Visual Synset: Towards a Higher-level Visual Representation

Yan-Tao Zheng
National University of Singapore
yantaozheng@comp.nsu.edu.sg

Ming Zhao
Google Inc. U.S.A
zhaoming@zhaoming.name

Shi-Yong Neo
National University of Singapore
neoshiy@comp.nus.edu.sg

Tat-Seng Chua
National University of Singapore
chuats@comp.nus.edu.sg

Qi Tian
Institute for Infocomm Research, Singapore
tian@i2r.a-star.edu.sg

Abstract

We present a higher-level visual representation, visual synset, for object categorization. The visual synset improves the traditional bag of words representation with better discrimination and invariance power. First, the approach strengthens the inter-class discrimination power by constructing an intermediate visual descriptor, delta visual phrase, from frequently co-occurring visual word-set with similar spatial context. Second, the approach achieves better intra-class invariance power, by clustering delta visual phrases into visual synset, based their probabilistic 'semantics', i.e. class probability distribution. Hence, the resulting visual synset can partially bridge the visual differences of images of same class. The tests on Caltech-101 and Pascal-VOC 05 dataset demonstrated that the proposed image representation can achieve good accuracies.

1. Introduction

In the task of visual object recognition, the bag-of-words (BoW) methods have achieved many significant results [6, 21, 22, 26, 9, 8], due to its simplicity, effectiveness and good practical performance. Analogous to document representation in terms of words in text domain, the bag-of-words approach models an image as a geometric-free unordered bag of visual words, which are formed by vector quantization of local region descriptors, such as Scale Invariant Feature Transform (SIFT) [11]. By coding the statistics of local image regions independently, the bag-of-words approach achieves the robustness in handling variable object appearances caused by changes in pose, image capturing conditions, scale, translation, clutter and occlusion, etc.

Though various systems [6, 21, 22, 26, 9, 8] have shown promising practical performances of bag-of-words approach, the accuracies of visual object categorization are still incomparable to its analogy in text domain, i.e. the document categorization. The reason is obvious. The textual word possesses semantic itself and the documents are well-structured data regulated by grammar, linguistic and lexicon rules. In contrast, there appears to be no well-defined rule in visual word composition of images. The objects of same class might have arbitrarily different shapes and visual appearances, while objects of different classes might share similar local appearances. Consequently, such huge object appearance diversities lead to scarce correlation between proximity of images in feature space and their semantic relevance, which renders statistical models ineffective in visual object recognition. The lack of such correlation is a form of ambiguity and uncertainty of visual word representation [23, 24], which are manifested by two phenomenons: polysemy and synonymy. The polysemous visual word is a one that might represent different semantic meanings in different image context, while the synonymous words are a set of visually dissimilar words representing the same semantic meaning. By sharing a set of polysemous visual words, the semantically dissimilar images might be close in feature space, while the synonymous visual words may cause the images with same semantic to be far apart in feature space.

To achieve more effective object recognition, the polysemy and synonymy issues must be tackled.

- **Polysemy issue:** The polysemy encumbers the distinctiveness of visual words and leads to under-representations [23, 24]. Its consequence is effectively low inter-class discrimination. The polysemy is rooted from two reasons. First, visual word is the result of vector quantization (clustering of region descriptors) and each visual word corresponds to a group of local regions. Due to visual diversity, it is impossible to make regions of one visual word with homogeneous appearances. Such quantization error inevitably results in ambiguity of visual word representation. Second, the regions represented in a visual word might come from the object parts with different semantic but similar local appearances. For example in Figure 1, visual word

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

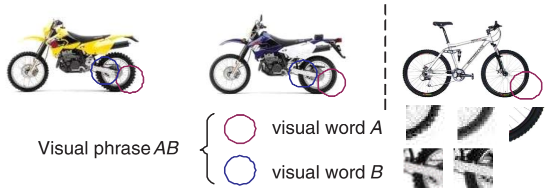

Figure 1. Examples of visual phrase

A is not able to distinguish motorbike from bicycle, as they share visually similar tires. However, the combination of visual word A and B, i.e. the visual phrase AB, can effectively distinguish motorbike from bicycle. The polysemy issue can, therefore, be resolved by mining inter-relation among visual words in certain neighborhood region. Yuan el at. [23, 24] and Quack el at. [16] proposed to utilize frequently co-occurring visual word-set to address the polysemy issue. Specifically, Yuan el at. denote such visual word-set as visual phrase. The major weakness of visual phrase approach is that it merely considers the co-occurrence information among visual words but neglect spatial information among them. To tackle such issue, we propose a new visual descriptor - delta visual phrase, which incorporates both co-occurrence and spatial scatter information of visual words.

