From Frequent Itemsets to Semantically Meaningful Visual Patterns

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

ABSTRACT

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. It is not a trivial task to discover meaningful visual patterns in image databases, because the content variations and spatial dependency in the visual data greatly challenge most existing methods. This paper presents a novel approach to coping with these difficulties for mining meaningful visual patterns. Specifically, the novelty of this work lies in the following new contributions: (1) a principled solution to the discovery of meaningful itemsets based on frequent itemset mining; (2) a self-supervised clustering scheme of the high-dimensional visual features by feeding back discovered patterns to tune the similarity measure through metric learning; and (3) a pattern summarization method that deals with the measurement noises brought by the image data. The experimental results in the real images show that our method can discover semantically meaningful patterns efficiently and effectively.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining, Image databases*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms

Keywords

image data mining, meaningful itemset mining, pattern summarization, self-supervised clustering

1. INTRODUCTION AND RELATED WORK

Meaningful patterns can be those that appear frequently, thus an important task for data mining and pattern discovery is to identify repetitive patterns. Frequent itemset mining (FIM) and its extensions [9] [21] [7] have been extensively studied. However, a highly repetitive pattern may not be informative or semantically meaningful. Therefore a more important task is to extract informative and potentially interesting patterns (e.g. semantically meaningful patterns) in 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 [3] [19] [23], approximate patterns [29] [1] [14] and pattern summarization [28] [26] [27]. These data mining techniques may discover meaningful frequent itemsets and represent them in a compact way.

Such research in structured data (e.g., transaction data) and semi-structured data (e.g., text) has aroused our curiosity in finding meaningful patterns in non-structured multimedia data like images and videos [20] [24] [31] [10]. For example, once we can extract some invariant visual primitives such as interest points [15] or salient regions [17] 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" through clustering the high-dimensional features [24] [30], each image is represented by a set of transaction records, with each transaction corresponds to a local image patch and describes its composition of visual primitive classes (items). After that, data mining techniques like FIM can be applied to such a transaction database induced from images for discovering meaningful visual patterns.

Although this idea appears to be quite exciting, the leap from transaction data to images is not trivial, because of two fundamental differences between them. Above all, unlike transaction and text data that are composed of discrete elements without 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 possible, 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 vector quantization [18] or k-means clustering [24]), such "visual items" tend to be much more ambiguous than the case of transaction and text data. Such imperfect clustering of visual items brings large challenges when directly applying traditional data mining methods into image data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'07, August 12–15, 2007, San Jose, California, USA.
Copyright 2007 ACM 978-1-59593-609-7/07/0008 ...$5.00.

In addition to the continuous high-dimensional features, visual patterns have more complex structure than transaction and text pattern. The difficulty of representing and discovering spatial patterns in images prevents straightforward generalization of traditional frequent pattern 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 dependency. Although there exist methods [12] [32] [11] for spatial collocation pattern discovery from geo-spatial data, they cannot be 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 based on FIM may not always output meaningful and informative patterns in the image data.

Given a collection of unlabeled images, the objective of image data mining is to discover (if there is any) semantically meaningful spatial patterns that appear repetitively among the images. For example, given a set of images each of which contains an identical object (e.g. a book or a logo) but with possibly different locations, scales and views, the task is to efficiently discover and locate them in the images. This is a challenging problem because we have no prior knowledge of the object's size, location and pose, or whether such object exists at all. Some existing methods based on graph matching are computational demanding and the solution is prone to local minimum [25] [10]. Thus more efficient and robust algorithm is desirable. In this paper, we aim at an even more challenging problem: given a category of images, for example each image contains a frontal face but from *different* persons, we expect to discover some meaningful patterns like eyes and noses that have semantic meanings and can well interpret the face category. To this end, the following three issues need to be further addressed.

- **Spatial dependency of visual primitives**. To discover frequent patterns in image data using FIM, we can induce a transaction database where each transaction consists of a set of visual items charactering a local image region. However, these induced transactions are not independent as the local patches have spatial overlaps in images. This phenomenon complicates the data mining process for spatial data, 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;
- **Ambiguities in visual items**. The unsupervised clustering of visual primitives is not perfect. A same visual item may convey different semantic meanings. Taking a circle-like visual primitive for example, it can represent a human eye or a car wheel under different context. Thus it brings ambiguities when discovering meaningful patterns. The polysemy word phenomena in text data also appears in images.
- **Incomplete patterns**. There are two kinds of imperfections when translating the image data into transaction data. First of all, the visual primitives can be miss detected in the feature extraction process, due to occlusion of the visual primitive, bad lighting condition or the unreliable detector. Secondly, even a visual primitive is extracted, it can be wrongly labeled into a visual item because of quantization error. These two types of errors will be reflected in the induced transaction database. Performing FIM in the noisy transaction database brings a big obstacle for recovering semantic patterns. For example, a semantically meaningful pattern may be split into a lot of incomplete sub-patterns.

This paper presents a novel approach to discovering semantically meaningful visual patterns from images. By addressing the above three difficulties, our contributions are three-fold:

- *new criteria for meaningful itemset discovery.* The co-occurrence frequency is no longer a sufficient condition for the meaningful collocation patterns in images. A more plausible **meaningful itemset mining** based on likelihood ratio test and traditional FIM is proposed to evaluate the significance of a visual itemset;
- *self-supervised refinement of visual items.* To reduce the ambiguities in visual items, a top-down refinement is proposed by taking advantage of the discovered visual patterns. They serve as self-supervision to tune the metric in the high-dimensional feature space of visual primitives for better visual item clustering.
- *pattern summarization.* To handle the possible imperfections from the image data, a pattern summarization method using normalized cut is proposed to further cluster these incomplete and synonymous meaningful itemsets into semantically-coherent patterns;

2. OVERVIEW

2.1 Notations and basic concepts

Each image in the database is described as 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. 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 of v_i (ϵ-NN) under Euclidean distance. 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 to a visual primitive v_i. 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 $\mathbf{\Omega}$ 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 transfered 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 $\mathbf{\Omega}$ ($|\mathbf{\Omega}| = M$), the induced transaction database $\mathbf{T}$ is defined as follows.

