Vis Comput (2009) 25: 13–23
DOI 10.1007/s00371-008-0294-0

ORIGINAL ARTICLE

Toward a higher-level visual representation for object-based image retrieval

Yan-Tao Zheng · Shi-Yong Neo · Tat-Seng Chua · Qi Tian

Published online: 1 November 2008

Abstract We propose a higher-level visual representation, visual synset, for object-based image retrieval beyond visual appearances. The proposed visual representation improves the traditional part-based bag-of-words image representation, in two aspects. First, the approach strengthens the discrimination power of visual words by constructing an intermediate descriptor, visual phrase, from frequently co-occurring visual word-set. Second, to bridge the visual appearance difference or to achieve better intra-class invariance power, the approach clusters visual words and phrases into visual synset, based on their class probability distribution. The rationale is that the distribution of visual word or phrase tends to peak around its belonging object classes. The testing on Caltech-256 data set shows that the visual synset can partially bridge visual differences of images of the same class and deliver satisfactory retrieval of relevant images with different visual appearances.

Keywords Visual representation · Object-based image retrieval · More

Y.-T. Zheng (✉) · S.-Y. Neo · T.-S. Chua
National University of Singapore, Singapore, Singapore
e-mail: yantaozheng@comp.nus.edu.sg

S.-Y. Neo
e-mail: neoshiy@comp.nus.edu.sg

T.-S. Chua
e-mail: chuats@comp.nus.edu.sg

Q. Tian
Institute for Infocomm Research, Singapore, Singapore
e-mail: tian@i2r.a-star.edu.sg

1 Introduction

Due to the explosive proliferation of digital images, the image retrieval based on their visual content has spurted much research attention, in order to effectively index, monitor and manage image databases. Here, we narrow down our focus to object-based image retrieval (OBIR), which aims to retrieve image l containing salient object of the same semantic class c as the given example query image q from an image collection $\mathbf{D}_{\mathcal{I}}$ of semantic classes $C = \{c_i\}_{i=1}^{m}$.

Recently, many image retrieval systems [4, 12, 24, 33] have shown that the part-based representation for image retrieval is much superior over traditional global features, as one single image feature computed over the entire image is not sufficient to represent important local characteristics of objects [9, 13, 15, 27, 29, 32]. Specifically, the bag-of-words image representation has drawn much attention, as it tends to code the local visual characteristics toward object level, which is closer to the perception of human visual systems [12]. 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 is formed by vector quantization of local region descriptors, such as Scale Invariant Feature Transform (SIFT) [18]. 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 [4, 12, 24, 33] have shown the superiority of part-based image representation in image retrieval task, the accuracies of image retrieval are still incomparable to its analogy in text domain, i.e. the document retrieval. The reason is obvious. Compared to textual word, the visual word does not possess any semantics, as it is only

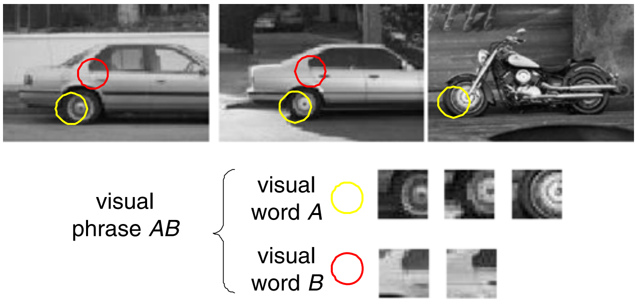


Fig. 1 Examples of visual phrase

a quantized vector of sampled local regions. However, if neglecting the semantic factor, what really distinguishes textual word from visual word is the discrimination and invariance power. Obviously, the textual words are more stable, indicative and representative of their belonging document topic, and therefore, possess much better discrimination and invariance power than visual words. On the other hand, the low discrimination and invariance power of visual words lead to low correlation between the topological proximity of images in feature space and their semantic relevance. Such low correlations between image features and its semantics render most statistical machine learning models ineffective in image retrieval. Hence, in order to achieve better image retrieval performance, the low discrimination and invariance issues of visual words must be tackled.