• **Synonymy issue:** The synonymy is attributed to the visual diversity of object of same semantic class. Such appearance diversity makes multiple visual words share same or similar semantic meaning. It is, in fact, an over-representation of semantics by visual words [23, 24]. The consequence is large intra-class variations. In this circumstance, both visual words and phrases become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of same semantic. To tackle this issue, a higher level visual content unit is needed. In text domain, when documents of same topic or categories are represented by different sets of words, the word synset (synonymy set) that link words of similar semantic are robust to model them [3]. Inspired by this, we propose a novel visual content unit, *visual synset*, on top of visual words and phrases. We define *visual synset* as a relevance-consistent group of visual words or phrases with similar semantic. However, it is hard to measure the semantic of a visual word or phrase, as they are only a quantized vector of sampled regions of images. Rather than in a conceptual manner, we define the 'semantic' probabilistically as semantic inferences $P(c_i|w)$ of visual word or phrase w towards image class c_i.

Intuitively, if several visual words or phrases from different images share similar class probability distribution, like the brand logos in car images shown in Figure 2, then the visual synset that clusters them together shall possess similar class probability distribution and distinctiveness towards image classes. The visual synset can then partially bridge the visual differences between these images and deliver a more coherent, robust and compact representation of images.

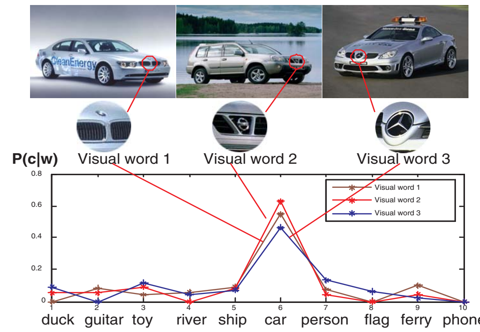

Figure 2. Examples of visual synset that clusters three visual words with similar image class probability distributions.

2. Overview and Preliminaries

As shown in Figure 3, the overall flow of the proposed approach consists of 3 phases. Phase 1 constructs visual words or visual codebook. Here we follow the notation of [23]. It first extracts regions from an image and computes visual features of regions a_i. It then performs clustering on a_i to generate visual code $\Omega = \{W_1, ..., W_M\}$, where W_i is a visual word. The image $\mathcal{I}$ is then represented by a bag of visual words $\{W_{(a_1)}, ..., W_{(a_i)}, ...\}$, where $W_{(a_i)}$ is the corresponding visual word of region a_i. Phase 1 can be regarded as a standard bag of words approach.

Phase 2 tackles the polysemy issue in visual words, by exploiting co-occurrence and spatial scatter information among visual words. For each local region $a_i \in \mathcal{I}$ from phase 1, its local spatial neighborhood $\mathcal{G}$ is defined as group of its K nearest neighbor regions $\{W_{(a_i)}, W_{(a_{i_1})}, W_{(a_{i_2})} ... W_{(a_{i_K})}\}$. By processing all image, a visual word group database $\mathbf{G} = \{\mathcal{G}_i\}_{i=1}^N$ will be generated. In the domain of data mining, the database $\mathbf{G}$ can be regarded as a transaction database [7]. Therefore, the discovery of frequently co-occurring visual word-sets, i.e. visual phrases, can be reduced to a task of frequent itemset mining (FIM) in the database $\mathbf{G}$ [7] [23]. We explore the FP-growth algorithm to perform the FIM task, as its prefix-tree structure enable it to store and search frequent itemsets in an extremely efficiently way. A visual word-set $\mathcal{P} \subset \Omega$ is counted as a frequently co-occurring set or a visual phrase, if its frequency $freq(\mathcal{P}) > \theta$. Specifically, the neighborhood $\mathcal{G}$ is called the support region of $\mathcal{P}$, as $\mathcal{P}$ is mined from the database of $\mathcal{G}$. By mining frequently co-occurring word-sets with different support regions, we propose a new collocation pattern - delta visual phrases, which incorporate both co-occurrence and spatial scatter information. Details of phase 2 will be introduced in Section 3. For simplicity, we denote both visual words and delta visual phrases as visual lexicons.

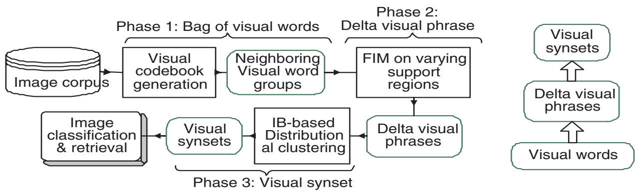

Figure 3. The overall framework of visual synsets generation

Phase 3 addresses the synonymy issue. Given image categories $\mathcal{C} = \{c_i\}_{i=1}^m$, the 'semantic' of a visual lexicon w is its contribution to the classification of its belonging image, which can be approximately measured by $P(c_i|w)$. Phase 3 clusters visual lexicons with consistent semantic into visual synsets, via Information Bottleneck-based distributional clustering. Details of phase 3 will be discussed in Section 4. Finally, the visual object recognition are performed based on image representations in terms of visual synsets.

