Mining Visual Collocation Patterns via Self-Supervised Subspace Learning

Junsong Yuan, *Member, IEEE,* and Ying Wu, *Senior Member, IEEE*

***Abstract*—Traditional text data mining techniques are not directly applicable to image data which contain spatial information and are characterized by high-dimensional visual features. It is not a trivial task to discover meaningful visual patterns from images, because the content variations and spatial dependency in visual data greatly challenge most existing data mining methods. This paper presents a novel approach to coping with these difficulties for mining visual collocation patterns. Specifically, the novelty of this work lies in the following new contributions: (1) a principled solution to the discovery of visual collocation patterns based on frequent itemset mining; and (2) a self-supervised subspace learning method to refine the visual codebook by feeding back discovered patterns via subspace learning. The experimental results show that our method can discover semantically meaningful patterns efficiently and effectively.**

***Index Terms*—image data mining, visual pattern discovery, visual collocation pattern.**

I. Introduction

Motivated by the previous success in mining structured data (e.g., transaction data) and semi-structured data (e.g., text), it has aroused our curiosity in finding meaningful patterns in non-structured multimedia data like images and videos [1] [2] [3] [4] [5]. For example, as illustrated in Fig. 1, once we can extract some invariant visual primitives such as interest points [6] or salient regions [7] from the images, we can represent each image as a collection of such visual primitives characterized by high-dimensional feature vectors. By further quantizing those visual primitives to discrete "visual items" (also known as "visual words") through clustering these high-dimensional features [1], each image is represented by a set of transaction records, where each transaction corresponds to a local image region and describes its composition of visual items. After that, data mining techniques can be applied to such a transaction database induced from images for discovering visual collocation pattern.

Although the discovery of visual patterns from images appears to be quite exciting, data mining techniques that are successful in transaction and text data may not be simply applied to image data that contain high-dimensional features and have spatial structures. Unlike transaction and text data that are composed of discrete elements without much ambiguity (*i.e.* predefined items and vocabularies), visual patterns generally exhibit large variabilities in their visual appearances. A same visual pattern may look very different under different views, scales, lighting conditions, not to mention partial occlusion. It is very difficult, if not impossible, to obtain invariant visual features that are insensitive to these variations such that they can uniquely characterize visual primitives. Therefore although a discrete item codebook can be forcefully obtained by clustering high-dimensional visual features (e.g., by k-means clustering), such "visual items" tend to be much more ambiguous than the case of transaction and text data. Thus the imperfect clustering of visual primitives brings large challenges when directly applying traditional data mining methods to image data. Specifically, the ambiguity lies in two aspects: *synonymy* and *polysemy* [8]. A synonymous visual item shares the same semantic meanings with other visual items. Because the corresponding underlying semantics is split and represented by multiple visual items, synonymy leads to over-representations. On the other hand, a polysemous visual item may mean different things under different contexts. Thus polysemy leads to under-representations. Both phenomena appear quite often when clustering visual primitives through an unsupervised way. The root of these phenomena is the large uncertainties within non-structured visual data in the high-dimensional space. 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 *visual collocation* (or *co-occurrence*) of several visual items is likely to be much less ambiguous. Therefore, it is of great interest to automatically discover these collocation visual patterns. Once such visual collocation patterns are discovered, they can help to learn a better representation for clustering visual primitives.

However, since visual patterns exhibit more complex structure than transaction and text pattern, the difficulty in representing and discovering spatial patterns in images prevents straightforward generalization of traditional data mining methods that are applicable for transaction data. For example, unlike traditional transaction database where records are independent of each other, the induced transactions generated by image patches can be correlated due to spatial overlap. This phenomenon complicates the data mining process for spatial data, because simply counting the occurrence frequencies is doubtable and thus a frequent pattern is not necessarily a meaningful pattern. Thus special care needs to be taken. Although there exist methods [9] [10] [11] for spatial collocation pattern discovery from geo-spatial data, they cannot be

Junsong Yuan is with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore (email: jsyuan@ntu.edu.sg), and Ying Wu is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, 60208 USA (email:yingwu@ece.northwestern.edu).

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

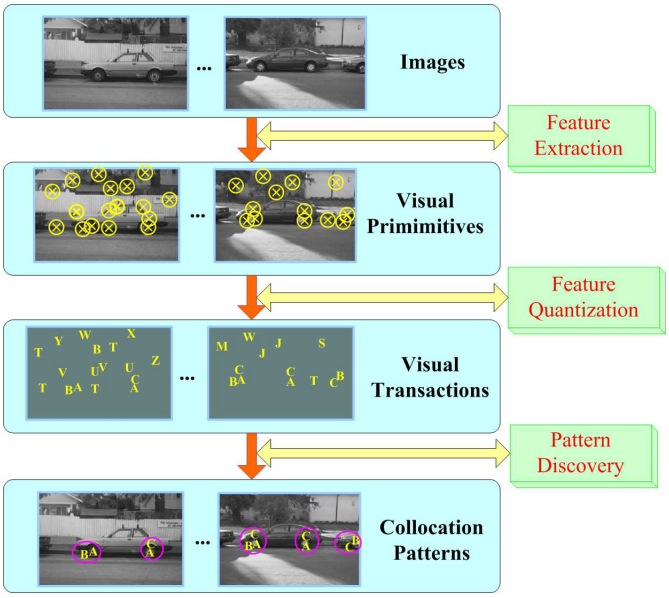

Fig. 1. The illustration of collocation visual patterns from images. There are two kinds of imperfectness when translating image data into transaction data for data mining. First, the visual primitives can be miss-detected in the feature extraction process, due to the occlusion, bad lighting conditions or the unreliable detectors. Secondly, even if a visual primitive is extracted, it can be wrongly clustered into a visual item due to visual polysemy and synonymy. A direct pattern mining on the noisy transaction database cannot obtain satisfactory results.

directly applied to image data which are characterized by high-dimensional features. Moreover, the spatial co-occurrences of the items do not necessarily indicate the real associations among them, because a frequent spatial collocation pattern can be generated by the self-repetitive texture in the image and thus is not semantically meaningful. Thus, finding frequent patterns may not always output meaningful and informative patterns in image data.

Given a collection of images, our objective of image data mining is to discover meaningful visual patterns that appear repetitively among the images. Compared with the background clutters, such visual patterns are of great interests thus should be well treated in clustering visual items. For example, given a few face photos of different persons, can we discover the visual collocations like eyes and noses that can interpret the face category? Moreover, once these visual patterns are discovered, can they help to learn a better feature representation via subspace learning?

To address these problems, this paper presents a novel bottom-up approach to discovering semantically meaningful visual patterns from images. As shown in Fig. 1, an image is represented by a collection of visual items after clustering visual primitives. To discover visual collocation patterns, a new data mining criterion is proposed. Instead of using the co-occurrence frequency as the criterion for mining the meaningful collocation patterns in images, a more plausible *visual collocation mining* based on likelihood ratio test is proposed to evaluate the significance of a visual itemset. Secondly, once the visual collocation patterns (foreground items) and background items are discovered, we can feed them back into the clustering procedure by learning a better subspace representation via metric learning, to distinguish the object of interests from the cluttered background. Then the visual primitives can be better clustered in the learned subspace. To this end, we propose a self-supervised subspace learning of visual items. By taking advantage of the discovered visual patterns, such a top-down refinement procedure helps to reduce the ambiguities among visual items and better distinguish foreground items from the background items. Our experiments on three object categories from the 101 Caltech dataset demonstrate the effectiveness and efficiency of our proposed method.

The rest of the paper is organized as follows. We discuss the related work in Section II, followed by the overview of our approach in Section III. The discovery of visual collocation patterns is presented in Section IV. After that we discuss how to refine the discovered patterns via metric learning in Section V. The experiments are conducted in Section VI followed by the conclusion in Section VII.

II. Related Work

By characterizing an image as a collection of primitive visual features that highlight the local image invariants, and clustering these primitive features into discrete visual words, we can translate an image to a visual document. Such a "bag of words" model bridges the text data mining and image data mining research, and has been extensive applied in image retrieval [12], recognition [13], as well as image data mining [1] [14] [15]. As a similar treatment of texts, previous text information retrieval and data mining methods can be applied to image data. For example, in [16], text-retrieval methods are applied to search visual objects in videos. Statistical natural language models, such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), are applied to discover object categories from images [17] [8].

Although it brings many benefits by representing images as visual documents, the induced visual vocabulary tend to be much more ambiguous than that in text. To learn a better visual vocabulary, [18] discusses the limitation of k-means clustering and propose a new strategy to build the codebook. However, unsupervised learning of a good visual vocabulary is difficult. As a treatment to resolve the ambiguity of polysemous and synonymous visual words, spatial context information in image is taken into consideration. It is noted that the *co-occurrence* of visual words, namely a composition of visual items, has a better representation power and is likely to be less ambiguous, such as the visual phrases [19] [20], visual synset [21] and dependent regions [22].

