we have the group database $\mathbf{G}$, a basic task is to discover the frequent word collocations from the induced transaction database. We define a *visual word-set* (also called itemset in data mining literature) by a set of visual words $\mathcal{P} \subset \Omega$. For a given word-set $\mathcal{P}$, the record $\mathcal{G}_i$ which includes $\mathcal{P}$ is called an *occurrence* of $\mathcal{P}$. Namely $\mathcal{G}_i$ is an occurrence of $\mathcal{P}$, if $\mathcal{P} \subseteq \mathcal{G}_i$. We denote $\mathbf{G}(\mathcal{P})$ as the set of all the occurrences of $\mathcal{P}$ in $\mathbf{G}$, and the *frequency* of a word-set $\mathcal{P}$ is denoted by:

$$frq(\mathcal{P}) = |\mathbf{G}(\mathcal{P})| = |\{i : \forall j \in \mathcal{P}, x_{ij} = 1\}|. \quad (1)$$

A word-set $\mathcal{P}$ is called *frequent* if $frq(\mathcal{P}) \geq \theta$, where the threshold θ is called the *minimum support*. Finding frequent word collocation patterns is closely related to the finding of frequent itemsets in this transaction database.

Finding frequent itemsets in transaction database (*i.e.*, frequent itemset mining or FIM) has been widely studied in data mining literature [8]. Given a transaction dataset, the task of FIM is to discover all the frequent itemsets (frequent word-set in our case) $\mathcal{P}_j \subseteq \mathbf{\Omega}$ such that $frq(\mathcal{P}_j) \geq \theta$. Because the possible itemsets are combinatorial, it is impossible to do an exhaustive check, but there exist FIM algorithms that are extremely efficient, such as the Aprior and the FP-growth algorithm. In this paper we apply the FP-growth algorithm [7] to implement the FIM. As the FP-tree has a prefix-tree structure and can store compressed information of frequent itemsets, it can quickly discover all the frequent sets from group dataset $\mathbf{G}$, without miss detection.

Based on the VWL, this paper presents a new solution to generate a higher-level lexicon of word-collocation patterns called *visual phrase lexicon* or VPL, as illustrated in Fig. 2.

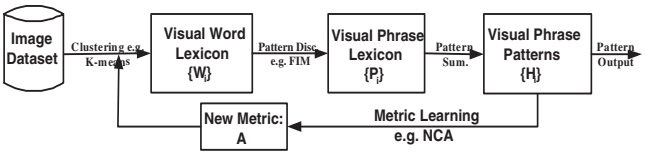


Figure 2. Framework overview for visual phrase discovery.

In Sec. 3, we present the fast and principled method to discover the meaningful word-sets and build VPL. In Sec. 4, we propose a pattern summarization method to further cluster those related phrases into phrase patterns in order to handle the compositional variations. In Sec. 5, a top-down self-supervision method is proposed by applying the discovered phrase patterns as feedback to further train a better similarity metric for representing visual primitives. Then new visual word lexicon is obtained through clustering by using the learned new metric.

3. Discovering Visual Phrase Lexicon

Given an image dataset $\mathbf{D}_{\mathcal{I}}$, the task is to discover the meaningful word-set $\mathcal{P} \subset \mathbf{\Omega}$ ($|\mathcal{P}| \geq 2$), and to build the *visual phrase lexicon* $\mathbf{\Psi} = \{\mathcal{P}_j\}$, where each visual phrase $\mathcal{P}_j$ represents a meaningful word-set.

Simply checking the occurrence frequency in $\mathbf{G}$ is far from sufficient, because of three difficulties: (1) the dependency among the records in $\mathbf{G}$, (2) redundant high-order word-sets, and (3) meaningless frequent word-sets. We will analyze and propose our solutions to these issues shortly, after we introduce a measure of the statistical significance of the frequent co-occurrence.