3. Discovering Delta Visual Phrase

The major shortcoming of visual phrase proposed in [23, 24] is that it neglects the spatial inter-relation among visual words. To tackle this issue, the proposed delta visual phrase is mined not only from co-occurrence information, but also the local proximity of visual words. Such spatial proximity information defines the specificity of the visual phrase, which can be determined by the size of support region that visual phrase is mined from. Specifically, a **delta visual phrase** is defined in 2 dimensions: its member visual word-set $\mathcal{P}$ and its scatter $\mathcal{R}$, namely, how spread the visual phrase crosses over image.

Prior to presenting the proposed delta visual phrase, we first introduce the concept of **minimal support region**. The support region of visual phrase $\mathcal{P}$ is the visual word group $\mathcal{G}$ of size K, where K is the number of visual words in the neighborhood $\mathcal{G}$. Let $\mathcal{G}^1$, $\mathcal{G}^2$,..., $\mathcal{G}^{k-1}$, $\mathcal{G}^k$,... be a series of support regions with same centroid and growing size. The minimal support region is then defined as below.

Definition 3.1. *The region $\mathcal{G}^k$ is called **minimal support region** of visual phrase $\mathcal{P}$, if any smaller region $\mathcal{G}^{k-i}, \forall i > 0$ is not large enough to discover the visual phrase $\mathcal{P}$.*

With respect to each support region $\mathcal{G}^k$, the delta visual phrase is defined as below.

Definition 3.2. *The **delta visual phrase** (dVP) of region $\mathcal{G}^k$ is the visual phrase that has $\mathcal{G}^k$ as minimum support region. In other words, the delta visual phrase of region $\mathcal{G}^k$ is the newly discovered visual phrases when the support region just grows from $\mathcal{G}^{k-1}$ to $\mathcal{G}^k$. The size of $\mathcal{G}^k$ is therefore the **scatter** $\mathcal{R}$ of delta visual phrase and $\mathcal{R} = |\mathcal{G}^k|$.*

Intuitively, the delta visual phrase is mined from the changes of support regions. This is also why the word

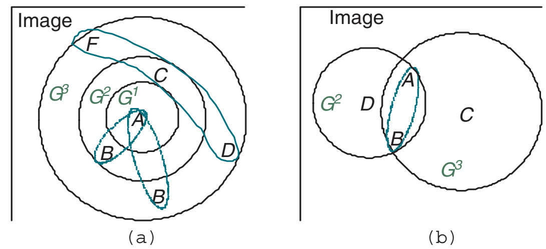

Figure 4. Examples of delta visual phrases. (a) Visual word-set *'CDF'* is a dVP with $\mathcal{R} = |\mathcal{G}^3|$. (b) Visual word-set *'AB'* can not be counted as a dVP with $\mathcal{R} = |\mathcal{G}^3|$

"delta" is in its name. The visual word-set $\mathcal{P}$ is deemed to be delta visual phrase [$\mathcal{P}$, $\mathcal{R}$], if it satisfies one of the following condition:

$$freq^{\mathcal{G}^k}(\mathcal{P}) - freq^{\mathcal{G}^{k-1}}(\mathcal{P}) > \theta_k, \tag{1}$$

where $\mathcal{R} = |\mathcal{G}^k|$, $freq^{\mathcal{G}^k}(\mathcal{P})$ is the frequency of a visual word-set $\mathcal{P}$ for support region $\mathcal{G}^k$ and θ_k is the threshold. For example in Fig. 4 (a), the visual word-set '*CDF*' will be considered as dVP with scatter $\mathcal{R} = |\mathcal{G}^3|$, if the number of newly discovered instances of '*CDF*' resulted from the increase of support region (from $\mathcal{G}^2$ to $\mathcal{G}^3$) is greater than the threshold. The Eq. (1) also ensures that the visual words of a dVP are scattered over its support region. For example in Fig. 4 (b), the instance of visual word-set '*AB*' will not be counted for dVP with $\mathcal{R} = |\mathcal{G}^3|$, as it lies in region $\mathcal{G}^2$ as well and will be offsetted by Eq. (1). If we define the size of first support region $\mathcal{G}^1$ to be 1, the resulted delta visual phrases are actually visual words with scatter $\mathcal{R} = 1$. In this way, we can combine visual words and delta visual phrases into a unified representation.

▽ Statistical Significance Measure

Yuan el at. [23] proposed to measure the statistical significance of visual phrase based on its frequency and its component visual word frequencies. This measurement, however, neglects the coherency of component visual words in visual phrase. We measure the significance on the basis that the delta visual phrase should be a visual word-set that is frequently and coherently occurring together, with respect to certain semantic meaning. Specifically, the significance score $L([\mathcal{P}, \mathcal{R}])$ of a dVP [$\mathcal{P}$, $\mathcal{R}$] is defined as:

$$L([\mathcal{P}, \mathcal{R}])) = freq([\mathcal{P}, \mathcal{R}]) \cdot \frac{P(\mathcal{P}, \mathcal{R}|\mathbf{D}_{\mathcal{I}})}{1 + P(\mathcal{P}^{-}|\mathbf{D}_{\mathcal{I}})} \tag{2}$$

where $P(\mathcal{P}, \mathcal{R}|\mathbf{D}_{\mathcal{I}})$ is the probability that the visual word-set $\mathcal{P}$ forms a valid dVP with scatter $\mathcal{R}$ by satisfying the condition of Eq. (1) and it can be approximated by $\frac{docfreq([\mathcal{P}, \mathcal{R}]}{T}$, where $docfreq([\mathcal{P}, \mathcal{R}])$ is the document frequency equal to number of images containing dVP $[\mathcal{P}, \mathcal{R}]$. $\mathcal{P}^{-}$ is the visual word-set $\mathcal{P}$ that does not form any