Discrimination issue A visual word might represent different semantic meanings in different image context. This encumbers the distinctiveness of visual words and leads to low discrimination. In fact, the discrimination issue is a problem of under-representations [31]. Its consequence is effectively small interclass distances. One of the major reasons for low discrimination issue is that the regions represented in a visual word might come from the object with different semantics but similar local appearances. For example in Fig. 1, the 'cars' and 'motorbike' share visually similar tires. The visual word *A* is, therefore, not able to distinguish motorbike from car. However, the combination of visual words *A* and *B*, i.e. the visual phrase *AB*, can effectively distinguish motorbike from car. The discrimination of representation can, therefore, be improved by mining inter-relation among visual words in certain neighborhood region. Specifically, we exploit the visual phrase, i.e. frequently co-occurring visual word-set, proposed in [31] to improve the discrimination power of visual word representation.

Invariance issue The images of the same semantic class can have arbitrarily different visual appearances and shapes. Such visual diversity of object causes one image semantics to be represented by different visual words. This leads to low invariance of visual words. The consequence is large intra-class variations. In this circumstance, the visual words become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of the same semantics. To tackle this issue, a higher-level visual content unit is needed. In text domain, when documents of the same topic or categories are represented by different sets of words, the word synset (**syn**onymy **set**) that links words of similar semantics is 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 semantics, in the spirit of [34]. However, it is hard to measure the semantics 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 'semantics' probabilistically as semantic inferences $P(c_i|w)$ of visual word or phrase w towards image class c_i.

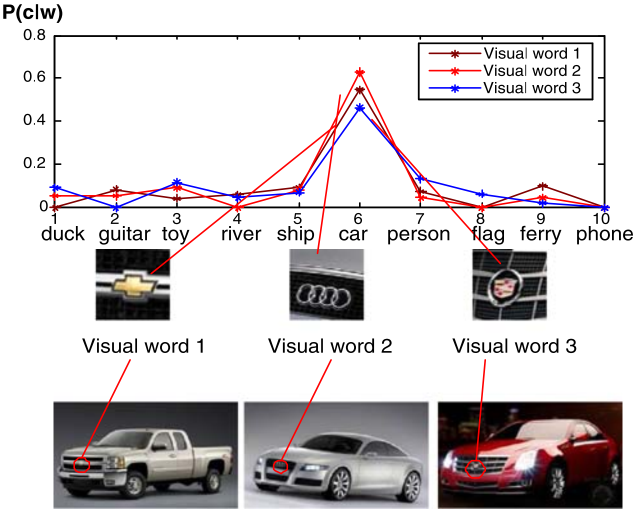


Fig. 2 Examples of visual synset that clusters three visual words with similar image class probability distributions

Intuitively, if several visual words or phrases from different images share similar class probability distribution, like the brand logos in car images shown in Fig. 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.

The major contribution of our work is: by improving discrimination and invariance power of visual word representation, we propose a higher level visual feature, visual synset, to retrieve images beyond their visual appearances. The testing on Caltech-256 data set [8] shows that the visual synset can partially bridge the visual difference of images of the same class and retrieve images beyond their visual appearances.