Definition 1. **Induced Transaction Database** *The induced transaction database $\mathbf{T} = \{\mathcal{T}_i\}_{i=1}^{N}$ contains a collection of N transactions with M visual items. A sparse binary matrix $X_{N\times M}$ can represent $\mathbf{T}$, where $x_{ij} = 1$ denotes the i_{th} transaction contains the j_{th} visual item in the codebook and $x_{ij} = 0$ otherwise.*

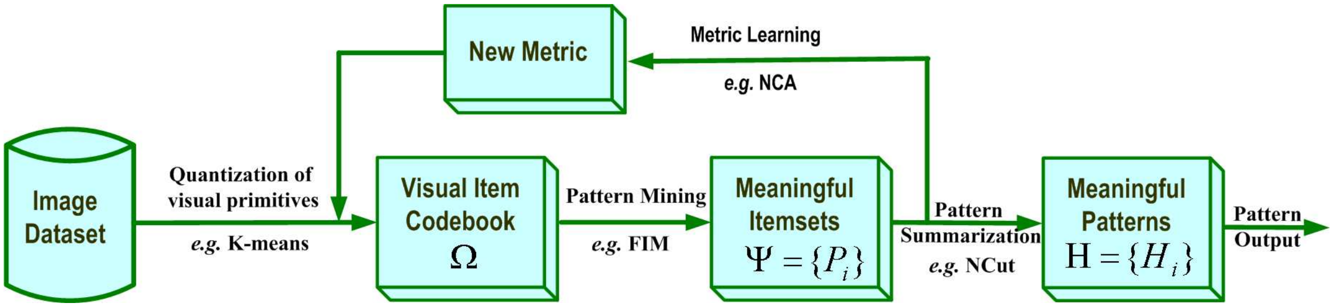

Figure 1: Overview for meaningful visual pattern discovery.

Such an induced transaction database is essentially based on the *centric reference feature model* for mining association rules [32], although collocation pattern models like [12] are also feasible in our approach. Given the visual item codebook $\mathbf{\Omega}$, a 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 denoted as:

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

For a given threshold θ, called a *minimum support*, itemset $\mathcal{P}$ is *frequent* if $frq(\mathcal{P}) > \theta$. If an itemset $\mathcal{P}$ appears frequently, then all of its sub-sets $\mathcal{P}' \subset \mathcal{P}$ will also appear frequently, *i.e.* $frq(\mathcal{P}) > \theta \Rightarrow frq(\mathcal{P}') > \theta$. To eliminate this redundancy, we tend to discover *closed frequent itemsets* [8]. The number of closed frequent itemsets can be 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. Thus this guarantees that no meaningful itemsets will be left out through FIM. The *closed frequent itemset* is defined as follows.

Definition 2. **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 $\mathcal{Q}$, $\mathbf{T}(\mathcal{P} \cup \mathcal{Q}) = \boldsymbol{T}(\mathcal{P}) \cap \mathbf{T}(\mathcal{Q})$, and if $\mathcal{P} \subseteq \mathcal{Q}$ then $\mathbf{T}(\mathcal{Q}) \subseteq \mathbf{T}(\mathcal{P})$.

In this paper we apply the modified FP-growth algorithm [6] to implement the closed FIM. As FP-tree has a prefix-tree structure and can store compressed information of frequent itemset, it can quickly discover all the closed frequent sets from transaction dataset $\mathbf{T}$.

2.2 Overview of our method

We present the overview of our visual pattern discovery method in Fig. 1. In Sec. 3, we present our new criteria for discovering meaningful itemsets $\mathbf{\Psi} = \{\mathcal{P}_i\}$, where each $\mathcal{P}_i \subset \mathbf{\Omega}$ is a meaningful itemset. Further in Sec. 4, a top-down self-supervised clustering method is proposed by feeding back the discovered meaningful itemsets $\mathbf{\Psi}$ to supervise the clustering process. A better visual item codebook $\mathbf{\Omega}$ is then obtained by applying the trained similarity metric for better representing visual primitives. Finally, in Sec. 5, in order to handle the incomplete sub-pattern problem, we propose a pattern summarization method to further cluster those meaningful itemsets (incomplete sub-patterns) and recover the integral semantically meaningful pattern $\mathcal{H}_j$.

3. DISCOVERING MEANINGFUL VISUAL ITEMSETS

3.1 Visual Primitive Extraction

We apply the PCA-SIFT points [13] 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 [13], 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 as a 35-dimensional feature vector $\vec{f_i}$. These visual primitives are 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. 4.

3.2 Meaningful Itemset Mining

Given an image dataset $\mathbf{D}_{\mathcal{I}}$ and its induced transaction database $\mathbf{T}$, the task is to discover the meaningful itemset (MI) $\mathcal{P} \subset \mathbf{\Omega}$ ($|\mathcal{P}| \geq 2$). To evaluate the significance of an itemset $\mathcal{P} \subseteq \mathbf{\Omega}$, 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 compute the likelihood ratio for an itemset $\mathcal{P} \subseteq \mathbf{\Omega}$ 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 an itemset $\mathcal{P} = \{W_i\}_{i=1}^{|\mathcal{P}|}$ can be calculated as:

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

where $P(\mathcal{T}_i|H_1) = \frac{1}{N}$ 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 $\mathcal{T}_i$, such that $P(\mathcal{P}|\mathcal{T}_i, H_1) = 1$, 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 itemset $\mathcal{P}$. In fact if $\mathcal{P} = \{W_A, W_B\}$ is a second-order itemset, then $L(\mathcal{P})$ is the mutual information criterion, *e.g.*, the lift criterion, to test the dependency.