Figure 5. An example of visual synset generated from Caltech-101 dataset, which groups two visual lexicons representing two salient parts of motorbikes.

valid dVP; and $P(\mathcal{P}^-|\mathbf{D}_\mathcal{I})$ is the probability that visual word-set $\mathcal{P}$ forms some random and sporadic patterns, which can be approximated by $\frac{docfreq(\mathcal{P}^-)}{T}$. $freq([\mathcal{P}, \mathcal{R}])$ is the frequency of dVP $[\mathcal{P}, \mathcal{R}]$. Intuitively, we want to penalize the delta visual phrases whose member visual words also frequently co-occur in a random and sporadic manner. In this way, we enforce the correlation among member visual words, and therefore, ensures the coherency of delta visual phrases.

▽ Unique Counting of Maximal Visual Word-set

The subsets of a frequent visual word-set $\mathcal{P}$ are frequent as well, and therefore, will be falsely counted as dVP. To address this problem, we exploit closed FIM algorithms to discover maximal frequent itemsets, in the way that any of its subsets will not be considered as frequent itemset, in the spirit of [23]. In the phase of FIM, a word-set might be over-counted, if it lies in the overlapping area of different neighborhood regions. To overcome this problem, we borrow the approach in [23] to re-count real instances of word-set through the original image database.

4. Generating Visual Synset

Though the co-occurrence and spatial scatter information make visual lexicons more distinctive, the synonymy issue remains. To tackle this issue, we propose to exploit the prior available semantic knowledge, i.e. semantic class labels of training images and their distributions, to generate a higher level visual content unit, called **visual synset**, using a supervised learning process.

4.1. Visual Synset: a Semantic-Consistent Cluster of Visual Lexicons

In text literature, the synonymous words are usually clustered into one synset (**syn**onymy **set**) to improve document categorization performance, based on word-document class distribution [3]. Such approach inspires us in solving the synonymy issue in visual lexicons. However, it is infeasible to define the semantic meaning of visual lexicon, as it is only a set of quantized vectors of sampled regions of images. Hence, rather than defining the semantic of a visual lexicon in a conceptual manner, we define it probabilistically, in the spirit of [3].

Definition 4.1. *Given image categories $\mathcal{C} = \{c_i\}_{i=1}^m$, the **semantic** of a visual lexicon $\mathcal{V}$ (visual word or phrase) is its contribution to the classification of its belonging image, which can be approximately measured by $P(c_i|\mathcal{V})$.*

As shown in Fig. 2, the probability distribution $P(c_i|\mathcal{V})$ implies the semantic inference of visual lexicon $\mathcal{V}$, namely how much $\mathcal{V}$ votes for each of the classes. We then define the *visual synsets* as below.

Definition 4.2. *The **visual synset** is a probabilistic concept or a semantic-consistent cluster of visual lexicons, in which the member visual lexicons might have different visual appearances but similar semantic inferences towards the image classes*

The rational of visual synset is that due to the visual heterogeneity and distinctiveness of objects, a considerable number of visual lexicons are intrinsic and highly indicative to certain classes. This implies that some visual lexicons tend to share similar probability distribution $P(c_i|\mathcal{V})$, which might peak around its belonging classes. By grouping these highly distinctive and informative visual lexicons into visual synsets, the visual differences of images from the same class can be partially bridged. Consequently, the image distribution in feature space will become more coherent, regular and stable. For example in Fig. 5, two visually different salient components (visual lexicons) of motorbikes can be grouped into one visual synset, based on their image class probability distribution. Consequently, the visually different motorbike images will now have some commonality in the feature space.

4.2. Information Bottleneck Principle

By formulating visual synset construction as a task of visual lexicon clustering based on their class probability distributions, the issue now is reduced to how to measure the 'right' distance between these distributions, namely the similarity metric in clustering. Pereira et al. [15] proposed to use the relative entropy or Kullback-Leibler (KL) distance to measure the distributional similarity. The KL distance is, however, not symmetric. To address this issue, Baker and McCallum [2] proposed to utilize the average of KL divergence of each distribution as the clustering similarity metric. Such metric, however, focuses merely on the distributional similarity but neglect the fact that clustering is also a process of data compression (compressings a group of data into one clustering). to address the issue above, we propose to utilize the Information Bottleneck (IB) principle to guide the clustering process. Given the joint distribution $P(\mathbf{V}, \mathcal{C})$ of the visual lexicons $\mathbf{V}$ and image classes $\mathcal{C}$, the goal of IB principle is to construct the optimal compact representation of $\mathbf{V}$, namely the visual synset clusters $\mathbf{S}$, such that $\mathbf{S}$ preserves as much information as possible about $\mathcal{C}$. In particular, the IB principle is reduced to the following Lagrangian optimization problem to maximize