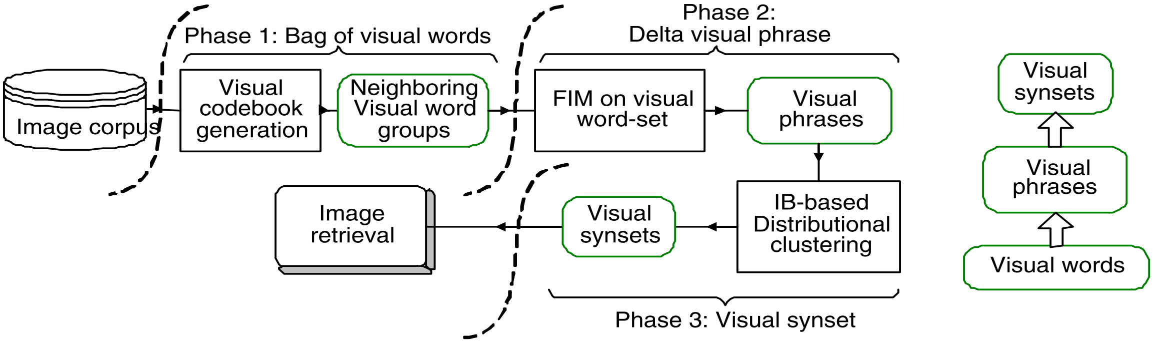

Phase 1: Bag of visual words
Image corpus
Visual codebook generation
Neighboring Visual word groups
Phase 2: Delta visual phrase
FIM on visual word-set
Visual phrases
IB-based Distributional clustering
Visual synsets
Image retrieval
Phase 3: Visual synset
Visual synsets
Visual phrases
Visual words

Fig. 3 The overall framework of visual synsets generation

2 Overview

As shown in Fig. 3, the overall flow of the proposed approach consists of 3 phases. Phase 1 can be considered as a standard bag-of-words (BoW) approach. In phase 1, the region sampling based on extremal scale saliency [6] is applied on gray-scale images to extract local regions. For each region, a 128D SIFT descriptor is computed. The vector quantization on SIFT descriptors is then performed to construct visual vocabulary by exploiting k-means clustering. The output of phase 1 is the image representations in terms of visual word vectors. Details of phase 1 will be introduced in Sect. 2.1.

Phase 2 tackles the low discrimination issue in visual words, by exploiting the spatial co-occurrence information among visual words. In the spirit of [31], phase 2 first generates a transaction database of visual word-sets. Each visual word-set is a group of spatially neighboring visual words. Then, the discovery of visual phrases, i.e. frequently co-occurring visual word-sets, can be reduced to the task of frequent itemset mining (FIM) in the transaction database [11, 31]. Details of phase 2 will be introduced in Sect. 3.

Phase 3 first takes a small set of labeled images from each class as training data set to select a set of highly informative and distinctive visual words and phrases. It then takes the joint probabilities of visual words or phrases and image classes (estimated from the labeled training set) as input to perform sequential Information Bottleneck (sIB) clustering algorithm [13] to group visual words/phrases with similar class probability distributions into visual synsets. Details of phase 3 will be introduced in Sect. 4.

2.1 Region extraction based on extremal scale saliency

We exploit the extremal scale-saliency [6] based region sampling strategy for visual word construction in phase 1. In fact, many region sampling algorithms are applicable here, e.g. keypoint detection like Laplacian of Gaussian, image segmentation, blobs of homogeneous regions [5], etc. The reason for utilizing extremal scale saliency [6] for region sampling is that we want to take the sampling of most repeatable image regions as the basis for visual word generation. This is to mitigate the statistical sparseness issue in bag-of-words (BoW) image representation, as the BoW image feature is usually a sparse vector with high dimensionality. We further argue that the repeatability of local regions of different images relies on the spatially integral similarity of the regions. Therefore, rather than using keypoint detection or image segmentation, we sample local regions by choosing the ones with extremal global saliency over the entire region.

In the scale-saliency algorithm [6], a region (circle) is determined by a point x in the image and its scale (radius). The scale-saliency algorithm defines the saliency of region in terms of the region's local signal complexity or unpredictability. More specifically, it exploits the Shannon Entropy of local attributes to estimate the region saliency H_{d,R_x} as below:

$$H_{d,R_x} = -\sum_i P_{D,R_x}(d_i)\log_2 P_{D,R_x}(d_i) \tag{1}$$

where R_x is the local region of point x, D is the local attribute vector with values $\{d_1,\dots,d_r\}$, $P_{D,R_x}(d_i)$ is the probability of D taking the value d_i, and H_{d,R_x} is the local entropy or saliency of region R_x. We select D as the color intensity histogram of local regions. Therefore, the regions with flatter intensity histogram distributions, namely more diverse color patterns, tend to have higher signal complexity and thus higher entropy and saliency.