In order to quantify the statistical significance of a frequently co-occurrent word-set $\mathcal{P}$, we need to compare the likelihood that $\mathcal{P}$ is generated by the hidden pattern versus the likelihood that $\mathcal{P}$ is randomly generated, *i.e.* by chance. More formally, we compute the following likelihood ratio for $\mathcal{P} = \{W_i\}$ based on the two hypotheses, where H_0: occurrences of $\mathcal{P}$ are randomly generated, and H_1: occurrences of $\mathcal{P}$ are generated by the hidden pattern.

$$L(\mathcal{P}) = \frac{P(\mathcal{P}|H_1)}{P(\mathcal{P}|H_0)} = \frac{\sum_{i=1}^{N} P(\mathcal{P}|\mathcal{G}_i, H_1)P(\mathcal{G}_i|H_1)}{\prod_{i=1}^{|\mathcal{P}|} P(W_i|H_0)}, \quad (2)$$

where $P(\mathcal{G}_i|H_1) = \frac{1}{N}$ is a constant; $P(\mathcal{P}|\mathcal{G}_i, H_1)$ is the likelihood that $\mathcal{P}$ is generated by a hidden pattern and is observed at a particular group $\mathcal{G}_i$, such that $P(\mathcal{P}|\mathcal{G}_i, H_1) = 1$ if $\mathcal{P} \subseteq \mathcal{G}_i$ and $P(\mathcal{P}|\mathcal{G}_i, H_1) = 0$ otherwise. Consequently, based on Eq. 1, we can calculate $P(\mathcal{P}|H_1) = \frac{frq(\mathcal{P})}{N}$. We also assume that the words $W_i \in \mathcal{P}$ are conditionally independent under the null hypothesis H_0, and $P(W_i|H_0)$ is the prior of word $W_i \in \mathbf{\Omega}$, *i.e.* the total number of visual primitives that are labeled with W_i in image database $\mathbf{D}_{\mathcal{I}}$. We thus refer $L(\mathcal{P})$ as the "significance" score to measure the importance of a word-set $\mathcal{P}$. In fact if $\mathcal{P} = \{W_A, W_B\}$ is a second-order word-set, then $L(\mathcal{P})$ is the mutual information criterion to test the pair-wise dependency. In addition, to assure that a visual word-set $\mathcal{P}$ is meaningful, we also require it to appear repetitively enough in the database, *i.e.* $frq(\mathcal{P}) \geq \theta$, such that we can avoid those phrases that appear rarely but happen to have strong spatial dependency among the words. With these criteria, we need to overcome the following three issues before the visual phrases can be mined.

◊ Frequency over-counting of word-sets

Different from transaction database where each record is independent of each other, every group $\mathcal{G}_i$ has spatial overlap with its neighborhood groups, therefore their contents are dependent on one another. It can cause over-counting when calculating $frq(\mathcal{P})$ from Eq. 1 directly, *i.e.* $frq(\mathcal{P})$ can have duplicate counts, which is illustrated in Fig. 3.

Figure 3. Frequency over-counting caused by the spatial overlap of groups. The word-set {A,B} is counted twice by groups $\mathcal{G}_1 = \{A, B, D\}$ and $\mathcal{G}_2 = \{A, B, C, E\}$, although it has only one instance in the image. There is only one pair of A and B that co-occur in a local region such that $d(A, B) < 2r$, with r the radius of $\mathcal{G}_1$. In the texture region where visual primitives are densely sampled, such over-count can largely exaggerate the number of occurrences for a texton pattern.

In order to address the group dependency problem, we apply a two-phase mining scheme. First, without considering the spatial overlap problem, we apply FIM based on Eq. 1 to obtain a candidate set of frequent phrases. For these candidates $\{\mathcal{P}_i : frq(\mathcal{P}_i) \geq \theta\}$, we re-count the number of their real instances exhaustively through the original image database, not allowing duplicate counts. The computational cost is largely saved by taking such a two-phase mining. After FIM, we only need to re-count for a small candidate set, whose size is much smaller than that of the original candidate set ($2^{|\mathbf{\Omega}|}$). Without causing confusion, we denote $\hat{frq}(\mathcal{P})$ as the real instance number of word-set $\mathcal{P}$. Accordingly, we adjust the calculation of $P(\mathcal{P}|H_1) = \frac{\hat{frq}(\mathcal{P})}{\hat{N}}$, where $\hat{N} = N/K$ denotes the approximated independent group number $\hat{N}$, where K is the cardinality of the spatial neighborhood.

◊ **Redundant high-order word-sets.**
If a word-set $\mathcal{P}$ appears frequently, then all of its sub-sets $\mathcal{P}' \subset \mathcal{P}$ will also appear frequently, *i.e.* $frq(\mathcal{P}) \geq \theta \Rightarrow frq(\mathcal{P}') \geq \theta$. Thus high-order word-sets bring redundancy in the visual phrase lexicon. For instance, a frequent word-set $\mathcal{P}$ composed with n words can generate 2^n sub-sets which are all frequent word-sets.