It is worth noting that $L(\mathcal{P})$ may favor high-order itemsets even though they appear less frequently. Table 1 gives 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 1: 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 1, 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:

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

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 with 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 itemset $\mathcal{P}$ ($|\mathcal{P}| > 2$), we perform the *Student 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 8 for details.) A high-order itemset $\mathcal{P}_i$ is meaningful only if all of its pairwise subsets can pass the test individually: $\forall i, j \in \mathcal{P}$, $t(\{W_i, W_j\}) > \tau$, 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 itemset $\mathcal{P}$ 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 itemsets that appear rarely but happen to exhibit strong spatial dependency among items. With these three criteria, a meaningful visual itemset is defined as follows.

Definition 3. **Meaningful Itemset (MI)**
An itemset $\mathcal{P} \subseteq \mathbf{\Omega}$ is (θ, τ, γ)-meaningful 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**: $L(\mathcal{P}) > \gamma$.

3.3 Spatial Dependency

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. 2 illustrates this phenomena where $frq(\mathcal{P})$ contains duplicate counts.

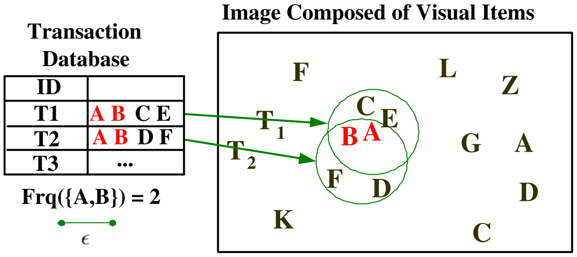

Figure 2: Illustration of the frequency over-counting caused by the spatial overlap of transactions. The itemset $\{A, B\}$ is counted twice by $\mathcal{T}_1 = \{A, B, C, E\}$ and $\mathcal{T}_2 = \{A, B, D, F\}$, although it has only one instance in the image. Namely there is only one pair of A and 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 the 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 meaningful 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 real instance 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 contents in this section, our meaningful itemsets mining (MIM) algorithm is outlined in Algorithm 1.

Algorithm 1: Meaningful Itemset Mining (MIM)

```
input  : Transaction dataset $\mathbf{T}$, MI parameters: $(\theta, \tau, \gamma)$
output: a collection of meaningful itemsets: $\mathbf{\Psi} = \{\mathcal{P}_i\}$
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 $\mathbf{\Psi}$
```

4. SELF-SUPERVISED CLUSTERING OF VISUAL ITEM CODEBOOK

Toward discovering meaningful visual patterns in images, it is critical to obtain optimal visual item codebook $\mathbf{\Omega}$. A bad clustering of visual primitives brings large quantization errors when translating the continuous high-dimensional visual features $\vec{f} \in \mathcal{R}^d$ into discrete labels $W_i \in \mathbf{\Omega}$. Such quantization error reflected in the induced transaction database can affect the data mining results significantly, and thus needs to be minimized.

To improve the clustering results, one possible method is to provide some supervisions, *e.g.* partially label some instances or give some constrains for pairs of instances belonging to the same or different clusters. Such a semi-supervised clustering method has demonstrated its ability in greatly improving the clustering results [2]. However, in our unsupervised clustering setting, there does not exist apparent supervisions. Thus an interesting question is: *is it possible to obtain some supervisions from the completely unlabeled visual primitives ?* Although it is amazing to see the answer is yes, we can explain the reason based on the hidden structure of the image data. It is worth noting that those visual primitives are *not* independently distributed in the images and appearing in the transactions. There are hidden patterns that bring structures in the visual primitive distributions. And such structures can be observed and recovered from the transaction database. For example, if we observe that item W_i always appears together with item W_j in a local region, we can infer that they should be generated from a hidden pattern rather than randomly generated. Each pair of W_i and W_j is thus an instance of the hidden pattern. When such hidden patterns (structures) of the data are discovered through our meaningful itemsets mining, we can apply them as supervision to further improve the clustering results.

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

Definition 4. **Meaningful Item Codebook $\mathbf{\Omega}^+$** *Given a set of meaningful itemsets $\mathbf{\Psi} = \{\mathcal{P}_i\}$, an item $W_i \in \mathbf{\Omega}$ is meaningful if it belongs to any $\mathcal{P} \in \mathbf{\Psi}$: $\exists \mathcal{P} \in \mathbf{\Psi}$, such that $W_i \subset \mathcal{P}$. All of the meaningful items form the meaningful item codebook $\mathbf{\Omega}^+ = \bigcup_{i=1}^{|\mathbf{\Psi}|} \mathcal{P}_i$.*

Based on the concept of meaningful item codebook, the original $\mathbf{\Omega}$ can be partitioned into two disjoined subsets: $\mathbf{\Omega} = \mathbf{\Omega}^+ \cup \mathbf{\Omega}^-$, where $\mathbf{\Omega}^- = \mathbf{\Omega} \backslash \mathbf{\Omega}^+$. For any $\mathcal{P}_i \in \mathbf{\Psi}$, we have $\mathcal{P}_i \subseteq \mathbf{\Omega}^+$ and $\mathcal{P}_i \nsubseteq \mathbf{\Omega}^-$. Since only $\mathbf{\Omega}^+$ can compose MI, $\mathbf{\Omega}^+$ is the meaningful item codebook. Correspondingly we denote $\mathbf{\Omega}^-$ as the *meaningless item codebook*, because an item $W_i \in \mathbf{\Omega}^-$ never appears in any $\mathcal{P}_i \in \mathbf{\Psi}$. In such a case, $W_i \in \mathbf{\Omega}^-$ should be a noisy or redundant item that is not of interests, for example, located in the clutter background of the image.