By (1), each point x in the image will have one saliency value for each scale region. We then select the extremal saliency on two dimensions. First, for each point, the maximal and minimal saliency and their respective scales are selected, which leads to two scale-saliency maps, SMap_{min} and SMap_{max}. Second, rather than clustering points with similar spatial locations and saliency together as in [14], we spatially detect the local extremity of saliency maps (minimum for SMap_{min} and maximum for SMap_{max}) by using Difference of Gaussian (DoG) function

$$D(x,\sigma) = \big(G(x,k\sigma) - G(x,\sigma)\big) \cdot \text{SMap}(x) \tag{2}$$

where $D(x,\sigma)$ is the DoG saliency of point x, $\text{SMap}(x)$ is the saliency value for point x, k is the multiplicative factor, σ is the blur factor, and

$$G(x,\sigma) = \frac{1}{2\pi\sigma^2} e^{-x^2/\sigma^2}. \tag{3}$$

This local saliency extremity method based on DoG is inspired by the salient keypoint detection by DoG in [18]. The result of DoG function is effectively a new saliency map $D(x,\sigma)$ whose values are the differences between two blurred saliency maps with different sharpness (σ). If $D(x,\sigma)$ is larger or smaller than all its 8 spatial neighbors, then point x is deemed to be the local extremal (maximum or minimum) in its surrounding neighborhood and the local region specified by x and its scale will be selected for subsequent visual word generation. Intuitively, the selected extremal regions are the ones with either largest color pattern diversity or smallest diversity, like homogeneous color regions, in the neighborhood.

3 Constructing visual phrase

In order to improve the discrimination of visual word image representation, we exploit the spatial and co-occurring information of visual words to construct visual phrases from spatially neighboring visual word-sets, in the spirit of [31].

3.1 Mining frequently co-occurring visual word-sets

As introduced in Sect. 2, in the visual word or visual codebook construction phase, the approach first extracts regions from an image and computes visual features of regions a_i to generate visual code $\Omega = \{W_1, \ldots, W_M\}$, where W_i is a visual word. The image $\mathcal{I}$ is then represented by a bag of visual words $\{W_{(a_1)}, \ldots, W_{(a_i)}, \ldots\}$, where $W_{(a_i)}$ is the corresponding visual word of region a_i.

To discover visual phrases, the approach first extracts the spatially neighboring visual word group for each region. For each local region $a_i \in \mathcal{I}$ from visual code construction phase, its local spatial neighborhood $\mathcal{G}$ is defined as a group of its K nearest neighbor regions $\{W_{(a_i)}, W_{(a_{i_1})}, W_{(a_{i_2})} \ldots W_{(a_{i_K})}\}$. By processing all the images in the database $\mathbf{D}_{\mathcal{I}}$, 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 [11]. 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 $\mathbf{G}$ [11, 31]. We explore the FP-growth algorithm to perform the FIM task, as its prefix-tree structure enables 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 $\text{freq}(\mathcal{P}) > \theta$.

3.2 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 visual phrase. 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 [31]. 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 [31] to re-count real instances of word-set through the original image database.

3.3 Statistical significance measure

Yuan et al. [31] 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 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 $L(\mathcal{P})$ of a visual phrase $\mathcal{P}$ is defined as:

$$L(\mathcal{P}) = \text{freq}(\mathcal{P}) \cdot \frac{P(\mathcal{P}|\mathbf{D}_{\mathcal{I}})}{1 + P(\mathcal{P}^{-}|\mathbf{D}_{\mathcal{I}})} \tag{4}$$

where $P(\mathcal{P}|\mathbf{D}_{\mathcal{I}})$ is the probability that the visual word-set $\mathcal{P}$ forms a valid visual phrase and it can be approximated by $\frac{\text{docfreq}(\mathcal{P})}{T}$, where $\text{docfreq}(\mathcal{P})$ is the document frequency equal to number of images containing visual phrase $\mathcal{P}$; and $\mathcal{I}$ is the size of $\mathbf{D}_{\mathcal{I}}$. $\mathcal{P}^{-}$ is the visual word-set $\mathcal{P}$ that does not form any valid visual phrase; 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 $\frac{\text{docfreq}(\mathcal{P}^{-})}{T}$. $\text{freq}(\mathcal{P})$ is the frequency of visual phrase $\mathcal{P}$. Intuitively, we want to penalize the 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, ensure the coherency of visual phrases.