$$\mathcal{L}[P(\mathcal{S}|c)] = I(\mathbf{S};\mathcal{C}) - \beta I(\mathbf{V};\mathbf{S}) \tag{3}$$

with respect to $P(\mathcal{S}|c)$ and subject to the Markov condition $\mathbf{S} \leftarrow \mathbf{V} \leftarrow \mathcal{C}$. The term $I(\mathbf{S};\mathcal{C})$ measures the information that $\mathbf{S}$ contains about $\mathcal{C}$ and $\beta I(\mathbf{V};\mathbf{S})$ measures the information loss in clustering $\mathbf{V}$ into $\mathbf{S}$. Intuitively, Eq. 3 aims to cluster or compress the visual lexicons into visual synsets through a compact bottleneck, under the constraint that this compression keeps the information about image classes as much as possible and the information loss in the clustering as small as possible.

The IB optimization in Eq. 3 yields the solution of: (1) the prior probability $P(\mathcal{S})$ for each visual synset cluster $\mathcal{S} \in \mathbf{S}$; (2) the membership probability $P(\mathcal{S}|\mathcal{V})$ of visual lexicon $\mathcal{V}$ to its visual synset cluster $\mathcal{S}$; and (3) the visual synset distribution $P(c|\mathcal{S})$ over image classes, which are specifically defined in the equations below:

$$\begin{cases} P(\mathcal{S}) = \sum_{\mathcal{V}} P(\mathcal{S}|\mathcal{V})P(\mathcal{V}) \\ P(c|\mathcal{S}) = \frac{1}{P(\mathcal{S})}\sum_{\mathcal{V}} P(\mathcal{S}|\mathcal{V})P(\mathcal{V})P(c|\mathcal{V}) \\ P(\mathcal{S}|\mathcal{V}) = \frac{P(\mathcal{S})}{Z(\beta,\mathcal{V})} exp(-\beta D_{KL}[P(c|\mathcal{V})||P(c|\mathcal{S})]) \end{cases} \tag{4}$$

where $Z(\beta,\mathcal{V})$ is the normalization factor, β is a lagrange parameter that determines the cluster resolution and $D_{KL}[P(c|\mathcal{V})||P(c|\mathcal{S})]$ is the Kulback-Libeler divergence [19] between $P(c|\mathcal{V})$ and $P(c|\mathcal{S})$.

There exist several implementations of IB principle. Here, we adopt the sequential Information Bottleneck (sIB) clustering algorithm [18] to generate the optimal visual synset clusters in our approach, as it is reported to outperform other IB clustering techniques [18]. The target principled function that sIB algorithm exploits to guide the clustering process is $\mathcal{F}(\mathbf{S}) = \mathcal{L}[P(\mathcal{S}|c)]$ as in Eq. 3. The sIB algorithm takes visual synset cluster cardinality $|\mathbf{S}|$, and joint probability $P(\mathcal{V}, c)$ as input, and starts with some initial random clustering $\mathbf{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K\}$ on $\mathbf{V}$. It then simulates the process of K-means clustering to iteratively reach a local maximum of $\mathcal{F}(\mathbf{S})$. Specifically, the cost $d_{\mathcal{F}}(\mathcal{V}, \mathcal{S}^{new})$ of moving visual word $\mathcal{V}$ to a new cluster $\mathcal{S}^{new}$ can be defined as (cf. [18] for more details):

$$d_{\mathcal{F}}(\mathcal{V}, \mathcal{S}^{new}) = (P(\mathcal{V}) + P(\mathcal{S}^{new})) \cdot JS(P(c|\mathcal{V}), p(c|\mathcal{S}^{new})) \tag{5}$$

where $JS(x, y) =$ is the *Jensen-Shannon* divergence [19].

5. Experiments and Discussion

5.1. Testing Dataset and Experimental Setup

We evaluate the proposed image representation on object categorization task using two datasets: 1) Caltech-101 dataset [10]; and 2) Pascal-VOC 2005 [5]. The classifier used is Support Vector Machines (SVM) [20] with generalized RBF kernel.

Generation of Visual Lexicon Codebook:

The local region extraction is accomplished by Difference of Gaussian (DoG) [11] and Hessian Laplace [12] algorithms. DoG corresponds to high contrast structures, and Hessian Laplace samples blob-like regions, which tend to complement each other. For each region, the SIFT and Spin [26] features are computed as region descriptor. We then perform k-means clustering to obtain 1010 primitive visual words in total. To discover delta visual phrase, we perform FIM on the database $\mathbf{G}$ of approximately 3 million visual word groups with support region size of 1, 4, 8 and 12 respectively. Based on the significance score in Equation 1, we construct the visual lexicon codebook by selecting the top K delta visual phrases (dVP) with highest scores. In the experiments, K is set to 1100, 1200, 1300, 1400, 1500, 1700, 1800 and 2000 respectively.