To control this redundancy from the high-order frequent word-set, we apply closed FIM algorithms to discover *closed frequent itemsets* [8] instead of *frequent itemsets*. The number of closed frequent word-sets can be much less than the frequent word-sets. Formally, we call a word-set $\mathcal{P}$ is a *closed word-set* if there does *not* exist a word-set $\mathcal{Q}$ such that (1) $\mathcal{P} \subset \mathcal{Q}$ and (2) $\forall\, \mathcal{G}_i, \mathcal{P} \subseteq \mathcal{G}_i \Rightarrow \mathcal{Q} \subseteq \mathcal{G}_i$, *i.e.* $\mathbf{G}(\mathcal{P}) = \mathbf{G}(\mathcal{Q})$. Note the closed frequent word-sets compress information of frequent word-sets in a lossless form, *i.e.* the full list of frequent word-sets $\mathbf{F} = \{\mathcal{P}_i\}$ and their corresponding frequency counts can be exactly recovered from the compressed representation. Thus this guarantees that no meaningful word-sets will be left out.

Furthermore, for a closed high-order word-set $\mathcal{P}_i (|\mathcal{P}_i| > 2)$, we perform the *Student t-test* for each pair of its words $t(\{W_i, W_j\}), \forall i, j \in \mathcal{P}$, as in Eq. 3:

$$t(\{A,B\}) = \frac{P(\{A,B\}) - \mu_x}{\sqrt{\frac{S^2}{\hat{N}}}} \quad (3)$$

$$= \frac{P(\{A,B\}) - P(A)P(B)}{\sqrt{\frac{P(\{A,B\})(1-P(\{A,B\}))}{\hat{N}}}} \quad (4)$$

$$\approx \frac{\hat{frq}(\{A,B\}) - \frac{1}{\hat{N}}\hat{frq}(A)\hat{frq}(B)}{\sqrt{\hat{frq}(\{A,B\})}}, \quad (5)$$

where μ_x is the mean of Gaussian distribution x and S^2 is the estimated variance of x from the observation data. The high-order word-set $\mathcal{P}_i$ is possibly meaningful only if all of its pairwise subsets can pass the test individually.

◊ **Meaningless frequent visual phrases.**
After closed FIM and t-test for further filtering the redundant high-order word-sets, we obtain a collection of frequent word-sets $\mathbf{F} = \{\mathcal{P}_i\}$. However, a frequent word-set $\mathcal{P}_i = \{W_j\}$ is still not necessarily a meaningful pattern, because it is not clear whether the word co-occurrences are statistically significant or just by chance. To justify the frequent word-set, we need to further perform a hypothesis testing in Eq. 2 for each word-set $\mathcal{P}_i \in \mathbf{F}$.

By integrating all the above together, instead of selecting word-sets $\mathcal{P}_i$ simply by their occurrence frequency $\hat{frq}(\mathcal{P})$, we can rank $\mathcal{P}_i \in \mathbf{F}$ based on their likelihood ratio $L(\mathcal{P})$ according to Eq. 2. The top-k most meaningful word-sets with largest likelihood ratio will be selected and thus form the *visual phrase lexicon* $\mathbf{\Psi} = \{\mathcal{P}_j\}_{j=1}^{k}$.

4. Pattern Summarization of Visual Phrases

As discussed before, synonymy exists in VPL $\mathbf{\Psi}$ because of many factors. Suppose there exists a *visual phrase pattern* $\mathcal{H}$ (*e.g.* a visual pattern associated with the common objects) that repetitively generates a lot of instances which are captured by some groups $\mathcal{G}_i$. We can certainly observe such meaningful repetitive patterns from the induced transaction database, *e.g.* discovering meaningful word-sets $\mathcal{P}_i \in \mathbf{\Psi}$ through FIM. However, instead of observing a complete pattern $\mathcal{H}$, we tend to observe many incomplete patterns with compositional variations, *i.e.* many visual phrases $\mathcal{P}_i$ that correspond to the same $\mathcal{H}$ (see Fig. 4). This can be caused by many reasons, including the missing detection of visual primitives, imperfect clustering of visual primitives in constructing VWL, and partial occlusion of the hidden pattern itself.