For each class $W_i \in \mathbf{\Omega}^+$, its positive training set $\mathbf{D}^+_{W_i}$ contains the visual primitives $v_i \in \mathbf{D}_\mathcal{I}$ that satisfy the following two conditions simultaneously:

1. $Q(v_i) = W_i$, where $Q(\cdot)$ is the quantization function from the continuous high-dimensional feature to the discrete item.
2. $v_i \in \mathbf{T}(\mathcal{P}_1) \cup \mathbf{T}(\mathcal{P}_2) \cup ... \cup \mathbf{T}(\mathcal{P}_c)$, where $\mathcal{P}_j$ is the meaningful itemset that contains W_i, namely $\forall j = 1, ..., c, W_i \subset \mathcal{P}_j$.

In summary, not all v_i labeled with W_i are qualified as positive training samples for item class $W_i \in \mathbf{\Omega}^+$. We only choose those visual primitives that can constitute meaningful itemsets. Such visual primitives are very likely generated from the hidden pattern $\mathcal{H}$ that explains the MI.

With these self-labeled training data for each meaningful item $W_i \in \mathbf{\Omega}^+$, we transfer the originally unsupervised clustering problem into semi-supervised clustering. Still, our task is to cluster all the visual primitives $v_i \in \mathbf{D}_\mathcal{I}$. But now some of the visual primitives are already labeled after MIM. Thus many semi-supervised clustering methods are feasible to our task. Here we apply the nearest component analysis (NCA) [5] to improve the clustering results by learning a better Mahalanobis distance metric in the feature space.

Neighborhood Component Analysis (NCA)
Similar to linear discriminative analysis (LDA), NCA targets at learning a global linear projection matrix A for the original features. 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:

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

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):

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

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 $A\vec{f_i}$.

5. PATTERN SUMMARIZATION OF MEANINGFUL ITEMSETS

As discussed before, there are imperfections when translating the image data into transactions. Suppose there exists a hidden visual pattern $\mathcal{H}_j$ (*e.g.* a semantic pattern "eye" in the face category) that repetitively generates a number of instances (eyes of different persons) in the image database. We can certainly observe such meaningful repetitive patterns in the image database, for example, discovering meaningful itemsets $\mathcal{P}_i$ based on Def. 3. However, instead of observing a unique integral pattern $\mathcal{H}_j$, we tend to observe many incomplete sub-patterns with compositional variations due to noise, *i.e.* many synonyms itemsets $\mathcal{P}_i$ that correspond to the same $\mathcal{H}_j$ (see Fig. 3). Again, this can be caused by many reasons, including the missing detection of visual primitives, quantization error of visual primitives, and partial occlusion of the hidden pattern itself. Therefore, we need to cluster those correlated MIs (incomplete sub-patterns) in order to recover the complete pattern $\mathcal{H}$.

According to [28], if two itemsets $\mathcal{P}_i$ and $\mathcal{P}_j$ are correlated, then their transaction set $\mathbf{T}(\mathcal{P}_i)$ and $\mathbf{T}(\mathcal{P}_j)$ (Eq. 1) should

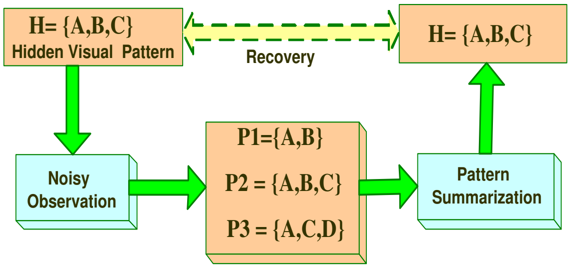

Figure 3: Motivation for pattern summarization. An integral hidden pattern may generate incomplete and noisy instances. The pattern summarization is to recover the unique integral pattern through the observed noisy instances.

also have a large overlap, implying that they may be generated from the same pattern $\mathcal{H}$. As a result, $\forall i, j \in \mathbf{\Psi}$, their similarity $s(i, j)$ should 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 $\mathbf{T}(\mathcal{P}_j)$. Given two itemsets, there are many methods to measure their similarity including KL-divergence between pattern profiles [28], mutual information criterion and Jaccard distance [16]. We apply the Jaccard distance here although others are certainly applicable. The corresponding similarity between two MI $\mathcal{P}_i$ and $\mathcal{P}_j$ is defined as:

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

Based on this, our pattern summarization problem can be stated as follows: given a collection of meaningful itemsets $\mathbf{\Psi} = \{\mathcal{P}_i\}$, we want to cluster them into unjoined K-clusters. Each cluster $\mathcal{H}_j = \{\mathcal{P}_i\}_{i=1}^{|\mathcal{H}_j|}$ is defined as a **meaningful visual pattern**, where $\cup_j \mathcal{H}_j = \mathbf{\Psi}$ and $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset, \forall i, j$. The observed MI $\mathcal{P}_i \in \mathcal{H}$ are instances of the visual pattern $\mathcal{H}$, with possible variations due to imperfections from the images. We propose to apply the normalized cut algorithm [22] for clustering MI. Normalized cut is a well-known algorithm in machine learning and computer vision community. Originally it is applied for clustering-based image segmentation.