5.2. The Caltech-101 Dataset

The Caltech-101 dataset [10] contains 102 image categories and a total of 9233 images. For benchmark purpose, we follow the setup of [25] and [9] by selecting 30 images from each category as training set. The evaluation criteria is the mean classification accuracy, which is the average of evenly weighed recognition rate of each category.

As most visual objects in Caltech-101 are dominant objects positioned at the centre of the images, we divide an image into 2×2 grides and extract visual lexicons and synsets on each grid. The result image representation H is the concatenated vector of each grid and whole image. As the object images do not have large scale changes, we also incorporate global visual information into the distance function of RBF kernel, so as to complement the part-based local features. The distance function of RBF kernel is, therefore, defined as:

$$D(\mathcal{I}_L, \mathcal{I}_R) = ||H_L - H_R||_{L2} + \lambda||T_L - T_R||_{L2}, \tag{6}$$

where H_L is the normalized visual lexicon/synset feature vector of image L (H_R for image R respectively) and T_L is the normalized wavelet texture (WT) histogram (T_L for image R respectively). For WT, an image is divided over 3×3 grids and the variance in 9 Haar wavelet sub-bands for each grid are computed to form a 81D feature vector. We empirically set $\lambda = 1/6$, based on the classification of 12 classes.

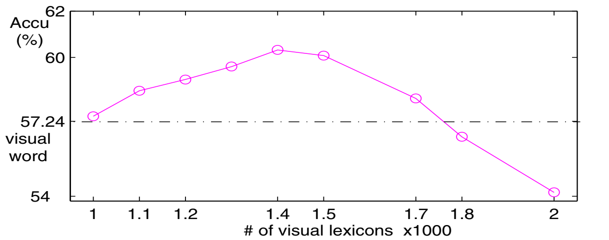

Figure 6. The average classification accuracy by visual lexicons on Caltech-101 dataset.

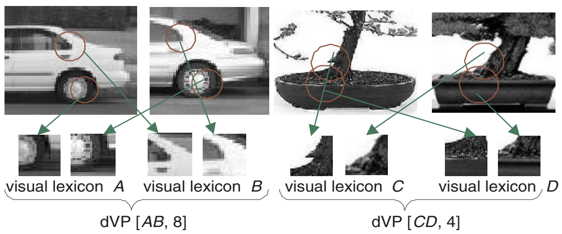

Figure 7. The examples of delta visual phrases generated from Caltech-101 dataset. The first dVP consists of disjoint visual words A and B with a scatter of 8 and the second has joint visual words C and D with a scatter of 4

Performance of Visual Lexicons: We first perform classification, based on 1010 visual words. This yields a mean classification accuracy of 57.24%. This classification is used as the baseline of our experiments. Next, we perform object categorization, based on 1100, 1200, 1300, 1400, 1500, 1700, 1800 and 2000 visual lexicons respectively. As shown in Fig. 6, the performance increases as more visual lexicons are incorporated up to 1400. In particular, the codebook with 1400 visual lexicons gives the highest accuracy of 60.32%. This demonstrates that by incorporating co-occurrence and spatial scatter information, the visual lexicons do carry more distinctiveness than visual words. Fig. 7 shows some examples of delta visual phrases with different scatter. As shown, when objects share some appearance similarity in a large scope, the delta visual phrase can combine the ambiguous visual words scattered in such area into one more distinctive unit, which can contribute to distinguishing objects of different classes with larger inter-class distance and better classification.

However, we also observe that when the number of lexicons is above 1700, the performance drops drastically and even becomes inferior to original visual word representation. We attribute such performance degradation to the fact that the newly incorporated visual lexicons with lesser significance score might not be statistically substantial. Though these visual lexicons might still be distinctive patterns, their statistical sparseness renders image distributions in feature space more incoherent, sporadic or even noisy.

Performance of Visual Synset: We evaluate the effectiveness of visual synset, by performing IB-based distributional clustering on the codebook of 1400 visual lexicons (best run from previous section). Specifically, we set the cardinality of visual synsets $|\mathbf{S}|$ to 50, 100, 200, 400, 600, 800, 1000 and 1200. Fig. 8 displays the average classification accuracies. From Fig. 8, we observe that with proper cardinality, the visual synset representation can deliver superior results over both visual lexicons and visual words with a more compact representation. For example, the run with only 50 visual synsets can achieve an accuracy of 55.21%, while the runs with 600 visual synsets has achieved superior accuracies over the run with 1400 visual lexicons. This representation compactness does not only enable high computational efficiency but also alleviate the issue of curse of dimensionality.