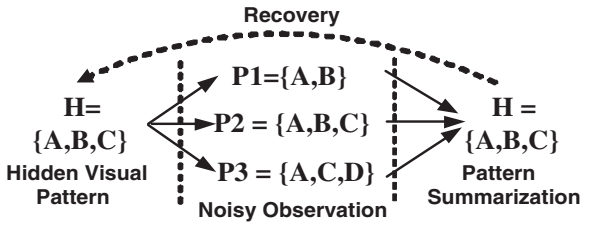


Figure 4. Pattern summarization.

Therefore, we need to cluster those correlated visual phrases $\{\mathcal{P}_i\}$ (incomplete patterns) in order to recover the complete one $\mathcal{H}$. We call this task as **pattern summarization**. The problem can be stated as follows: given a collection of meaningful word-sets, *i.e.* the visual phrase lexicon $\mathbf{\Psi} = \{\mathcal{P}_i\}_{i=1}^{k}$, we want to further cluster the related phrases $\mathcal{P}_i$ into phrase classes, where each class $\mathcal{H}_j = \{\mathcal{P}_i\}_{i=1}^{|\mathcal{H}_j|}$

is defined as a *visual phrase pattern*. Consequently, after phrase summarization, we will obtain a small set of visual phrase patterns from VPL.

Pattern summarization is a difficult task. One reason is that the polysemy in VWL brings many ambiguities in this clustering process. Typically, visual phrases with similar word compositions do not always correspond to the same phrase pattern $\mathcal{H}$. For example, are two phrases $\mathcal{P}_i = \{W_A, W_B\}$ and $\mathcal{P}_j = \{W_A, W_C\}$ always generated from the same visual phrase pattern $\mathcal{H}$? The answer is not necessarily. If $W_A \in \mathbf{\Omega}$ is a polysemous visual word, then these two visual phrases $\mathcal{P}_i$ and $\mathcal{P}_j$ can correspond to different visual phrase patterns, and it is up to the spatial context W_B and W_C to resolve the ambiguity of W_A.

However, if two visual phrases $\mathcal{P}_i$ and $\mathcal{P}_j$ are generated from the same pattern $\mathcal{H}$, then their group sets $\mathbf{G}(\mathcal{P}_i)$ and $\mathbf{G}(\mathcal{P}_j)$ (Eq. 1) should have a large overlap. As a result, the similarity between two visual phrases $s_{ij} = S(\mathcal{P}_i, \mathcal{P}_j), \forall i, j \in \mathbf{\Psi}$ should not be only based on their frequencies $\hat{frq}(\mathcal{P}_i)$ and $\hat{frq}(\mathcal{P}_j)$, but also the correlation between their group set $\mathbf{G}(\mathcal{P}_i)$ and $\mathbf{G}(\mathcal{P}_j)$ that support these two phrases. In order to measure the pair-wise similarity between visual phrases, we apply the "profile-based pattern representation" [24] to address the VWL polysemy in our pattern summarization. According to [24], the affinity between a pair of visual phrases can be defined by the KL-divergency between their group profiles. Once this affinity matrix is obtained, we use the normalized cut algorithm [18] for clustering the visual phrases.

5. Top-Down Refinement of VWL

By discovering VPL $\mathbf{\Psi}$ and summarizing it, we obtain a small set of meaningful visual phrase patterns $\mathbf{H} = \{\mathcal{H}_i\}$. Compared with visual words, these visual phrase patterns have much less ambiguities because they are semantically coherent. The discovered instances of the phrase patterns can thus serve as supervision to retrieve other instances of the same patterns in the database.

Based on VPL $\mathbf{\Psi}$, we can partition VWL $\mathbf{\Omega}$ into two unjoined subsets, $\mathbf{\Omega} = \mathbf{\Omega}^+ \cup \mathbf{\Omega}^-$, where for any visual phrase $\mathcal{P}_i \in \mathbf{\Psi}$, we have $\mathcal{P}_i \subseteq \mathbf{\Omega}^+$ and $\mathcal{P}_i \nsubseteq \mathbf{\Omega}^-$. Hence we denote $\mathbf{\Omega}^+ = \bigcup_{i=1}^{|\mathbf{\Psi}|} \mathcal{P}_i$ as the *foreground word lexicon*, which is the basis for composing $\mathbf{\Psi}$. Correspondingly we denote $\mathbf{\Omega}^- = \mathbf{\Omega} \backslash \mathbf{\Omega}^+$ as the *background word lexicon*, because a word $W_i \in \mathbf{\Omega}^-$ can not compose any phrase $\mathcal{P}_i \in \mathbf{\Psi}$. In such a case, $W_i \in \mathbf{\Omega}^-$ should be a noisy or redundant word that is not of interests. According to such a partition of $\mathbf{\Omega}$, we thus treat the instances of visual words $W_i \in \mathbf{\Omega}^-$ as negative training samples to represent the background, while the instances of visual phrases $\mathcal{P}_i \in \mathbf{\Psi}$ as positive training samples to represent foreground patterns. These self-labeled training data can then be utilized to train a better feature representation of visual primitives and refine VWL.