To discover visual patterns, some other methods consider the spatial configuration among the visual features when modeling the spatial image pattern. In [23], attributed relational graphs (ARG) are applied to model the spatial image pattern. EM algorithm is used for parameters learning. In [24], spatial relations are defined in terms of the distance and direction between pair of detected parts. In [25], both geometry and appearance information of generic visual object category is coded in a hierarchical generative model. A constellation model is used to model a visual object category in [26]. The proposed probabilistic model is able to handle all kinds of

variations of the objects, such as shape, appearance, occlusion and relative scale. However, the structure of the model and the object parts need to be manually selected. In [27], a tree model is proposed to discover and segment visual objects that belong to the same category. Despite its powerful modeling ability, the training of these generative models is usually time-consuming. When using the EM algorithm for learning, it is easy to be trapped by the local optimality. In [28], a generative model, called active basis model, is proposed to learn a deformable template to sketch the common object from images. The active basis consists of a small number of Gabor wavelet elements at selected locations and orientations.

Instead of using top-down generative model to discover visual patterns, these are also data-driven bottom-up approaches. To perform efficient image data mining, frequent itemset mining is applied in [29] [30] by translating each local image region into a transaction. In order to consider the spatial configuration among the visual items, in [31], semantics are represented as visual elements and geometric relationships between them. An efficient mining method is proposed to find pair-wise associations of visual elements that have consistent geometric relationships sufficiently often. In [32], an efficient image pattern mining approach is proposed to handle the scale, translation, and rotation variations in mining frequent spatial patterns in images. In [33], contextual visual word is proposed to improve the image search and near duplicate copy detection.

There are also related works in data mining. It is of great interests to discover frequent patterns in data mining research. For example, frequent itemset mining (FIM) and its extensions [34] [35] [36] have been extensively studied. However, a highly frequent pattern may not be informative or interesting, thus a more important task is to extract informative and potentially interesting patterns from the possibly noisy data. This can be done by mining meaningful patterns either through post-processing the FIM results or proposing new data mining criteria, including mining compressed patterns [37] [38] [39], approximate patterns [40] [41] [42] and pattern summarization [43] [44] [45]. These data mining techniques may discover meaningful frequent itemsets and represent them in a compact way.

III. Overview

A. Notations and basic concepts

Each image in the database is described by a set of visual primitives: $\mathcal{I} = \{v_i = \left(\vec{f_i}, x_i, y_i\right)\}$, where $\vec{f_i}$ denotes the high-dimensional feature and $\{x_i, y_i\}$ denotes the spatial location of v_i in the image. We treat these visual primitives as the *atomic visual patterns*. For each visual primitive $v_i \in \mathcal{I}$, its local spatial neighbors form a *group* $\mathcal{G}_i = \{v_i, v_{i_1}, v_{i_2}, \cdots, v_{i_K}\}$. For example, $\mathcal{G}_i$ can be the spatial K-nearest neighbors (K-NN) or ϵ-nearest neighbors (ϵ-NN) of v_i under the Euclidean distance. As illustrated in Fig. 2, the image database $\mathbf{D}_\mathcal{I} = \{\mathcal{I}_t\}_{t=1}^T$ can generate a collection of such groups, where each group $\mathcal{G}_i$ is associated with a visual primitive v_i. We want to mention that two spatially neighbored groups may share some visual primitives

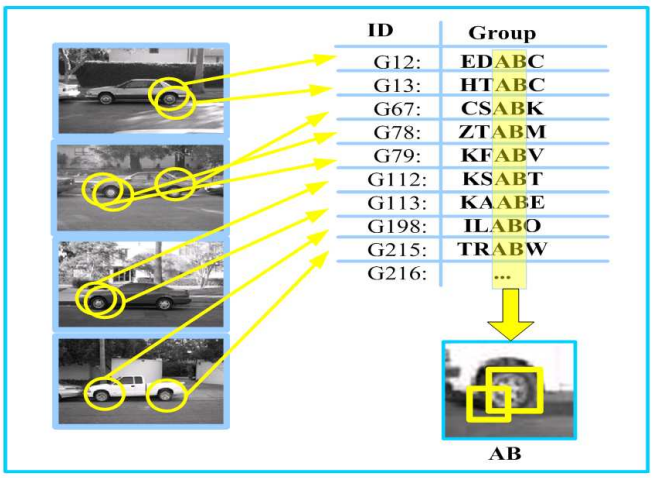

Fig. 2. Illustrations of the visual groups and the discovery of visual collocations. Each circle corresponds to a spatial group (namely a transaction), which is composed of 5-NN visual items. An image can generate a collection of such groups for data mining. A and B are discovered visual collocation patterns.

due to their spatial overlap. By further quantizing all the high-dimensional features $\vec{f_i} \in \mathbf{D}_\mathcal{I}$ into M classes through k-means clustering, a codebook of visual primitives $\mathbf{\Omega}$ can be obtained. We call every prototype W_k in the codebook $\mathbf{\Omega} = \{W_1, ..., W_M\}$ a *visual item*. Because each visual primitive is uniquely assigned to one of the visual items W_i, the group $\mathcal{G}_i$ can be transformed into a *transaction* $\mathcal{T}_i$. More formally, given the group dataset $\mathbf{G} = \{\mathcal{G}_i\}_{i=1}^N$ generated from $\mathbf{D}_\mathcal{I}$ and the visual item codebook $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = M$), we can induce a transaction database $\mathbf{T} = \{\mathcal{T}_i\}$. Such an induced transaction database is essentially based on the *centric reference feature model* for mining association rules [10]. Given the visual item codebook $\mathbf{\Omega}$, a sub-set $\mathcal{P} \subset \mathbf{\Omega}$ is called a *visual itemset* (itemset for short). For a given itemset $\mathcal{P}$, the transaction $\mathcal{T}_i$ which includes $\mathcal{P}$ is called an *occurrence* of $\mathcal{P}$, *i.e.* $\mathcal{T}_i$ is an occurrence of $\mathcal{P}$, if $\mathcal{P} \subseteq \mathcal{T}_i$. Let $\mathbf{T}(\mathcal{P})$ denote the set of all the occurrences of $\mathcal{P}$ in $\mathbf{T}$, and the *frequency* of $\mathcal{P}$ is the number of its occurrences denoted as:

$$frq(\mathcal{P}) = |\mathbf{T}(\mathcal{P})| = |\{i : \forall j \in \mathcal{P}, t_{ij} = 1\}|, \qquad (1)$$

where $t_{ij} = 1$ denotes that the j_{th} item appears in the i_{th} transaction, and $t_{ij} = 0$ otherwise.

For a given threshold θ, called a *minimum support*, itemset $\mathcal{P}$ is *frequent* if $frq(\mathcal{P}) > \theta$. It is not a trivial task to discover all the frequent itemsets given dataset $\mathbf{T}$, because the number of possible itemsets is exponentially large with respect to the codebook size. For example, the codebook $\mathbf{\Omega}$ has in total $2^{|\mathbf{\Omega}|}$ candidates for frequent itemsets, therefore exhaustive check is infeasible for large codebooks. Also, if an itemset $\mathcal{P}$ appears frequently, then all of its sub-sets $\mathcal{P}' \subset \mathcal{P}$ will also appear frequently, *i.e.* $frq(\mathcal{P}) > \theta \Rightarrow frq(\mathcal{P}') > \theta$. For example, a frequent itemset $\mathcal{P}$ composed with n items can generate 2^n frequent sub-itemsets including itself and the null itemset. To eliminate this redundant representation, *closed frequent itemsets* is introduced [46]. Thus this guarantees that no visual collocations will be left out. The *closed frequent itemset* is defined as follows.

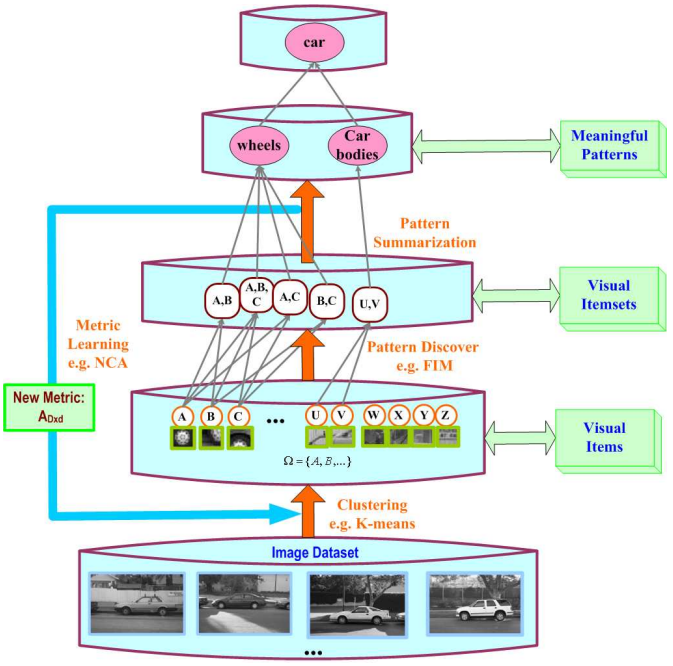

Fig. 3. The overview for the proposed method for mining visual collocation patterns. We propose a hierarchical and self-supervised visual pattern discovery method to handle the imperfectness from the visual vocabulary and can reveal the hierarchical structure of visual patterns

Definition 1: **closed frequent itemset**
If for an itemset $\mathcal{P}$, there is no other itemset $\mathcal{Q} \supseteq \mathcal{P}$ that can satisfy $\mathbf{T}(\mathcal{P}) = \mathbf{T}(\mathcal{Q})$, we say $\mathcal{P}$ is *closed*. For any itemset $\mathcal{P}$ and $\mathcal{Q}$, $\mathbf{T}(\mathcal{P} \cup \mathcal{Q}) = \mathbf{T}(\mathcal{P}) \cap \mathbf{T}(\mathcal{Q})$, and if $\mathcal{P} \subseteq \mathcal{Q}$ then $\mathbf{T}(\mathcal{Q}) \subseteq \mathbf{T}(\mathcal{P})$.