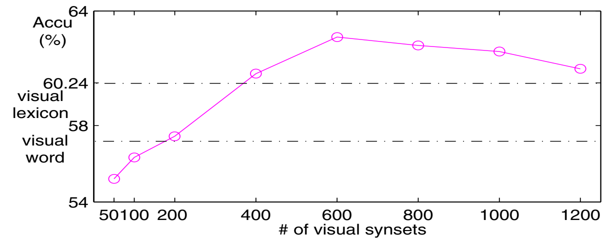

Figure 8. The average classification accuracy by visual synsets on Caltech-101 dataset.

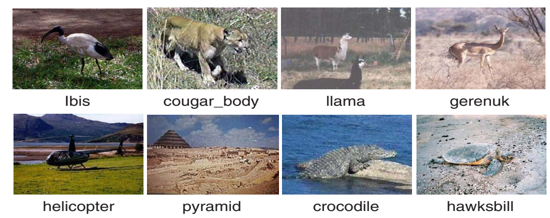

Figure 9. Example images of classes that have visual lexicon outperforming visual synsets.

The best run is the one with 600 visual synsets and it achieves an accuracy of 62.64%. We attribute such improvements to two factors: (1) by fusing semantic-consistent visual lexicons together, the visual synset reduces the intra-class variations and renders the image distribution in feature space more coherent and manageable; and (2) the visual synset is a result of supervised dimensionality reduction and the properly reduced dimensionality can partially resolve the statistical sparseness problem of visual lexicons and also enable better classification. However, after a detailed comparison, we find that 16 classes have visual lexicons delivering better classification performance than visual synsets. Fig. 9 shows some example images from these classes. With close examination, we find that the images of these classes are not visually distinctive from images of other classes, either due to their cluttered backgrounds or neutral textures and color of objects. This leads to the lack of visual lexicons distinctive to these classes. Consequently, these non-distinctive visual words might be

Table 1. Benchmark of classification performance on Caltech-101 dataset. (VW: visual word; VL: visual lexicon; VS: visual synset)

run	VL	VS	[26]	[13]	[9]	[25]	[4]
Accu(%)	60.2	62.6	53.9	56	64.6	66.2	81.3

Table 2. Pascal-VOC 2005 EER results based on visual words (VW), best run of visual lexicons (VL), best run of visual synsets (VS) and publicly reported systems.

run	VW	VL	VS	[26]	[14]
Mean EER	0.92	0.936	0.948	0.928	0.954

clustered together with visual lexicons indicative of other classes and resulted in non-distinctive visual synsets that effectively link images of different classes together.

We also observe that the number of visual synsets plays an important role in its performance. A too small number of visual synsets usually gives bad performance. This is because a small number of visual synsets will force the distinctiveness-inconsistent visual words together and generate noninformative and nondistinctive visual synsets. Overall, the experimental results show that the number of visual synsets between $1/3$ and $2/3$ of visual lexicon codebook size usually gives a reasonably good performance. Fig. 5 shows an example of visual synset generated from Caltech-101 dataset.

Benchmark on Caltech-101: In the run of 600 visual synsets generated from 1400 visual lexicons, we achieve an accuracy of 62.64%. Table 1 summarizes the accuracies of other reported systems. As shown in Table 1, the proposed visual synset approach outperforms most of existing systems and delivers a comparable result with the state-of-arts one, with more compact image representation.

5.3. The Pascal-2005 Dataset

The Pascal-VOC 2005 contains four object classes: bicycles, cars, people and motorbike. It has one training dataset of 684 images and two testing sets with 689 images (test set 1) and 956 images (test set 2). Here, we use test set 1 for our evaluation, as many recent works utilize this set. We follow the same experimental setup for Caltech-101 and the evaluation criteria here is equal error rate (EER). The EER is a point on the Receiver Operating Characteristic (ROC) curve, which measures the accuracy when the number of false positives and negatives are equal.

Table 2 summarizes the results of classification based on visual words only, visual lexicons and visual synsets. The baseline classification with visual words give an EER of 0.92. Similar to Caltech-101, the EER increases as more visual lexicons are incorporated and reaches its peak of 0.936, when the number of visual lexicons is 1300. The optimum number of visual lexicons here is lesser than 1400 in Caltech-101. We attribute this to the fact that the images of same category in Pascal-VOC 2005 are more visually diverse. Therefore, the resulting delta visual phrases are less statistically stable. Based on the best run of visual lexicons, we generate visual synsets and perform the classifications. Consistent to the observation in Caltech-101, the visual synset achieves both compactness and superior performance. Specifically, the run with 600 visual synsets delivers the best EER of 0.948. Table illustrates the ERR benchmark with other published approaches. The visual synset delivers a comparable result with the state-of-arts system [14], which however extracts quite a large amount (10k) of regions per image for classification.

6. Related Work

To improve the bag-of-words approach, many researchers have proposed various systems. Lazebnik el at. [9] proposed a spatial pyramid model to incorporate spatial information hierarchically. Agarwal and Triggs [1] proposed a hyperfeature to code the local visual information in a multi-resolution way. To address the polysemy issue of visual words, Juan el at. [23] and Quack el at. [16] proposed visual phrase, i.e. frequently co-occurring visual and spatial configurations. Different from the approaches above, our proposed delta visual phrase attempts to exploit both co-occurrence and spatial scatter information of visual words, by utilizing a series of varying support regions, so as to deliver more distinctive primitive visual features.