Normalized Cut (NCut)
Let $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ denote a fully connected graph, where each vertex $\mathcal{P}_i \in \mathbf{V}$ is an MI, and the weight $s(i, j)$ on each edge represents similarity between two MIs $\mathcal{P}_i$ and $\mathcal{P}_j$. Normalized cut can partition the graph $\mathbf{G}$ into clusters. In the case of bipartition, $\mathbf{V}$ is partitioned into two disjoined sets $\mathbf{A} \cup \mathbf{B} = \mathbf{V}$. The following cut value needs to be minimized to get the optimal partition:

$$Ncut(\mathbf{A}, \mathbf{B}) = \frac{cut(\mathbf{A}, \mathbf{B})}{assoc(\mathbf{A}, \mathbf{V})} + \frac{cut(\mathbf{A}, \mathbf{B})}{assoc(\mathbf{B}, \mathbf{V})}, \qquad (7)$$

where $cut(\mathbf{A}, \mathbf{B}) = \sum_{i\in\mathbf{A}, j\in\mathbf{B}} s(i, j)$ is the cut value and $assoc(\mathbf{A}, \mathbf{V}) = \sum_{i\in\mathbf{A}, j\in\mathbf{V}} s(i, j)$ is the total connection from the vertex set $\mathbf{A}$ to all vertices in $\mathbf{G}$. To minimize the $Ncut$ in Eq. 7, we need to solve the following standard eigenvector problem:

$$\mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{S})\mathbf{D}^{-\frac{1}{2}}z = \lambda z, \qquad (8)$$

where $\mathbf{D}$ is a diagonal matrix with $\sum_j s(i, j)$ on its diagonal and otherwise are 0; $\mathbf{S}$ is a symmetric matrix with $s(i, j)$ its element. The eigenvector corresponding to the second smallest eigenvalue can be used to partition $\mathbf{V}$ into $\mathbf{A}$ and $\mathbf{B}$. In the case of multiple K-class partitioning, the bipartition can be utilized recursively or just apply the eigenvectors corresponding to the $K + 1$ smallest eigenvalues.

We summarize our visual pattern discovery algorithm as follows.

Algorithm 2: Main Algorithm

```
input : Image dataset $\mathbf{D}_\mathcal{I}$,
        $\epsilon$ or $K$ for searching spatial $\epsilon$-NN or K-NN,
        MIM parameter: $(\theta, \tau, \gamma)$,
        number of meaningful patterns: $|\mathbf{H}|$,
        number of maximum iteration $l$
output: A set of meaningful 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($\mathbf{T}^i_\mathbf{\Omega}$);        /*get meaningful itemsets */
    $\mathbf{\Omega}^i_+ = \cup_j \mathcal{P}_j$, 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 $\mathbf{H}$;
```

6. EXPERIMENTS

6.1 Setup

Given a large image dataset $\mathbf{D}_\mathcal{I} = \{\mathcal{I}_i\}$, we first extract the PCA-SIFT points [13] 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 position, 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 two categories from the Caltech 101 database [4] for the experiments: faces (435 images from 23 persons) and cars (123 images of different cars). 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 $|\mathbf{\Omega}| = 160$ and 500 for car and face database respectively when doing k-means clustering. For generating the transaction databases $\mathbf{T}$, we set $K = 5$ for searching spatial K-NN to constitute each transaction. All the experiments were conducted on a Pentium-4 3.19GHz PC with 1GB RAM running window XP.

6.2 Evaluation of Meaningful Itemset Mining

To test whether our MIM algorithm can output meaningful patterns, we want to check if the discovered MI are associated with the frequently appeared foreground objects (*e.g.* faces and cars) while not located in the clutter backgrounds.

The following two criteria are proposed for the evaluation: (1) the precision of $\mathbf{\Psi}$: ρ^+ denotes the percentage of discovered meaningful itemsets $\mathcal{P}_i \in \mathbf{\Psi}$ 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. 4 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 MI.

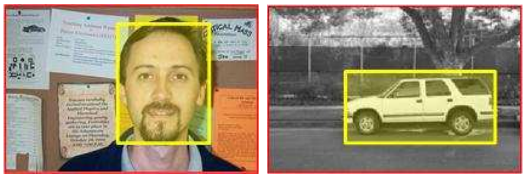

Figure 4: Evaluation of meaningful itemsets mining. The highlight bounding box (yellow) represents the foreground region where the interesting object is located. In the idea case, all the MI $\mathcal{P}_i \in \mathbf{\Psi}$ should locate inside the bounding boxes while all the meaningless items $W_i \in \mathbf{\Omega}^-$ are located outside the bounding boxes.

In Table 2, we present the results of discovering meaningful itemsets from the car database. The first row indicates the number of meaningful itemsets ($|\mathbf{\Psi}|$), selected by their $L(\mathcal{P})$. It is shown that when adding more meaningful itemsets into $\mathbf{\Psi}$, 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 discovered MI 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 actually are not meaningful as they do not constitute the foreground objects. Therefore it is reasonable to get rid of those noisy items from the background. Examples of meaningful itemsets are shown in Fig. 9 and Fig. 10.

Table 2: 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 meaningful itemsets $\mathcal{P}$ into $\mathbf{\Psi}$, against the baseline of selecting the individual visual items $W_i \in \mathbf{\Omega}$ to build $\mathbf{\Psi}$. The three MI selection criteria are: (1) occurrence frequency: $\hat{frq}(\mathcal{P})$ (2) T-score (Eq. 9) (only select the second order itemsets, $|\mathcal{P}| = 2$) and (3) likelihood ratio: $L(\mathcal{P})$ (Eq. 2). The results are presented in Fig. 5. It shows the changes of ρ^+ and ρ^- with increasing size of $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 1, ..., 30$). We can see that all three MI selection criteria perform significantly better than the baseline of choosing the most frequent individual items as meaningful patterns. This demonstrates that FI and MI are more informative features than the singleton items in discriminating the foreground objects from the clutter backgrounds. This is because the most frequent items $W_i \in \mathbf{\Omega}$ usually correspond to common features (*e.g.* corners) which appear frequently in both foreground objects and clutter backgrounds, thus lacking the discriminative power. On the other hand, the discovered MI is the composition of items that function together as a single visual pattern (incomplete pattern though) which corresponds to the foreground object that repetitively appears in the database. 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. 5 that when only selecting a few number of MI, *i.e.* $\mathbf{\Psi}$ has a small size, all the three criteria yield similar performances. However, when more MI are added, the proposed likelihood ratio test method performs better than the other two, which shows our MIM algorithm can discover meaningful visual patterns.