To find frequent itemsets, we apply the FP-growth algorithm to discover *closed* frequent itemsets [47]. The number of closed frequent itemsets is much less than the frequent itemsets, and they compress information of frequent itemsets in a lossless form, *i.e.* the full list of frequent itemsets $\mathbf{F} = \{\mathcal{P}_i\}$ and their corresponding frequency counts can be exactly recovered from the compressed representation of closed frequent itemsets. As FP-tree has a prefix-tree structure and can store compressed information of frequent itemset, it can efficiently discover all the closed frequent sets from transaction dataset $\mathbf{T}$.

B. Overview of our method

We present the overview of our visual pattern discovery method in Fig. 3. Given a collection of images, we detect the local interest features, followed by clustering them into a group of visual items. The spatial dependences of these visual items are discovered via using the proposed data mining methods. Once these spatial collocation patterns are discovered, it can guide the subspace learning in finding a better feature space for visual item clustering. Finally, the discovered visual items are further grouped to recover the visual patterns.

In Section IV, we present our new criteria for discovering visual collocation patterns $\mathcal{P}_i \subset \mathbf{\Omega}$. After that in Section V, we feed back the discovered visual collocations $\mathbf{\Psi}$ to refine the data mining via metric learning. The experiments are conducted in Section VI and we conclude in Section VII.

IV. Discovering Visual Collocation Patterns

A. Visual Primitive Extraction

We apply the PCA-SIFT points [48] as the *visual primitives*. Such visual primitives are mostly located in the informative image regions such as corners and edges, and the features are invariant under rotations, scale changes, and slight viewpoint changes. Normally each image may contain hundreds to thousands of such visual primitives based on the size of the image. According to [48], each visual primitive is a 41×41 gradient image patch at the given scale, and rotated to align its dominant orientation to a canonical direction. Principal component analysis (PCA) is applied to reduce the dimensionality of the feature. Finally each visual primitive is described by a 35-dimensional feature vector $\vec{f}_i$. These visual primitives are initially clustered into visual items through k-means clustering, using Euclidean metric in the feature space. We will discuss how to obtain a better visual item codebook $\mathbf{\Omega}$ based on the proposed self-supervised metric learning scheme in Sec. V.

B. Finding Meaningful Visual Collocations

Given an image dataset $\mathbf{D}_{\mathcal{I}}$ and its induced transaction database $\mathbf{T}$, the task is to discover the visual collocation patterns $\mathcal{P} \subset \mathbf{\Omega}$ ($|\mathcal{P}| \geq 2$). Each visual collocation is composed by a collect of visual items that occur together spatially. To evaluate the qualification of a $\mathcal{P} \subseteq \mathbf{\Omega}$, simply checking its frequency $frq(\mathcal{P})$ in $\mathbf{T}$ is far from sufficient. For example, even if an itemset appears frequently, it is not clear whether such co-occurrences among the items are statistically significant or just by chance. In order to evaluate the statistical significance of a frequent itemset $\mathcal{P}$, we propose a new likelihood ratio test criterion. We compare the likelihood that $\mathcal{P}$ is generated by the meaningful pattern versus the likelihood that $\mathcal{P}$ is randomly generated, *i.e.* by chance.

More formally, we perform the likelihood ratio test to measure a visual collocation based on the two hypotheses, where

$\mathbf{H_0}$: occurrences of $\mathcal{P}$ are randomly generated;
$\mathbf{H_1}$: occurrences of $\mathcal{P}$ are generated by the hidden pattern.

Given a transaction database $\mathbf{T}$, the likelihood ratio $L(\mathcal{P})$ of a visual collocation $\mathcal{P} = \{W_i\}_{i=1}^{|\mathcal{P}|}$ can be calculated as:

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

Here $P(\mathcal{T}_i|H_1) = \frac{1}{N}$ is the prior, and $P(\mathcal{P}|\mathcal{T}_i, H_1)$ is the likelihood that $\mathcal{P}$ is generated by a hidden pattern and is observed at a particular transaction $\mathcal{T}_i$. Therefore $P(\mathcal{P}|\mathcal{T}_i, H_1) = 1$, if $\mathcal{P} \subseteq \mathcal{T}_i$, and $P(\mathcal{P}|\mathcal{T}_i, H_1) = 0$, otherwise. Consequently, based on Eq. 1, we can calculate $P(\mathcal{P}|H_1) = \frac{frq(\mathcal{P})}{N}$. We also assume that the items $W_i \in \mathcal{P}$ are conditionally independent

under the null hypothesis H_0, and $P(W_i|H_0)$ is the prior of item $W_i \in \mathbf{\Omega}$, *i.e.* the total number of visual primitives that are labeled with W_i in the image database $\mathbf{D}_\mathcal{I}$. We thus refer $L(\mathcal{P})$ as the "significance" score to evaluate the deviation of a visual collocation pattern $\mathcal{P}$. If $\mathcal{P}$ is a second-order itemset, then $L(\mathcal{P})$ degenerates to the pointwise mutual information criterion, *e.g.*, the lift criterion [46].

It is worth noting that $L(\mathcal{P})$ may favor high-order collocations even though they appear less frequently. Table I presents such an example, where 90 transactions have only items A and B; 30 transactions have A,B and C; 61 transactions have D and E; and 19 transactions have C and E.

TABLE I
TRANSACTION DATABASE $\mathbf{T}_1$.

transaction	number	$L(\mathcal{P})$
AB	90	1.67
ABC	30	1.70
DE	61	2.5
CE	19	0.97

From Table I, It is easy to evaluate the significant scores for $\mathcal{P}_1 = \{A, B\}$ and $\mathcal{P}_2 = \{A, B, C\}$ with $L(\mathcal{P}_1) = 1.67$ and $L(\mathcal{P}_2) = 1.70 > L(\mathcal{P}_1)$. This result indicates that $\mathcal{P}_2$ is a more significant pattern than $\mathcal{P}_1$ but counter-intuitive. This observation challenges our intuition because $\mathcal{P}_2$ is not a cohesive pattern. For example, the other two sub-patterns of $\mathcal{P}_2$, $\mathcal{P}_3 = \{A, C\}$ and $\mathcal{P}_4 = \{B, C\}$, contain almost independent items: $L(\mathcal{P}_3) = L(\mathcal{P}_4) = 1.02$. Actually, $\mathcal{P}_2$ should be treated as a variation of $\mathcal{P}_1$ as C is more likely to be a noise. The following equation explains what causes the incorrect result. We calculate the significant score of $\mathcal{P}_2$ as:

$$L(\mathcal{P}_2) = \frac{P(A, B, C)}{P(A)P(B)P(C)} = L(\mathcal{P}_1) \times \frac{P(C|A, B)}{P(C)}. \quad (3)$$

Therefore when there is a small disturbance with the distribution of C over $\mathbf{T}_1$ such that $P(C|A, B) > P(C)$, $\mathcal{P}_2$ will compete $\mathcal{P}_1$ even though $\mathcal{P}_2$ is not a cohesive pattern (*e.g.* C is not related to either A or B). To avoid those free-riders such as C for $\mathcal{P}_1$, we perform a more strict test on the itemset. For a high-order $\mathcal{P}$ ($|\mathcal{P}| > 2$), we perform the t-test for each pair of its items to check if items W_i and W_j ($W_i, W_j \in \mathcal{P}$) are really dependent (see Appendix A for details.) A high-order collocation $\mathcal{P}_i$ is meaningful only if all of its pairwise subsets can pass the test: $\forall i, j \in \mathcal{P}$, $t(\{W_i, W_j\}) > \tau$, where τ is the confidence threshold for the t-test. This further reduces the redundancy among the discovered itemsets.

Finally, to assure that a visual collocation $\mathcal{P}$ is meaningful, we also require it to appear relatively frequent in the database, *i.e.* $frq(\mathcal{P}) > \theta$, such that we can eliminate those collocations that appear rarely but happen to exhibit strong spatial dependency among items. With these three criteria, a visual collocation pattern is defined as follows.

Definition 2: **Visual Collocation Pattern**
An itemset $\mathcal{P} \subseteq \mathbf{\Omega}$ is a (θ, τ, γ)-meaningful visual collocation, if it is:

1) **frequent**: $frq(\mathcal{P}) > \theta$;
2) **pair-wisely cohesive**: $t(\{W_i, W_j\}) > \tau, \forall i, j \in \mathcal{P}$;
3) **significant**: $L(\mathcal{P}) > \gamma$.

C. Spatial Dependency among Induced Transactions

Suppose primitives v_i and v_j are spatial neighbors, their induced transaction $\mathcal{T}_i$ and $\mathcal{T}_j$ will have large spatial overlap. Due to such spatial dependency among the transactions, it can cause over-counting problem if simply calculating $frq(\mathcal{P})$ from Eq. 1. Fig. 4 illustrates this phenomena where $frq(\mathcal{P})$ contains duplicate counts.