Specifically, we apply nearest component analysis (NCA) [6] here to learn a better distance metric other than the commonly used Euclidean distance in clustering visual primitives into VWL. NCA learns a global linear transformation in the feature space to improve the leave-one-out accuracy of the K-NN classifier. NCA is feasible to multiple classes learning and does not assume the distribution of each class is a single Gaussian and thus can be applied in our problem, *e.g.* visual word classes that have mixture-Gaussian or non-Gaussian distributions. To fit the supervised learning method NCA into our unsupervised clustering case, we take those discovered visual primitives that can compose visual phrases as multiple-class positive training samples, while those background words as negative training samples. It is important to note that our top-down refinement takes advantage of the spatial nature of image patterns: those visual primitives are *not* independent in the feature space, as there are 2-D spatial relations among them. By feeding back mined visual patterns as supervision, we tend to cluster those visual primitives into the same class if (1) they have similar features and (2) their local 2-D spatial neighbors also have similar features.

6. Experiments

6.1. Setup

Given a large image dataset $\mathbf{D}_\mathcal{I} = \{\mathcal{I}_i\}$, we first detect the PCA-SIFT points [11] in each image $\mathcal{I}_i$ and treat these interest points as the visual primitives. Each interest point is a 41×41 patch at the given scale, and rotated to align its dominant orientation to a canonical direction. Note multiple visual primitives could be located at the same position, but with different scales or orientations. We select two categories from the Caltech 101 database [4] for the experiments: faces (435 images) and side views of cars (123 images). K-means algorithm is used to cluster these atomic visual features into VWL $\mathbf{\Omega}$. For each category, the support threshold for closed FIM is set to $\theta = \frac{1}{4}|\mathbf{D}_\mathcal{I}|$, where $|\mathbf{D}_\mathcal{I}|$ is the total number of images. We set word lexicon size $|\mathbf{\Omega}| = 160$ and 500 for the car and face category respectively. For generating the group databases, we set $K = 5$ for choosing K-NN groups. All the experiments were conducted on a Pentium-4 3.19GHz machine with 1GB RAM running window xp.

6.2. Evaluation of visual phrase lexicon

To test whether the discovered visual phrases are really meaningful in that they are associated with the frequently appeared foreground objects, we propose the following two criteria for evaluation: (1) the precision of visual phrase lexicon $\mathbf{\Psi}$: ρ^+ represents the percentage of visual phrases $\mathcal{P}_i \in \mathbf{\Psi}$ that are located in the foreground object, and (2) the

precision of background word lexicon $\mathbf{\Omega}^-$: ρ^- represents the percentage of background words $W_i \in \mathbf{\Omega}^-$ that are located in the background. In the ideal case, if $\rho^+ = \rho^- = 1$, then all the visual phrases $\mathcal{P}_i \in \mathbf{\Psi}$ are associated with the common objects, *i.e*. located inside the bounding box of the object, while all the visual words $W_i \in \mathbf{\Omega}^-$ are located in the background, *i.e*. located outside the bounding box. Then we can precisely discriminate the foreground common objects from the clutter background. Furthermore, we use retrieval rate η to denote the percentage of retrieved images that contain at least one visual phrase instance. The larger the η, the more visual phrases we retrieve.

Table 1 shows the results of discovering visual phrases from the car database. The first row indicates the size of the visual phrase lexicon $\mathbf{\Psi}$. It is shown that as the size of $\mathbf{\Psi}$ increases, its precision score ρ^+ decreases (from 1.00 to 0.86), while the percentage of retrieved images η increases (from 0.11 to 0.88). The high precision ρ^+ indicates that most of the discovered visual phrases appear within the foreground objects. It is also noted that foreground words $\mathbf{\Omega}^+$ is only a small subset with respect to $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = 160$), which implies that most visual words actually do not depict the foreground objects. Thus it is meaningful to get rid of those noisy words from the background. Examples of visual phrase lexicons are shown in Fig. 6 and Fig. 7.