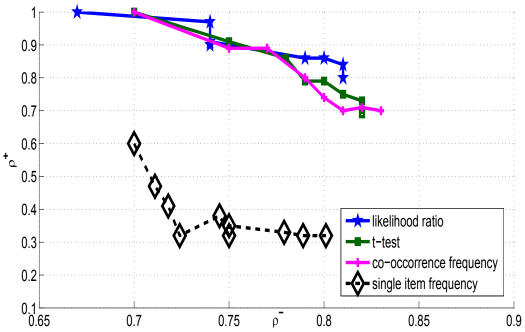

Figure 5: Performance comparison by applying three different meaningful itemset selection criteria, also with the baseline of selecting most frequent individual items to build $\mathbf{\Psi}$.

By taking advantage of the FP-growth algorithm for closed FIM, our pattern discovery is very efficient. It costs around 17.4 seconds for discovering meaningful itemsets from the face database containing over 60,000 transactions (see Table 3). It thus provides us a powerful tool to explore large object category database where each image contains hundreds of primitive visual features.

Table 3: CPU computational cost for meaningful itemsets mining in face database, with $|\mathbf{\Psi}| = 30$.

# images $\lvert\mathbf{D}_\mathcal{I}\rvert$	# transactions $\lvert\mathbf{T}\rvert$	closed FIM [6]	MIM Alg.1
435	62611	1.6 sec	17.4 sec

6.3 Refinement of visual item codebook

To implement NCA for metric learning, we select 5 meaningful itemsets from $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$). There are in total less than 10 items shared by these 5 meaningful itemsets for both face and car categories, *i.e.* $|\mathbf{\Omega}^+| < 10$. For each class, we select the qualified visual primitives as training samples. Our objective of metric learning is to obtain a better representation of the visual primitives, such that the the inter-class distance is enlarged while the intra-class distance is reduced among the self-labeled training samples.

After learning a new metric using NCA, we reconstruct the visual item codebook $\mathbf{\Omega}$ through k-means clustering again, with the number of clusters slightly less than before. The comparison results of the original visual item codebooks and those after refinement are shown in Fig. 6. It can be seen that the precision ρ^+ of $\mathbf{\Psi}$ is improved after refining the item codebook $\mathbf{\Omega}$.

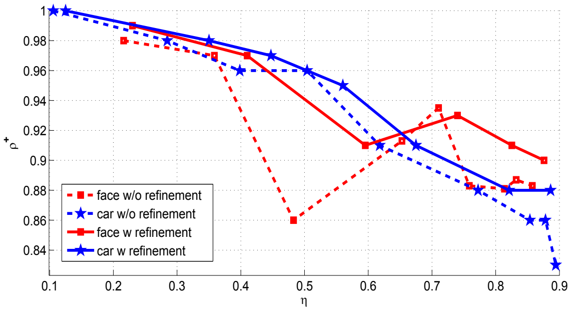

Figure 6: Comparison of visual item codebook before and after self-supervised refinement.

6.4 Visual Pattern Discovery through Pattern Summarization

For both car and face categories, we select the top-10 meaningful itemsets by their $L(\mathcal{P})$ (Eq. 2). All discovered MI are the second-order or third-order itemsets composed of spatially co-located items. We further cluster these 10 MI ($|\mathbf{\Psi}| = 10$) into meaningful visual patterns using the normalized cut. The best summarization results are shown in Fig. 7 and Fig. 8, with cluster number $|\mathbf{H}| = 6$ and $|\mathbf{H}| = 2$ for the face and car category respectively. For the face category, the semantic parts like eyes, noses and mouths are identified by various patterns. For the car category, the wheels and car bodies are identified.

To evaluate our pattern summarization results, we apply the precision and recall scores defined as follows: Recall = # detects / (# detects + # miss detects) and Precision = # detects /(# detects + # false alarms). From Fig. 7 and Fig. 8, it can be seen that the summarized meaningful visual patterns $\mathcal{H}_i$ are associated with semantic patterns with very high precision and reasonably good recall score.

7. CONCLUSION

Traditional data mining techniques are not directly applicable to image data which contain spatial information and are characterized by high-dimensional visual features. To discover meaningful visual patterns from image data, we present a new criterion for discovering meaningful itemsets based on traditional FIM. Such meaningful itemsets are statistically more interesting than the frequent itemsets. By further clustering these meaningful itemsets (incomplete sub-patterns) into complete patterns through normalized cut, we successfully discover semantically meaningful visual patterns from real images of car and face categories.

In order to bridge the gap between continuous high dimensional visual features and discrete visual items, we propose a self-supervised clustering method by applying the discovered meaningful itemsets as supervision to learn a better feature representation. The visual item codebook can thus be increasingly refined by taking advantage of the feedback from the meaningful itemset discovery.