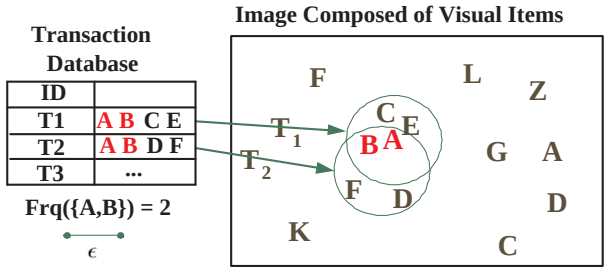

Fig. 4. Illustration of the frequency over-counting caused by the spatial overlap of transactions. The itemset $\{A, B\}$ is counted twice by $\mathcal{T}_1 = \{A, B, C, E\}$ and $\mathcal{T}_2 = \{A, B, D, F\}$, although it has only one instance in the image. Namely there is only one pair of A and B that co-occurs together, such that $d(A, B) < 2\epsilon$ with ϵ the radius of $\mathcal{T}_1$. In the texture region where visual primitives are densely sampled, such over-count will largely exaggerate the number of repetitions for a texture pattern.

In order to address this transaction dependency problem, we apply a two-phase mining scheme. First, without considering the spatial overlaps, we perform closed FIM to obtain a candidate set of frequent itemsets. For these candidates $\mathbf{F} = \{\mathcal{P}_i : frq(\mathcal{P}_i) > \theta\}$, we re-count the number of their real instances exhaustively through the original image database $\mathbf{D}_\mathcal{I}$, not allowing duplicate counts. This needs one more scan of the whole database. Without causing confusion, we denote $\hat{frq}(\mathcal{P})$ as the occurrence number of $\mathcal{P}$ and use it to update $frq(\mathcal{P})$. 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 transaction number with K the average size of transactions. In practice, as $\hat{N}$ is hard to estimate, we rank $\mathcal{P}_i$ according to their significant value $L(\mathcal{P})$ and perform the top-K pattern mining.

Integrating all the steps in this section, we present our algorithm to discover meaningful visual collocations in Algorithm 1.

Algorithm 1: Visual Collocation Mining

input : Transaction dataset $\mathbf{T}$, parameters: (θ, τ, γ)
output: a collection of meaningful visual collocations: $\mathbf{\Psi} = \{\mathcal{P}_i\}$

Init: closed FIM with $frq(\mathcal{P}_i) > \theta$: $\mathbf{F} = \{\mathcal{P}_i\}$, $\mathbf{\Psi} \longleftarrow \emptyset$;
foreach $\mathcal{P}_i \in \mathbf{F}$ **do** `GetRealInstanceNumber(`$\mathcal{P}_i$`)`
for $\mathcal{P}_i \in \mathbf{F}$ **do**
  if $L(\mathcal{P}_i) > \gamma \ \wedge$ `PassPairwiseTtest` $(\mathcal{P}_i)$ **then**
    $\mathbf{\Psi} \longleftarrow \mathbf{\Psi} \cup \mathcal{P}_i$
Return $\mathbf{\Psi}$

V. Self-supervised Refinement of Visual Item Codebook

A. Foreground V.S. Background

As discussed earlier, our image data mining method highly relies on the quality of the visual item codebook $\mathbf{\Omega}$. A bad clustering of visual primitives brings large quantization errors when generating the transactions. Such a quantization error will affect the data mining results significantly. Thus a good $\mathbf{\Omega}$ is required. To improve the codebook construction, we propose to use the discovered collocation patterns to supervise the clustering process. Although there is no supervision available initially, the unsupervised data mining process actually discover useful information for supervision. Thus it is called self-supervised refinement.

We notice that each image contains two layers: a foreground object and the background clutters. Given a collection of images containing the same category of objects, the foreground objects are similar, while the background clutters are different from each other. The discovered visual collocations are supposed to be associated with the foreground object. As each image is composed of the visual items, we can partition the codebook into two parts $\mathbf{\Omega} = \mathbf{\Omega}^+ \cup \mathbf{\Omega}^-$, where items in $\mathbf{\Omega}^+$ are more likely to appear in the foreground object, while items in $\mathbf{\Omega}^-$ are more likely to appear in the background. Thus it provides information to learn a better codebook.

By discovering a set of visual collocations $\mathbf{\Psi} = \{\mathcal{P}_i\}$, we define the *foreground item codebook* as follows:

Definition 3: **foreground codebook $\mathbf{\Omega}^+$**
Given a set of visual collocations $\mathbf{\Psi} = \{\mathcal{P}_i\}$, an item $W_i \in \mathbf{\Omega}$ is a foreground item if it belongs to any collocation pattern $\mathcal{P} \in \mathbf{\Psi}$, namely, $\exists \mathcal{P} \in \mathbf{\Psi}$, such that $W_i \subset \mathcal{P}$. All of the foreground items compose the foreground codebook $\mathbf{\Omega}^+ = \bigcup_{i=1}^{|\mathbf{\Psi}|} \mathcal{P}_i$.

With the foreground codebook, the *background codebook* becomes $\mathbf{\Omega}^- = \mathbf{\Omega} \backslash \mathbf{\Omega}^+$. Each visual primitive belongs to either the foreground object (positive class) or the background clutter (negative class).

Our goal now is to use the data mining results to refine the codebooks $\mathbf{\Omega}^+$ and $\mathbf{\Omega}^-$, such that they can better distinguish the two classes. For the negative class, any visual primitive that belongs to $\mathbf{\Omega}^-$ can be the negative training example. However, for the positive class $\mathbf{\Omega}^+$, not all of items in $\mathbf{\Omega}^+$ are qualified to be positive samples. We only choose those instances of the visual collocations.

B. Learning a better metric for clustering

With these training examples via data mining, we transfer the unsupervised clustering problem into semi-supervised clustering to obtain a better codebook $\mathbf{\Omega}$. Our task is to cluster all the visual primitives $v_i \in \mathbf{D}_\mathcal{I}$. Now some of the visual primitives are already labeled according to the discovered visual collocations and the background visual items. Thus we can use these labeled primitives to help cluster all of visual primitives.

We apply the nearest component analysis (NCA) [49] to improve the clustering results by learning a better Mahalanobis distance metric in the feature space. Similar to the linear discriminative analysis (LDA), NCA targets at learning a global linear projection matrix A. However, unlike LDA, NCA does not need to assume that each visual item class has a Gaussian distribution and thus can be applied to more general cases. Given two visual primitives v_i and v_j, NCA learns a new metric A and the distance in the transformed space is: $d_A(v_i, v_j) = (\vec{f_i} - \vec{f_j})^T A^T A (\vec{f_i} - \vec{f_j}) = (A\vec{f_i} - A\vec{f_j})^T (A\vec{f_i} - A\vec{f_j})$.

The objective of NCA is to maximize a stochastic variant of the leave-one-out K-NN score on the training set. In the transformed space, a point v_i selects another point v_j as its neighbor with probability:

$$p_{ij} = \frac{exp(-\|A\vec{f_i} - A\vec{f_j}\|^2)}{\sum_{k \neq i} exp(-\|A\vec{f_i} - A\vec{f_k}\|^2)}, \qquad p_{ii} = 0. \quad (4)$$

Under the above stochastic selection rule of nearest neighbors, NCA tries to maximize the expected number of points correctly classified under the nearest neighbor classifier (the average leave-one-out performance):

$$f(A) = \sum_i \sum_{j \in C_i} p_{ij}, \quad (5)$$

where $C_i = \{j | c_i = c_j\}$ denotes the set of points in the same class as i. By differentiating f, the objective function can be maximized through gradient search for optimal A. After obtaining the projection matrix A, we update all the visual features of $v_i \in \mathbf{D}_\mathcal{I}$ from $\vec{f_i}$ to $A\vec{f_i}$, and re-cluster the visual primitives based on their new features $A\vec{f_i}$.

C. Clustering of Visual Collocations

The discovered visual collocations may not be complete patterns. There is redundancy among them as well. Give a pattern $\mathcal{H} = \{A, B, C\}$, it is possible to obtain many incomplete visual collocations such as $\{A, B\}$, $\{A, C\}$, $\{B, C\}$ due to image noises and quantization errors. Therefore we need to handle this problem.

If two visual collocations $\mathcal{P}_i$ and $\mathcal{P}_j$ are correlated, their transaction set $\mathbf{T}(\mathcal{P}_i)$ and $\mathbf{T}(\mathcal{P}_j)$ (Eq. 1) should also have a large overlap [43], implying that they may be generated from the same pattern $\mathcal{H}$. As a result, $\forall i, j \in \mathbf{\Psi}$, we can measure their similarity $s(i, j)$, which depend not only on their frequencies $\hat{frq}(\mathcal{P}_i)$ and $\hat{frq}(\mathcal{P}_j)$, but also the correlation between their transaction set $\mathbf{T}(\mathcal{P}_i)$ and $\mathbf{T}(\mathcal{P}_j)$. We apply the Jaccard distance to estimate $s(i, j)$ [50]:

$$s(i,j) = \exp^{\frac{1}{1 - \frac{|\mathbf{T}(\mathcal{P}_i) \cap \mathbf{T}(\mathcal{P}_j)|}{|\mathbf{T}(\mathcal{P}_i) \cup \mathbf{T}(\mathcal{P}_j)|}}}. \quad (6)$$

Given a collection of visual collocations $\mathbf{\Psi} = \{\mathcal{P}_i\}$ and their pair-wise similarity $s(i, j)$, we cluster them into K classes using normalized cut [51]. Each class $\mathcal{H}_j = \{\mathcal{P}_i\}_{i=1}^{|\mathcal{H}_j|}$ is a group of visual collocations, called a *visual part*. Each $\mathcal{P}_i \in \mathcal{H}$ is an variation of $\mathcal{H}$, due to the image noises or the quantization error of the codebook. By grouping visual collocations, we end up with a collection of visual parts:

$\mathbf{H} = \{\mathcal{H}_i\}$. These visual parts can be further constructed to recover the whole visual object.