Table 1. Precision score ρ^+ and retrieval rate η for the discovered visual phrase lexicon $\mathbf{\Psi}$, corresponding to various phrase lexicon size. See text for descriptions of ρ^+ and η.

$\lvert\mathbf{\Psi}\rvert$	1	5	10	15	20	25	30
$\lvert\mathbf{\Omega}^+\rvert$	2	7	12	15	22	27	29
η	0.11	0.40	0.50	0.62	0.77	0.85	0.88
ρ^+	1.00	0.96	0.97	0.91	0.88	0.86	0.86

We further compare 3 different criteria for selecting meaningful word-sets $\mathcal{P} \in \mathbf{\Psi}$ as visual phrases, against the baseline of selecting the most frequent visual words $W_i \in \mathbf{\Omega}$ (first-order phrase) to build $\mathbf{\Psi}$. These 3 criteria are (1) occurrence frequency: $\hat{frq}(\mathcal{P})$ (2) t-score: $T(\mathcal{P})$ (only select second order phrases, $|\mathcal{P}| = 2$) and (3) likelihood ratio: $L(\mathcal{P})$. Fig. 5 shows the comparison results. It depicts the variations of ρ^+ and ρ^- with increasing phrase lexicon size ($|\mathbf{\Psi}| = 1, ..., 30$). We can see that all three phrase selection criteria perform significantly better than the baseline of choosing the most frequent visual words. This demonstrates that the discovered collocation patterns are more discriminative and informative than the singleton words which normally suffer from synonymy and polysemy problems. It is also shown from Fig. 5 that when selecting a small number of phrases $\mathcal{P}$ into $\mathbf{\Psi}$, all the three criteria yield similar performances. However, when more phrases are added, the proposed likelihood ratio method performs better than the other two. Moreover, the occurrence frequency $\hat{frq}(\mathcal{P})$ always gives the worst performance among the three criteria. This again validates that not all the frequently appeared word-sets are meaningful.

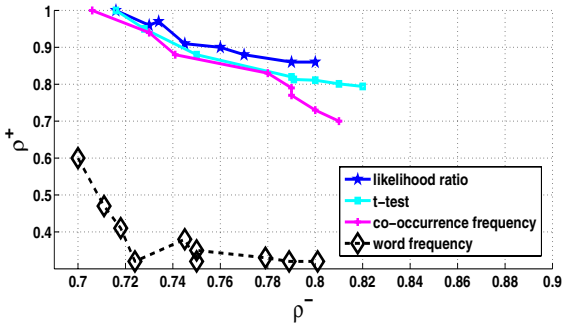


Figure 5. Performance comparison by applying three different visual phrase selection criteria, also with the baseline of selecting most frequent visual words to build $\mathbf{\Psi}$.

By taking advantage of the FP-growth algorithm for closed FIM, our pattern discovery is very efficient. It costs only 27 seconds for discovering VPL from the face database containing more than 60,000 groups (see table 2). It thus provides us a powerful tool to explore large object category database where each image contains hundreds of primitive visual features.

Table 2. CPU computational cost for discovering visual phrase lexicon in face database, with $|\mathbf{\Psi}| = 30$.

# images $\lvert\mathbf{D}_\mathcal{I}\rvert$	# groups $\lvert\mathbf{G}\rvert$	FIM only [7]	Improved FIM Sec. 3
435	62611	1.6 sec	27.1 sec

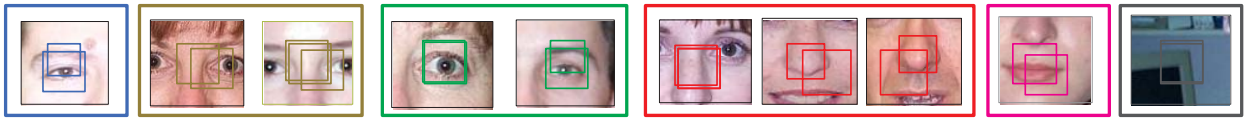
6.3. Summarization of visual phrases

We choose a visual phrase lexicon of size $|\mathbf{\Psi}| = 10$ and further cluster related phrases into semantically coherent phrase patterns. The best summarization results are shown in Fig. 6 and Fig. 7, with cluster number $|\mathbf{H}| = 6$ and $|\mathbf{H}| = 2$ for the face and car database respectively. Each visual phrase is composed of spatially co-located words, either the second-order or the third-order. As we allow multiple visual primitives to be located in the same position, it is possible that two visual primitives located in the same position (with different scales or orientations) are labeled with two different visual words. In order to evaluate the performance of the phrase summarization, we apply the precision and recall scores to measure the purity of discovered pattern classes $\mathcal{H}_i$:

$$Recall = \#detects/(\#detects + \#miss\ \ detects),$$

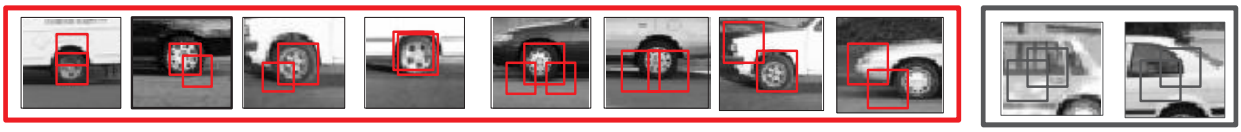
$$Precision = \#detects/(\#detects + \#false\ \ detects).$$

From Fig. 6 and Fig. 7, it can be seen that the summarized visual phrase patterns $\mathcal{H}_i$ are associated with semantic parts



H1: left eyes Prc.: 100% Rec.: 26.1%
H2: between eyes Prc.: 98.8% Rec.: 18.4%
H3: right eyes Prc.: 99.3% Rec.: 32.2%
H4: noses Prc.: 96.3% Rec.: 24.9%
H5: mouths Prc.: 100% Rec.: 10.1%
H6: coners Prc.: N.A. Rec.: N.A.

Figure 6. Visual phrase lexicon $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and its summarization results ($|\mathbf{H}| = 6$) for the face category. Each of the 10 sub-images represents a visual phrase, where rectangles denote the visual primitives (*e.g.*, a PCA-SIFT interest point at its scale). Compositionally similar visual phrases are clustered into a phrase pattern class (6 classes in total). Note the 3_{rd}, 4_{th}, 5_{th}, and 10_{th} phrases contain visual primitives that are co-located in the same position, but with different orientations. We use two highly overlapped rectangles to distinguish them, although they are actually located in the same position.



H1: wheels Prc.: 97.5% Rec.: 22.8%
H2: car bodies Prc.: 71.3% Rec.:N.A.

Figure 7. Visual phrase lexicon $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and its summarization results ($|\mathbf{H}| = 2$) for car database. The 4_{th} phrase contains two visual words co-located in the same position, but with different orientations.

with very high precision and reasonably good recall scores. This thus convinces us to apply the discovered visual phrase patterns as supervision to refine VWL, *i.e*. the instances of visual phrases provide training samples of the visual word classes.

6.4. Top-down refinement of Visual Word Lexicon

To implement NCA for metric learning, we manually select the best 5 phrases as positive training data from the top-10 phrases. Our objective of metric learning is to expand the inter-class distance while reducing the intra-class distance in the training data. It is important to note that not all the visual primitives that are labeled with foreground visual words $W_i \in \mathbf{\Omega}^+$ are qualified to serve as positive training samples. We only consider those visual primitives not only labeled with a foreground word $W_i \in \mathbf{\Omega}^+$, but also can compose into any of the 5 selected visual phrase with other visual primitives in its spatial neighborhood. Considering the large number of background words $W_i \in \mathbf{\Omega}^-$, we only select a small set of them which are more likely to be generated from the background class. For example, among the visual primitives labeled with background words $W_i \in \mathbf{\Omega}^-$, we only select those outliers that are located in the lowest density regions in the feature space.

After obtaining a new metric through NCA, we rebuild VWL $\mathbf{\Omega}$ based on the learned metric, with the number of clusters slightly less than before. Based on the new $\mathbf{\Omega}$, VPL $\mathbf{\Psi}$ can also be updated. The comparison results of the original $\mathbf{\Psi}$ and those after refinements are shown in Fig. 8. It can be seen that the precision ρ^+ of $\mathbf{\Psi}$ is improved after our top-down refinement of $\mathbf{\Omega}$.