8. APPENDIX

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.

$$t(\{W_i, W_j\}) = \frac{P(W_i, W_j) - \mu_x}{\sqrt{\frac{S^2}{\hat{N}}}} \quad (9)$$

$$= \frac{P(W_i, W_j) - P(W_i)P(W_j)}{\sqrt{\frac{P(\{W_i, W_j\})(1-P(W_i, W_j))}{\hat{N}}}} \quad (10)$$

$$\approx \frac{\hat{frq}(\{W_i, W_j\}) - \frac{1}{\hat{N}}\hat{frq}(W_i)\hat{frq}(W_j)}{\sqrt{\hat{frq}(\{W_i, W_j\})}} \quad (11)$$

Acknowledgment

This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222. The authors thank the anonymous reviewers for their helpful comments.

9. REFERENCES

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

[2] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proc. ACM SIGKDD*, 2004.

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

[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] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. In *Proc. of Neural Information Processing Systems*, 2004.

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

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

[8] J. Han and M. Kamber. Data mining: Concepts and techniques. In *Morgan Kaufmann Publishers.*, 2000.

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

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

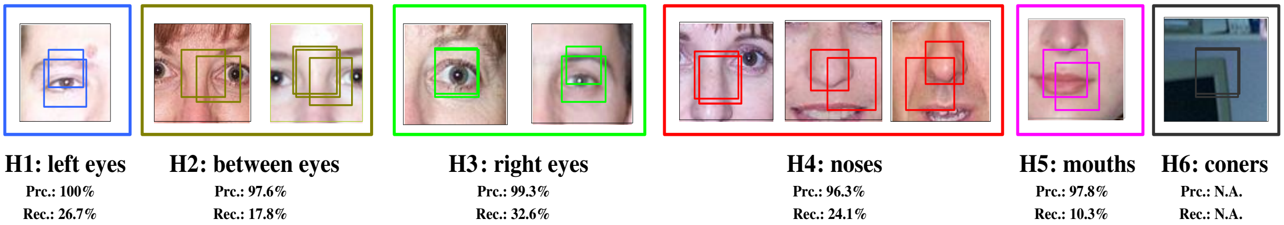

Figure 7: Selected meaningful itemsets $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and their summarization results ($|\mathbf{H}| = 6$) for the face database. Each one of the 10 sub-images contains a meaningful itemset $\mathcal{P}_i \in \mathbf{\Psi}$. The rectangles in the sub-images represent visual primitives (*e.g.* PCA-SIFT interest points at their scales). Every itemset, except for the 3_{rd} one, is composed of 2 items. The 3_{rd} itemset is a high-order one composed of 3 items. Five semantic visual patterns of the face category are successfully discovered: (1) left eye (2) between eyes (3) right eye (4) nose and (5) mouth. All of the discovered meaningful visual patterns have very high precision. It is interesting to note that left eye and right eye are treated as different semantic patterns, possibly due to the differences between their visual appearances. One extra semantic pattern that is not associated with the face is also discovered. It mainly contains corners from computers and windows in the office environment.

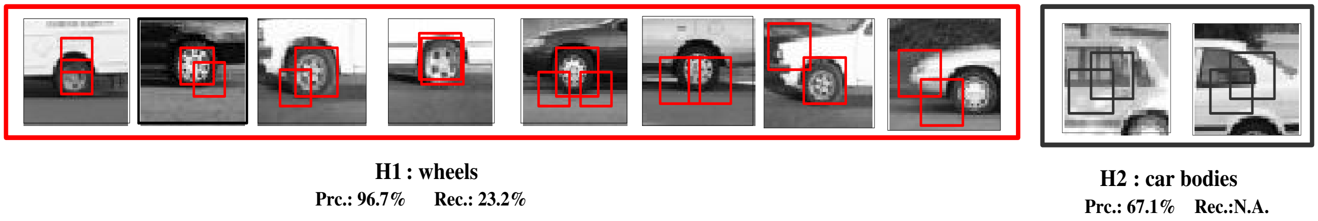

Figure 8: Selected meaningful itemsets $\mathbf{\Psi}$ ($|\mathbf{\Psi}| = 10$) and their summarization results ($|\mathbf{H}| = 2$) for the car database. Two semantic visual patterns that are associated with the car category are successfully discovered: (1) wheels and (2) car bodies (mostly windows containing strong edges). The 5_{th} itemset is a high-order one composed of 3 items.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] J. T. S. Newsam and B. Manjunath. Mining image datasets using perceptual association rules. In *SIAM Workshop on Mining Scientific and Engineering Datasets in conjunction with the SIAM International Conference (SDM)*, 2003.

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

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

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

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

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

[24] 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.

[25] K.-K. Tan and C.-W. Ngo. Common pattern discovery using earth mover's distance and local flow maximization. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2005.

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

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

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

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

[30] 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*, 2007.

[31] O. R. Zaiane, J. Han, and H. Zhu. Mining recurrent items in multimedia with progressive resolution refinement. In *Proc. ICDE*, 2000.