Algorithm 2: Main Algorithm

input : Image dataset $\mathbf{D}_\mathcal{I}$,
ϵ or K for searching spatial ϵ-NN or K-NN,
parameter: (θ, τ, γ),
number of semantic patterns: $|\mathbf{H}|$,
number of maximum iteration l

output: A set of semantic patterns: $\mathbf{H} = \{\mathcal{H}_i\}$

Init: Get visual item codebook $\mathbf{\Omega}^0$ and induced transaction DB $\mathbf{T}^0_\mathbf{\Omega}$; $i \longleftarrow 0$;
while $i < l$ **do**
  $\mathbf{\Psi}^i$ = `MIM(`$\mathbf{T}^i_\mathbf{\Omega}$`);` `/* visual collocation mining */`
  $\mathbf{\Omega}^i_+ = \cup_j \mathcal{P}_j$, where $\mathcal{P}_j \in \mathbf{\Psi}^i$;
  A^i = `NCA (`$\mathbf{\Omega}^i_+$,$\mathbf{T}^i_\mathbf{\Omega}$`);` `/* get new metric */`
  Update $\mathbf{\Omega}^i$ and $\mathbf{T}^i$ based on A^i`/* re-clustering */`
  if *little change of* $\mathbf{\Omega}^i$ **then**
    break;
  $i \longleftarrow i + 1$
$\mathbf{S}$ = `GetSimMatrix (`$\mathbf{\Psi}^i$`);`
$\mathbf{H}$ = `NCut (`$\mathbf{S}$, $|\mathbf{H}|$`);` `/* pattern summarization */`
Return $\mathbf{H}$;

VI. Experiments

A. Dataset Description

Given a large image dataset $\mathbf{D}_\mathcal{I} = \{\mathcal{I}_i\}$, we first extract the PCA-SIFT points [48] in each image $\mathcal{I}_i$ and treat these interest points as the visual primitives. We resize all images by the factor of $2/3$. The feature extraction is on average 0.5 seconds per image. Multiple visual primitives can be located at the same spatial location, but with various scales and orientations. Each visual primitives is represented as a 35-*d* feature vector after principal component analysis. Then k-means algorithm is used to cluster these visual features into a visual item codebook $\mathbf{\Omega}$. We select three categories from the Caltech 101 database [26] for the experiments: faces (435 images from 23 persons), cars (123 images of different cars), and airplanes (800 images of different airplanes). We set the parameters for MIM as: $\theta = \frac{1}{4}|\mathbf{D}_\mathcal{I}|$, where $|\mathbf{D}_\mathcal{I}|$ is the total number of images, and τ is associated with the confidence level of 0.90. Instead of setting threshold γ, we select the top phrases by ranking their $L(\mathcal{P})$ values. We set visual item codebook size $|\mathbf{\Omega}| = 160$, 500, and 300, for the car, face, and airplane categories, respectively. For generating the transaction databases $\mathbf{T}$, we set $K = 5$ for searching spatial K-NN to compose each transaction. All the experiments were conducted on a Pentium-4 3.19GHz PC with 1GB RAM running window XP.

B. Evaluation of Visual Collocation Patterns

We use two metrics to evaluate the discovered visual collocations: (1) the precision of $\mathbf{\Psi}$: ρ^+ denotes the percentage of discovered visual collocations $\mathcal{P}_i \in \mathbf{\Psi}$ that are located in the foreground objects, and (2) the precision of $\mathbf{\Omega}^-$: ρ^- denotes the percentage of meaningless items $W_i \in \mathbf{\Omega}^-$ that are located in the background. Fig. 5 illustrates the concepts of our evaluation. In the ideal situation, if $\rho^+ = \rho^- = 1$, then every $\mathcal{P}_i \in \mathbf{\Psi}$ is associated with the interesting object, *i.e.* located inside the object bounding box; while all meaningless items $W_i \in \mathbf{\Omega}^-$ are located in the backgrounds. In such a case, we can precisely discriminate the frequently appeared foreground objects from the clutter backgrounds, through an unsupervised learning. Finally, we use retrieval rate η to denote the percentage of retrieved images that contain at least one visual collocation. Since for the airplane category, most airplanes appear in the clean background, its precision will be high because there is much less interest points located in the background. To make a fair evaluation, we thus only test the accuracy of the car and face categories, where the objects are always located in a cluttered background. We only test the airplane category for the discovery of high-level visual patterns.

Fig. 5. Evaluation of visual collocations mining. The highlight bounding box (yellow) represents the foreground region where the interesting object is located. In the idea case, all the MI $\mathcal{P}_i \in \mathbf{\Psi}$ should locate inside the bounding boxes while all the meaningless items $W_i \in \mathbf{\Omega}^-$ are located outside the bounding boxes.

In Table II, we present the visual collocations from the car database. The first row indicates the number of visual collocations ($|\mathbf{\Psi}|$), selected by their $L(\mathcal{P})$. When selecting more visual collocations, the precision score of $\mathbf{\Psi}$, 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 of ρ^+ indicates that the discovered visual collocation are associated with the foreground objects. It is also noted that meaningful item codebook $\mathbf{\Omega}^+$ is only a small subset with respect to $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = 160$). This implies that most visual items do not belong to the foreground objects. They are noisy items from the backgrounds.

TABLE II
Precision score ρ^+ and retrieval rate η for the car database, corresponding to various sizes of $\mathbf{\Psi}$. 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	0.96	0.91	0.88	0.86	0.86

We further compare three types of criteria for selecting visual collocations $\mathcal{P}$ into $\mathbf{\Psi}$, against the baseline of selecting the individual visual items $W_i \in \mathbf{\Omega}$ to build $\mathbf{\Psi}$. The three visual collocation selection criteria are: (1) occurrence frequency: $\hat{frq}(\mathcal{P})$ (2) T-score (Eq. 7) (only select the second order itemsets, $|\mathcal{P}| = 2$) and (3) likelihood ratio: $L(\mathcal{P})$ (Eq. 2).

The comparison results are presented in Fig. 6. It shows how ρ^+ and ρ^- vary with increasing size of $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 1, ..., 30$). In general, the larger the size of $\mathbf{\Psi}$, the lower the precision ρ^+, but the higher the precision ρ^-, because $\mathbf{\Omega}^-$ becomes smaller and purer. Meanwhile, we notice that all of the three criteria perform significantly better than the baseline of choosing the most frequent individual items as meaningful patterns. This is not surprising because frequent items $W_i \in \mathbf{\Omega}$ correspond to common features (*e.g.* corners), therefore they appear in both foreground objects and clutter backgrounds and express little discriminative ability.

Finally, Among the three criteria, occurrence frequency $\hat{frq}(\mathcal{P})$ performs worse than the other two criteria, which further demonstrates that not all frequent itemsets are meaningful patterns. It is also shown from Fig. 6 that when only selecting a few number of visual collocations, *i.e.* $\mathbf{\Psi}$ has a small size, all the three criteria yield similar performances. However, when more visual collocations are added, the proposed likelihood ratio test method performs better than the other two, which shows our MIM algorithm can discover meaningful visual patterns.

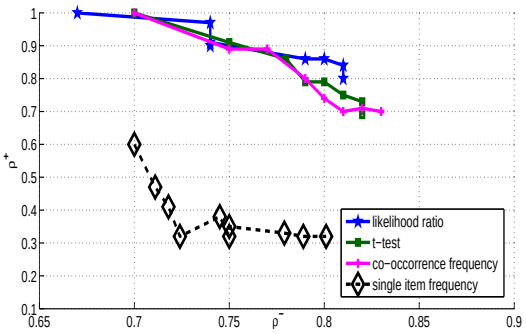

Fig. 6. Performance comparison by applying three different criteria to select visual collocations, also with the baseline of selecting most frequent individual items to build $\mathbf{\Psi}$.

By taking advantage of the FP-growth algorithm for closed FIM, our pattern discovery is very efficient. As presented in Table III, it costs around 17.4 seconds for discovering visual collocations from the face database containing over $60,000$ transactions. Compared with top-down models such as topic discovery for visual pattern discovery, such a bottom-up pattern mining method is more efficient. Moreover, as we only select a very small subset of visual primitives for subspace learning, *e.g.*, NCA, the computational cost for metric learning is acceptable.

TABLE III
CPU COMPUTATIONAL COST FOR VISUAL COLLOCATIONS MINING IN FACE DATABASE, WITH $|\mathbf{\Psi}| = 30$.

# images $\lvert\mathbf{D}_{\mathcal{I}}\rvert$	# transactions $\lvert\mathbf{T}\rvert$	closed FIM [47]	MIM Alg.1
435	62611	1.6 sec	17.4 sec

C. Refinement of visual item codebook

To implement NCA for metric learning, we select 5 visual collocations from $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$). There are in total less than 10 items shared by these 5 visual collocations for both face and car categories, *i.e.* $|\mathbf{\Omega}^+| < 10$. For the foreground class, we select all of the instances of top five visual collocations as training samples. Considering the large number of background items $\mathbf{\Omega}^-$, we only select a small number of them which have higher probability to generated from the background. Specifically, from the visual primitives belonging to $\mathbf{\Omega}^-$, we only select those uncommon visual primitives that cannot find many matches in the rest of images. The number of negative training samples is selected according to the number of positive training samples, to make balanced training examples.