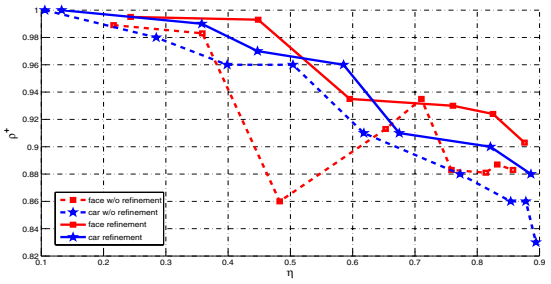


Figure 8. Comparison of VPL $\mathbf{\Psi}$ before and after the top-down refinement.

7. Related Work

Many existing work devote to seeking visual features that are spatially co-occurrent, such as "semi-local" parts [12], depedency regions [22], frequent spatial configurations [16], perceptual groups of local features [10], sparse flexible model [2] and hyperfeatures [1]. Various criteria are proposed to measure the spatial dependency among the primitive visual features. To avoid large computational complexity, some ad-hoc methods are proposed and others only consider the pair-wise dependency between features while ignoring the higher-order relations. Previous part-based methods such as the constellation model [4] and many extensions also aim to detect objects and learn object models in images, by reinforcing the number of parts and spatial constraints among them. In general, these methods are computationally demanding and prior knowledge of the object category is required. There are related work in modeling textons [25] and learning generic parts from multiple object categories for object recognition [20, 5, 23]. In data

mining domain, there are also related work in discovering spatial collocation patterns [9] and interpreting mined frequent patterns [24, 14]. These methods are concerned on discrete data, and may not be directly applied to visual data.

8. Conclusion

This paper devotes to the discovery of less ambiguous visual phrases lexicon (VPL) for better representing images than visual words lexicon (VWL) obtained from clustering local features. Several new data mining techniques such as closed FIM and pattern summarization are employed, with major modifications to fit the image data. By applying the top-down refinement, the lower-level VWL is updated by the feedback from the higher-level VPL. Our experimental results show that the discovered VPL can well distinguish the common foreground objects from the backgrounds. As a pure data-driven bottom-up approach, our method does not assume and depend on a top-down generative model in discovering compositions of visual primitives. In general, no prior knowledge is required, and the discovered VPL can actually be used to learn a generative model. Our method is computationally much more efficient than those methods based on generative models.

Acknowledgment

This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures: Multilevel local coding for visual recognition. In *Proc. European Conf. on Computer Vision*, 2006. 7

[2] G. Carneiro and D. Lowe. Sparse flexible models for local features. In *Proc. European Conf. on Computer Vision*, 2006. 7

[3] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorization with bags of words. In *Proc. Workshop on European Conf. on Compsuter Vision*, 2004. 1

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003. 5, 7

[5] S. Fidler, G. Berginc, and A. Leonardis. Hierarchical statistical learning of generic parts of object structure. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 7

[6] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. In *Proc. of Neural Information Processing Systemson*, 2005. 5

[7] G. Grahne and J. Zhu. Fast algorithms for frequent itemset mining using fp-trees. *IEEE Transaction on Knowledge and Data Engineering*, 2005. 3, 6

[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. 3, 4

[9] Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transaction on Knowledge and Data Engineering*, 16(12):1472–1485, 2004. 8

[10] M. Jamieson, S. Dickinson, S. Stevenson, and S. Wachsmuth. Using language to drive the perceptual grouping of local image features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 1, 7

[11] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. 5

[12] S. Lazebnik, C. Schmid, and J. Ponce. A discriminative framework for texture and object recognition using local image features. In *Proc. European Conf. on Computer Vision*, 2006. 1, 7

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 2004. 1

[14] Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai. Generating semantic annotations for frequent patterns with context analysis. In *Proc. ACM SIGKDD*, 2006. 8

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Intl. Journal of Computer Vision*, 2005. 1

[16] T. Quack, V. Ferrari, and L. V. Gool. Video mining with frequent itemset configurations. In *Proc. Int. Conf. on Image and Video Retrieval*, 2006. 1, 7

[17] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentation to discover objects and their extent in image collections. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 1

[18] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000. 5

[19] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. 1, 2

[20] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2005. 7

[21] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 1

[22] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 1, 7

[23] J. Winn, A.Criminisi, and T.Minka. Object categorization by learned universal visual dictionary. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2005. 7

[24] X. Yan, H. Cheng, J. Han, and D. Xin. Summarizing itemset patterns: a profile-based approach. In *Proc. ACM SIGKDD*, 2005. 5, 8

[25] S.-C. Zhu, C. en Guo, Y. Wang, and Z. Xu. What are textons? *Intl. Journal of Computer Vision*, 2005. 7