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

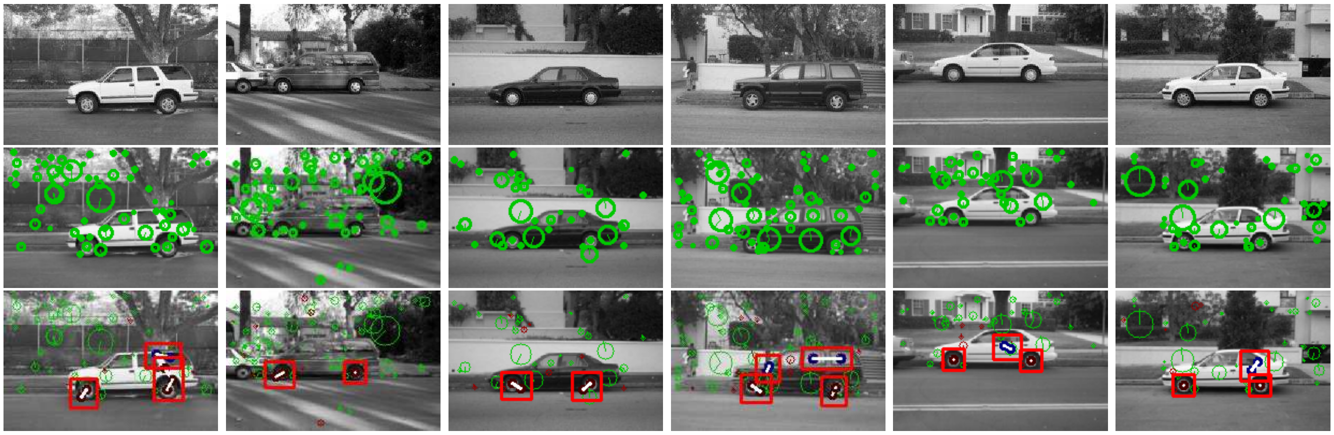

Figure 9: Examples of meaningful itemsets from car category (6 out of 123 images). The cars are all side views, but are of different types and colors and located in various clutter backgrounds. The first row shows the original images. The second row shows their visual primitives (PCA-SIFT points), where each green circle denotes a visual primitive with corresponding location, scale and orientation. The third row shows the meaningful itemsets. Each red rectangle in the image contains a meaningful itemset (it is possible two items are located at the same position). Different colors of the items denote different semantic meanings. For example, wheels are dark red and car bodies are dark blue. The precision and recall scores of these semantic patterns are shown in Fig. 8.

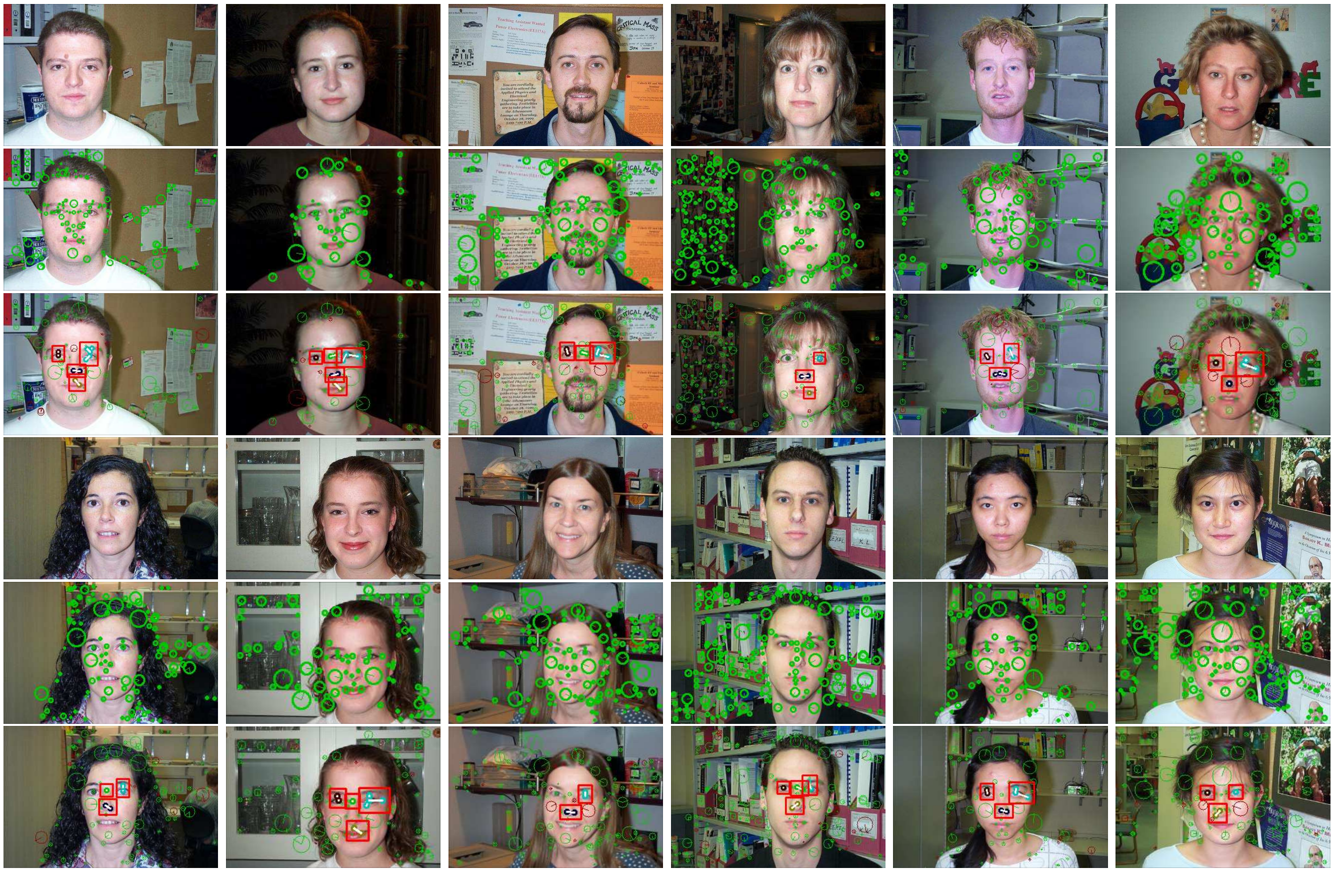

Figure 10: Examples of meaningful itemsets from face category (12 out of 435 images). The faces are all front views but are of different persons. Each red rectangle contains a meaningful itemset. Different colors of the visual primitives denote different semantic meanings, *e.g.* green visual primitives are between eyes *etc*. The precision and recall scores of these semantic patterns are shown in Fig. 7.