After learning a new metric using NCA, the inter-class distance is enlarged while the intra-class distance is reduced among the training samples. We then reconstruct the visual item codebook $\mathbf{\Omega}$ using k-means clustering, and perform the visual collocation discovery again. To compare the visual collocations using the original codebook and the refined one, Fig. 7 shows the results in both car and face datasets. It can be seen that the precision ρ^+ of visual collocations is improved with the refined codebook.

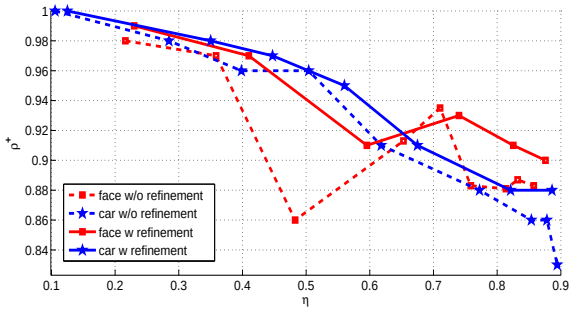

Fig. 7. Comparison of visual item codebook before and after self-supervised refinement.

D. Clustering of visual collocations

For each object category, we select the top-10 visual collocations by their $L(\mathcal{P})$ (Eq. 2). All of the discovered visual collocations are the second-order, third-order, or fourth-order itemsets. Each collocation pattern is a local composition of visual items. These items function together as a single visual lexical entity. By further clustering these top-10 visual collocation ($|\mathbf{\Psi}| = 10$) into visual parts, the clustering results are presented in Fig. 8 and Fig. 9, for the face and car categories respectively. For the face category, we cluster visual collocations into $|\mathbf{H}| = 6$ visual parts. Five of the six visual parts are semantically meaningful: (1) left eye (2) between eyes (3) right eye (4) nose and (5) mouth. All of the discovered visual parts have very high precision. It is interesting to note that left eye and right eye are discovered separately, due to the differences of their visual appearances. The other visual part that is not associated with the face. It corresponds to corners from computers and windows in the office environment. For the car category, we cluster them into $|\mathbf{H}| = 2$ visual parts: (1) car wheels and (2) car bodies (mostly windows containing strong edges). For the airplane category, we also cluster them into $|\mathbf{H}| = 3$ visual parts, while two of them are semantically meaningful: (1) airplane heads and (2) airplane wings.

To evaluate the clustering of visual collocations, we apply the precision and recall scores defined as follows: Recall = # detects / (# detects + # miss detects) and Precision = # detects /(# detects + # false alarms). For each visual part, the ground truths are manually labeled in the images. We evaluate both the car and face categories in Fig. 8 and Fig. 9. It can be seen that the discovered visual parts are of high precision but low recall rate. The high precision rate validates the quality of the discovered patterns. The miss detection of many visual parts is mainly caused by the interest point miss detection and the quantization errors in the codebook.

E. From visual parts to visual objects

The discovered visual parts in Fig. 8 and Fig 9 can be further composed into a high-level pattern that describes the whole object. We treat each part $\mathcal{H}_i$ as a high-level item and build another codebook $\mathbf{\Omega}' = \mathbf{H} = \{\mathcal{H}_i\}$. Based on the new codebook $\mathbf{\Omega}'$, each image is composed of a few high-level items $\mathcal{H}_i$, *i.e.*, visual parts. Thus each image generates a single transaction for data mining. Then we perform visual collocation mining on these transactions again, but with a much smaller codebook $\mathbf{\Omega}'$. In general, our approach can be easily extended to a multi-level discovery of visual patterns. Let l denotes the level, a typical semantic pattern in layer l, $\mathcal{H}_i^l \subset \Omega^l$ is a composition of simpler subpatterns (items) from the lower level $\Omega^{l-1} = \{\mathcal{H}_i^{l-1}, \mathcal{H}_2^{l-1}, ..., \mathcal{H}_{M_{l-1}}^{l-1}\}$ which, in turn, are built from even simpler subpatterns $\Omega^{l-2} = \{\mathcal{H}_1^{l-2}, \mathcal{H}_2^{l-2}, ..., \mathcal{H}_{M_{l-2}}^{l-2}\}$. The most primitive subpatterns $\Omega^1 = \{\mathcal{H}_1^1, \mathcal{H}_2^1, ..., \mathcal{H}_{M_1}^1\}$ are the primitive visual items. Such a multi-level pattern discovery can help to reveal the hierarchical structure of visual patterns and is computationally efficient.

To explain the procedure of detecting visual parts, Fig. 10, Fig. 12, and Fig 14 show some exemplar results, for the car, face, and airplane categories, respectively. Visual primitives are highlighted as the green circles in each image. Then the discovered visual collocations are highlighted by the bounding boxes. The colors of the visual primitives inside the bounding box distinguish different types of the visual parts. Once the visual parts are obtained, Fig. 11, Fig. 13 and Fig. 15 show how these visual parts can be further constructed to represent the whole object, for the car, face, and airplane categories, respectively. Since a visual part corresponds to a group of similar visual collocations $\mathcal{H} = \{\mathcal{P}_j\}$, in each image, we treat any occurrence of $\mathcal{P}_j \in \mathcal{H}$ as the occurrence of $\mathcal{H}$. For example, according to the visual parts in Fig. 8, there are 5 visual parts in the face category: $\mathbf{\Omega}' = \{left\ eye, right\ eye, between\ eyes, nose, mouth\}$ in the face category. While for the car and airplane categories, there are only two visual parts: $\mathbf{\Omega}' = \{car\ wheel, car\ body\}$ for the car category and $\mathbf{\Omega}' = \{airplane\ head, airplane\ wing\}$ for the airplane category, respectively. It is interesting to notice that although we do not reinforce geometrical relationship among the visuals parts, the geometrical configuration among them is consistent among different images.

VII. Conclusion

Text-based data mining techniques are not directly applicable to image data, which exhibit much larger variabilities and uncertainties. To leap from text data mining to image data mining, we present a systematic study on mining visual collocation patterns from images. A new criterion for discovering visual collocations based on traditional FIM is proposed. Such visual collocations are statistically more interesting than the frequent itemsets. To obtain a better visual codebook, a self-supervised subspace learning method is proposed by applying the discovered visual collocations as supervision to learn a better similarity metric through subspace learning. By further clustering these visual collocations (incomplete sub-patterns), we successfully extract semantic visual patterns despite the intra-pattern variations and the cluttered backgrounds. As a pure data-driven bottom-up approach, our method does not depend on a top-down generative model in discovering visual patterns. It is computationally efficient and requires no prior knowledge of the visual pattern. Our future work will consider how to apply the discovered visual collocations for image search and categorization.

Appendix A
Pair-wise Dependency Test

If W_i, $W_j \in \mathbf{\Omega}$ are independent, then the process of randomly generating the pair $\{W_i, W_j\}$ in a transaction $\mathcal{T}_i$ is a (0/1) Bernoulli trial with probability $P(W_i, W_j) = P(W_i)P(W_j)$. According to the central limit theory, as the number of trials (transaction number N) is large, the Bernoulli distribution can be approximated by the Gaussian random variable x, with mean $\mu_x = P(W_i)P(W_j)$. At the same time, we can measure the average frequency of $\{W_i, W_j\}$ by counting its real instance number in $\mathbf{T}$, such that $P(W_i, W_j) = \hat{frq}(W_i, W_j)/\hat{N}$. In order to verify if the observation $P(W_i, W_j)$ is drawn from the Gaussian distribution x with mean μ_x, the following T-score is calculated; S^2 is the estimation of variance from the observation data.

$$\begin{aligned} t(\{W_i, W_j\}) &= \frac{P(W_i, W_j) - \mu_x}{\sqrt{\frac{S^2}{\hat{N}}}} \\ &= \frac{P(W_i, W_j) - P(W_i)P(W_j)}{\sqrt{\frac{P(\{W_i, W_j\})(1-P(W_i, W_j))}{\hat{N}}}} \\ &\approx \frac{\hat{frq}(\{W_i, W_j\}) - \frac{1}{\hat{N}}\hat{frq}(W_i)\hat{frq}(W_j)}{\sqrt{\hat{frq}(\{W_i, W_j\})}}. \end{aligned}$$

References

[1] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 488–495, 2004.

[2] J. Yuan, Z. Li, Y. Fu, Y. Wu, and T. S. Huang, "Common spatial pattern discovery by efficient candidate pruning," in *Proc. IEEE Conf. on Image Processing*, 2007.

[3] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *TPAMI*, 2010.

[4] G. Zhao and J. Yuan, "Mining and cropping common objects from images," in *Proc. ACM Multimedia*, pp. 975–978, 2010.