The performance of primitive visual features, however, depends highly on the visual similarity and regularity. To mitigate such problem, Sivic el at. [17] proposed to model images with some higher level latent topic features by exploiting probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). Agarwal and Triggs also demonstrated the effectiveness of LDA in image classification in [1]. pLSA and LDA are similar to the proposed visual synset in the way that they are all some kinds of intermediate features derived from primitive visual lexicons. However, the proposed visual synset is different from pLSA and LDA in the way that visual synset is not a result of a generative model.

Unlike pLSA and LDA, the proposed visual synset is not a latent or hidden semantic variable in the middle of visual lexicons and image semantics. pLSA assumes a set of latent topic variable to tie up documents/images and words, while LDA treats a latent topic as a multinomial distribution over words and the mixture of latent topics per document/image [17]. The Markov condition in pLSA and LDA is be $\mathbf{V} \leftarrow \mathbf{S} \leftarrow \mathcal{C}$ [19], where $\mathbf{S}$ denotes the latent topic variable. On the contrary, the visual synset is the results of supervised data-mining process of compressing visual lexicons via distributional clustering based on IB principle. Thus, it is only conditional on visual lexicons, which follow the joint distribution of visual lexicons and image classes. Consequently, the Markov chain condition here is

$\mathbf{S} \leftarrow \mathbf{V} \leftarrow \mathcal{C}$, where $\mathbf{S}$ denotes visual synset variable.

7. Conclusion and Future Work

In order to address the polysemy and synonymy issue of visual words, we proposed a novel image feature, *visual synsets*, for visual object categorization. To address the polysemy issue, we exploit the co-occurrence and spatial scatter information of visual words to generate a more distinctive visual configuration, i.e. delta visual phrase. The improved distinctiveness leads to better inter-class distance. To tackle the synonymy issue, we proposed to group delta visual phrase with similar 'semantic' into a visual synset. Rather than in conceptual manner, the 'semantic' of a delta visual phrase is probabilistically defined as its image class probability distraction. The visual synset is therefore a probabilistic relevance-consistent cluster of delta visual phrases, which is learned by Information Bottleneck based distributional clustering. The effect of visual synset is to reduce the intra-class variations. The tests on Caltech-101 and Pascal-VOC 05 datasets demonstrated that the proposed image representation can achieve good accuracies for object categorization.

Several open issues remain. First, the generation of delta visual phrase is a time-consuming task. A more efficient algorithm is demanded. Second, how the number of classes changes the semantic inference distribution of visual lexicons and how this affects the visual synset generation and final classification have not been investigated.

References

[1] A. Agarwal and W. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2006. 7

[2] L. Baker and A. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of ACM SIGIR*, pages 96–103, Melbourne, AU, 1998. 4

[3] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research g*, 3:1183–1208, 2003. 2, 4

[4] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007. 7

[5] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The 2005 pascal visual object classes challenge. In *Machine Learning Challenges. PASCAL Workshop 05*, number 3944, pages 117–176, Southampton, UK, 2006. 5

[6] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of International Conference on Computer Vision*, pages 1458–1465, USA, 2005. IEEE Computer Society. 1

[7] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 14(1), 2007. 2

[8] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of International Conference on Computer Vision*, Washington, DC, USA, 2005. 1

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. 1, 5, 7

[10] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach aested on 101 object categories. In *Proceedings of CVPR Workshop*, Washington, DC, USA, 2004. 5

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003. 1, 5

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005. 5

[13] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *CVPR '06*, pages 11–18, Washington, DC, USA, 2006. 7

[14] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*. Springer, 2006. 7

[15] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of ACL*, pages 183–190, Morristown, NJ, USA, 1993. 4

[16] T. Quack, V. Ferrari, B. Leibe, and L. Van-Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007. 2, 7

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005. 7

[18] N. Slonim, N. Friedman, and N. Tishby. Agglomerative multivariate information bottleneck. In *Advances in Neural Information Processing Systems (NIPS)*, 2001. 5

[19] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. 5, 7

[20] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, USA, 1995. 5

[21] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proceedings of International Conference on Computer Vision*, page 257, Nice, France, 2003. IEEE Computer Society. 1

[22] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proceedings of ICPR Workshop on Learning for Adaptiable Visual Sysmtems*, 2004. 1

[23] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Proceedings of the international conference on Knowledge discovery and data mining*, 2007. 1, 2, 3, 4, 7

[24] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of conference on Knowledge discovery and data mining*, pages 864–873, New York, NY, USA, 2007. ACM Press. 1, 2, 3

[25] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: discriminative nearest neighbor classification for visual category recognition. In *Proceedings of CVPR*, volume 2, pages 2126–2136, 2006. 5, 7

[26] J. Zhang, M. Marsza, S. Lazebnik, and C. Schmid. Local features and kernels for cassification of texture and object categories: a comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, 2007. 1, 5, 7