[5] G. Zhao, J. Yuan, J. Xu, and Y. Wu, "Discovering the thematic object in commercial videos," *IEEE Multimedia Magazine*, vol. 18, no. 3, pp. 56–65, 2011.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

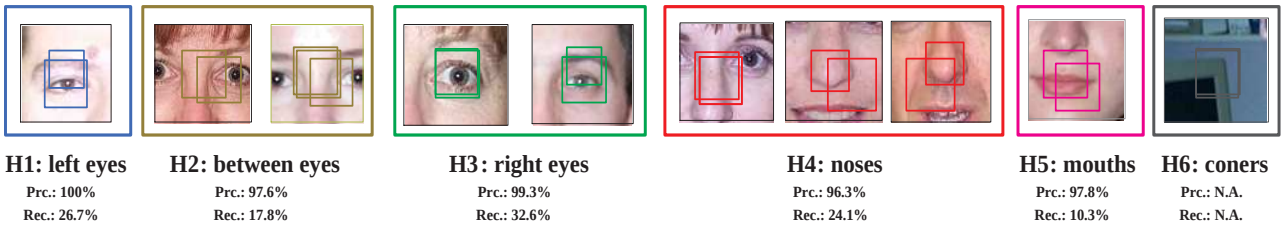

Fig. 8. Face category: top-10 visual collocations $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and their clustering results into five visual parts ($|\mathbf{H}| = 6$). Each image patch shows a visual collocation $\mathcal{P}_i \in \mathbf{\Psi}$. The rectangles in the images are visual primitives (*e.g.* PCA-SIFT interest points at their scales). Every visual collocation, except for the 3_{rd} one, is composed of 2 items. The 3_{rd} visual collocation is a high-order one composed of 3 items. For each visual collocation, we show its precision and recall rates.

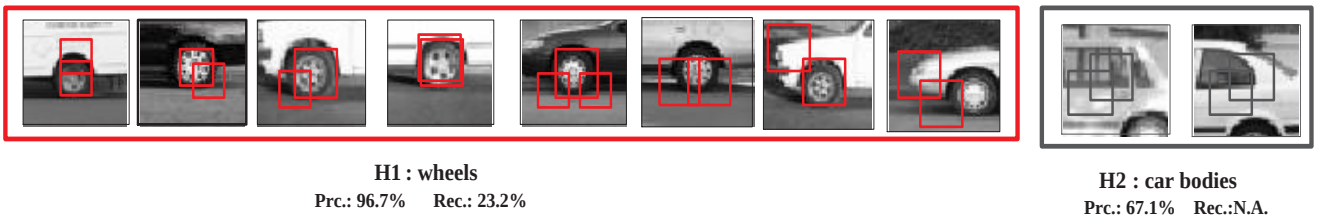

Fig. 9. Car category: top-10 visual collocations $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and their clustering results into two visual parts ($|\mathbf{H}| = 2$). Each image patch shows a visual collocation $\mathcal{P}_i \in \mathbf{\Psi}$. The rectangles in the images are visual primitives. Every visual collocation is composed of 2 items, except for the 5_{th} itemset, which is composed of 3 items. For each visual collocation, we show its precision and recall rates.

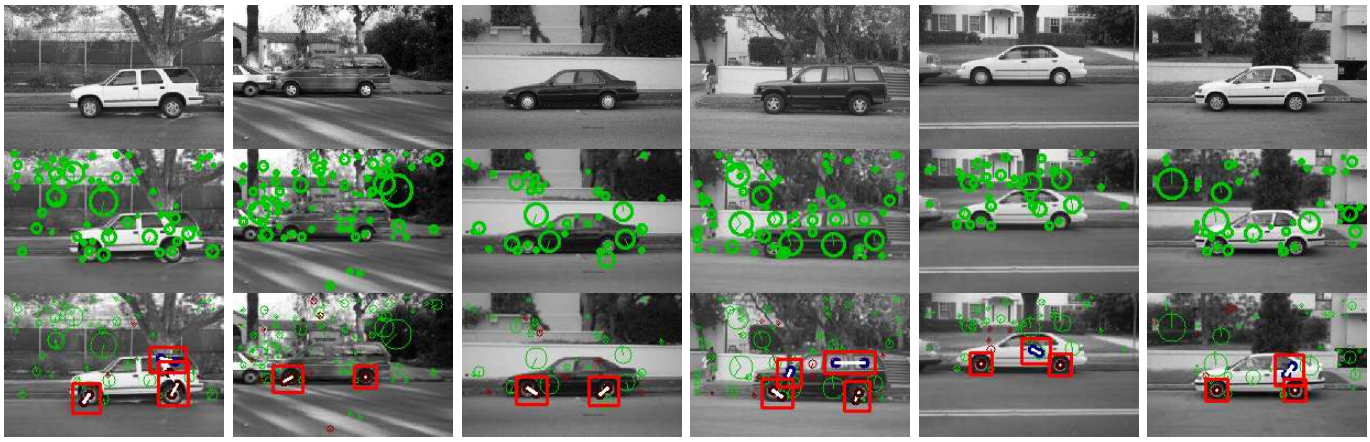

Fig. 10. Car category: from visual primitives to visual parts. The first row shows the original images. The second row shows their visual primitives (PCA-SIFT points). Each green circle denotes a visual primitive with corresponding location, scale and orientation. It is possible that two visual primitives are located at the same position but with different scales. The third row shows the visual collocations. Each red rectangle shows a visual collocation. The colors of the visual primitives inside the red rectangle distinguish different types of visual parts. For example, wheels are red and car bodies are blue.

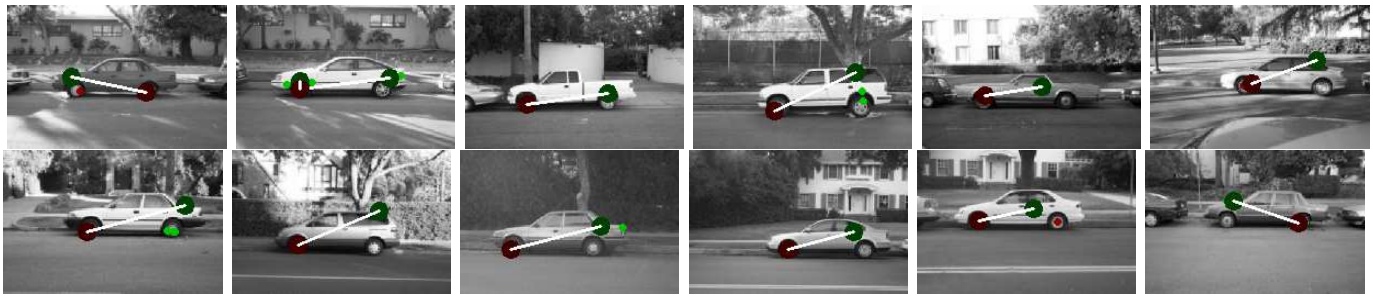

Fig. 11. Car category: from visual parts to visual object. The two visual parts are discovered in Fig. 9. Different colors show different visual parts: car wheels are red ($\mathcal{H}_1$) and car bodies are green ($\mathcal{H}_2$). Different types of cars are abstracted by a composition of car-wheel and car-body.

[7] 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*, vol. 65, no. 1-2, pp. 43–72, 2005.

[8] 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*, pp. 1605–1614, 2006.

[9] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: a general approach," *IEEE Transaction on*

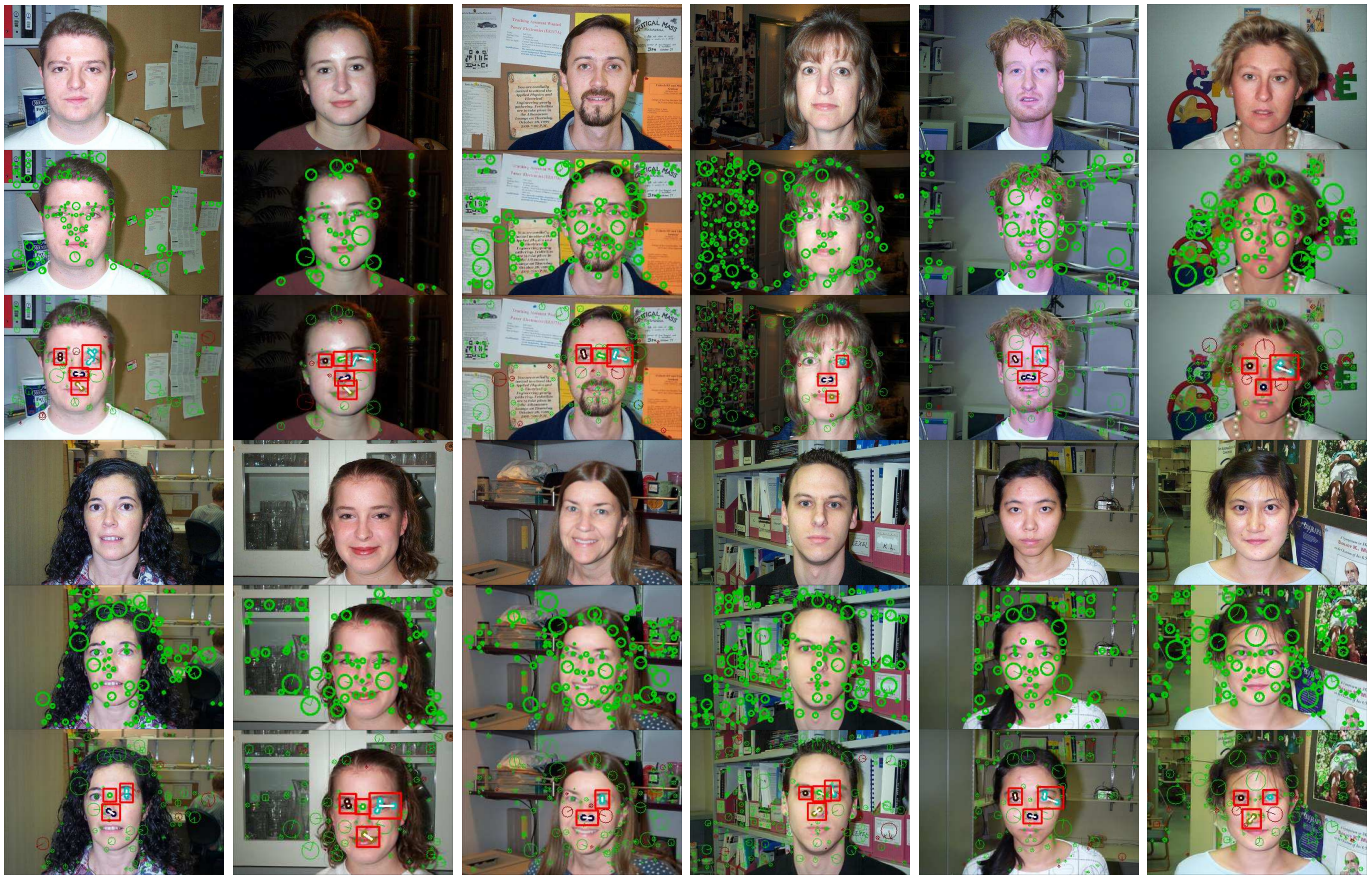

Fig. 12. Face category: from visual primitives to visual parts. The first and forth rows show the original images. The second and fifth rows show their visual primitives. Each green circle denotes a visual primitive with corresponding location, scale and orientation. The third and sixth rows show the visual collocations. Each red rectangle shows a visual collocation. The colors of the visual primitives inside the red rectangle distinguish different types of visual parts, *e.g.* green primitives are between eyes.

Fig. 13. Face category: from visual parts to visual object. Five visual parts are discovered in Fig. 9. Different colors show different visual parts. Different faces are abstracted by a composition of the visual parts.

Knowledge and Data Engineering, vol. 16, no. 12, pp. 1472–1485, 2004.

[10] X. Zhang, N. Mamoulis, D. W. Cheung, and Y. Shou, "Fast mining of spatial collocations," in *Proc. ACM SIGKDD*, 2004.

[11] W. Hsu, J. Dai, and M. L. Lee, "Mining viewpoint patterns in image databases," in *Proc. SIGKDD*, 2003.

[12] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 1470 – 1477, 2003.

[13] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop on European Conf. on Compsuter Vision*, pp. 1–22, 2004.

[14] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *International Journal of Computer Vision*, vol. 88, pp. 284–302, 2010.

[15] L. Cao and F.-F. Li, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 1–8, 2007.

[16] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proc. of the IEEE*, vol. 96, no. 4, pp. 548–566, 2008.

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 370–377, 2005.

[18] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 604–610, 2005.

[19] 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*, pp. 1–8, 2007.

[20] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. ACM Multimedia*, 2009.

[21] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Visual synset: a higher-level visual representation for object-based image retrieval," *The Visual Computer*, vol. 25, no. 1, pp. 13–23, 2009.

[22] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Proc. IEEE Conf. on*

Fig. 14. Airplane category: from visual primitives to visual parts. The colors of the visual primitives inside the rectangle distinguish different types of visual parts.

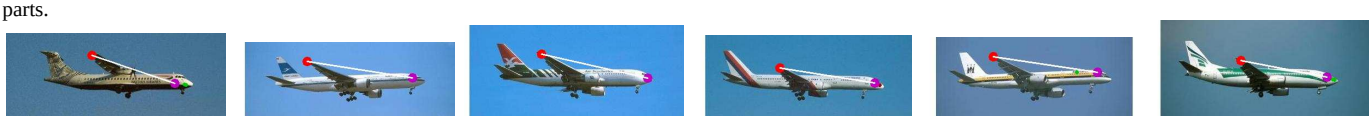

Fig. 15. Airplane category: from visual parts to visual object. Two visual parts are discovered, highlighted by two different colors.

Computer Vision and Pattern Recognition, pp. 1597–1604, 2006.

[23] P. Hong and T. S. Huang, "Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs," *Discrete Applied Mathematics*, pp. 113–135, 2004.

[24] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[25] G. Bouchard and B. Triggs, "Hierarchical part-based visual object categorization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 710 – 715, 2005.

[26] 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*, pp. 264 – 271, 2003.

[27] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2158–2174, 2008.

[28] Y. N. Wu, Z. Si, H. Gong, , and S.-C. Zhu, "Learning active basis model for object detection and recognition," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 198–235, 2010.

[29] 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*, pp. 1–8, 2007.

[30] J. Yuan, Y. Wu, and M. Yang, "From frequent itemsets to semantically meaningful visual patterns," in *Proc. ACM SIGKDD*, pp. 864–873, 2007.

[31] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *Proc. IEEE Intl. Conf. on Computer Vision*, pp. 2122 – 2129, 2009.

[32] S. Kim, X. Jin, and J. Han, "Sparclus: Spatial relationship pattern-based hierarchical clustering," in *Proc. 2008 SIAM Int. Conf. on Data Mining*, 2008.

[33] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, "Building contextual visual vocabulary for large-scale image applications," in *Proc. ACM Multimedia*, 2010.

[34] J. Han, J. Pei, and W. Yi, "Mining frequent patterns without candidate generation," in *Proc. SIGMOD*, 2000.

[35] R.Agrawal, T.Imielinski, and A.Swami, "Mining association rules between sets of items in large databases," in *Proc. SIGMOD*, 1993.

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

[37] T. Calders and B. Goethals, "Depth-first non-derivable itemset mining," in *Proc. SIAM International Conference on Data Mining*, 2005.

[38] N. Pasquiera, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proc. ICDT*, 1999.

[39] A. Siebes, J. Vreeken, and M. van Leeuwen, "Item sets that compress," in *Proc. SIAM International Conference data mining (SDM)*, 2006.

[40] C. Yang, U. Fayyad, and P. S.Bradley, "Efficient discovery of error-tolerant frequent itemsets in high dimensions," in *Proc. ACM SIGKDD*, 2001.

[41] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in *Proc. ACM SIGKDD*, 2004.

[42] J. Liu, S. Paulsen, X. Sun, W. Wang, A. Nobel, and J. Prins, "Mining approximate frequent itemsets in the presence of noise: Algorithm and analysis," in *Proc. SIAM International Conference on Data Mining*, 2006.

[43] X. Yan, H. Cheng, J. Han, and D. Xin, "Summarizing itemset patterns: a profile-based approach," in *Proc. ACM SIGKDD*, 2005.

[44] C. Wang and S. Parthasarathy, "Summarizing itemset patterns using probabilistic models," in *Proc. ACM SIGKDD*, 2006.

[45] D. Xin, H. Cheng, X. He, and J. Han, "Extracting redundancy-aware top-k patterns," in *Proc. ACM SIGKDD*, 2006.

[46] J. Han and M. Kamber, "Data mining: Concepts and techniques," in *Morgan Kaufmann Publishers.*, 2000.

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

[48] 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*, pp. 506–513, 2004.

[49] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *Proc. of Neural Information Processing Systems*, 2004.

[50] 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.

[51] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

PLACE PHOTO HERE

Junsong Yuan (M'08) is currently a Nanyang Assistant Professor at Nanyang Technological University. He received his Ph.D. from Northwestern University, USA and his M.Eng. from National University of Singapore. Before that, he graduated from the special program for the gifted young in Huazhong University of Science and Technology, P.R.China. He was a research intern at Microsoft Research Redmond, Kodak Research Laboratories, Rochester, and Motorola Applied Research Center, Schaumburg, USA. His current research interests include computer vision, video analytics, multimedia search and mining, vision-based human computer interaction, biomedical image analysis, etc. He is the program director of Video Analytics in the Infocomm Center of Excellence at Nanyang Technological University.

Junsong Yuan received the Outstanding Ph.D. Thesis award from the EECS department in Northwestern University, and the Doctoral Spotlight Award from IEEE Conf. Computer Vision and Pattern Recognition Conference (CVPR'09). He was also a recipient of the elite Nanyang Assistant Professorship in 2009. He has filed 3 US patents and is a member of IEEE.

PLACE
PHOTO
HERE

Ying Wu (SM'06) received the B.S. from Huazhong University of Science and Technology, Wuhan, China, in 1994, the M.S. from Tsinghua University, Beijing, China, in 1997, and the Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001.

From 1997 to 2001, he was a research assistant at the Beckman Institute for Advanced Science and Technology at UIUC. During summer 1999 and 2000, he was a research intern with Microsoft Research, Redmond, Washington. In 2001, he joined the Department of Electrical and Computer Engineering at Northwestern University, Evanston, Illinois, as an assistant professor. He is currently an associate professor of Electrical Engineering and Computer Science at Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He serves as associate editors for IEEE Transactions on Image Processing, SPIE Journal of Electronic Imaging, and IAPR Journal of Machine Vision and Applications. He received the Robert T. Chien Award at UIUC in 2001, and the NSF CAREER award in 2003. He is a senior member of the IEEE.