4 Generating visual synset

Though the co-occurrence and spatial scatter information make the image representation more distinctive, the invariance power of visual words or phrases is still low and their effectiveness on image retrieval relies highly on the visual similarity and regularity of images. 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 semantics-consistent cluster of visual lexicons

In text literature, the synonymous words are usually clustered into one synset (**syn**onymy **set**) to improve the document categorization performance [3]. Such approach inspires us to enhance the invariance power of visual lexicons (visual words or phrases). 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 semantics of a visual lexicon in a conceptual manner, we define it probabilistically as follows.

Definition 1 Given image categories $\mathcal{C} = \{c_i\}_{i=1}^{m}$, the **semantics** 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 follows.

Definition 2 The **visual synset** is a probabilistic concept or a semantics-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 rationale 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.

4.2 Information Bottleneck principle

By formulating visual synset construction as a task of visual lexicons 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. [19] proposed to use the relative entropy or Kullback–Leibler (KL) distance to measure the distributional similarity. The KL distance, however, does not possess symmetry property, which is necessary for similarity metric. 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 neglects the fact that clustering is also a process of data compression (compressing a group of data into one clustering).

Here, 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

$$\mathcal{L}\big[P(\mathcal{S}|c)\big] = I(\mathbf{S};\mathcal{C}) - \beta I(\mathbf{V};\mathbf{S}) \tag{5}$$

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, (5) 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 (5) 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:

$$\begin{cases} P(\mathcal{S}) = \sum_{\mathcal{V}} P(\mathcal{S}|\mathcal{V})P(\mathcal{V}) \\ P(c|\mathcal{S}) = \dfrac{1}{P(\mathcal{S})} \displaystyle\sum_{\mathcal{V}} P(\mathcal{S}|\mathcal{V})P(\mathcal{V})P(c|\mathcal{V}) \\ P(\mathcal{S}|\mathcal{V}) = \dfrac{P(\mathcal{S})}{Z(\beta,\mathcal{V})} \exp\big(-\beta D_{\mathrm{KL}}\big[P(c|\mathcal{V})||P(c|\mathcal{S})\big]\big) \end{cases} \tag{6}$$

where $Z(\beta,\mathcal{V})$ is the normalization factor, β is a Lagrange parameter that determines the cluster resolution, and $D_{\mathrm{KL}}[P(c|\mathcal{V})||P(c|\mathcal{S})]$ is the Kullback–Leibler divergence [26] 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 [22] to generate the optimal visual synset clusters in our approach, as it is reported to outperform other IB clustering techniques [22]. 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 (5). 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 $\mathbf{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K\}$ 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 $d_{\mathcal{F}}(\mathcal{V}, \mathcal{S}^{\text{new}})$ of moving visual word $\mathcal{V}$ to a new cluster $\mathcal{S}^{\text{new}}$ can be defined as (cf. [22] for more details):

$$d_{\mathcal{F}}(\mathcal{V}, \mathcal{S}^{\text{new}}) = \big(P(\mathcal{V}) + P(\mathcal{S}^{\text{new}})\big) \cdot \text{JS}(P(c|\mathcal{V}), p(c|\mathcal{S}^{\text{new}})) \qquad (7)$$

where $\text{JS}(x, y) =$ is the *Jensen–Shannon* divergence [26].

4.3 Image retrieval, indexing and similarity measure

With images represented by visual synsets, we index them by exploiting the inverted file scheme [30], due to its simplicity, efficiency and practical effectiveness. The similarity measure adopted here is the L-norm distance defined as

$$\text{Sim}(I_{\text{Q}}, I_{\text{D}}) = \left(\sum_i \left|v_i(I_{\text{Q}}) - v_i(I_{\text{D}})\right|^l\right)^{1/l}, \qquad (8)$$

where I_{Q} is the query image, I_{D} is an image in the database, v_i is the ith dimension of image feature vector, and l is set to 2 in the experiments. In retrieval, all the candidate answer images are ranked by their similarity value with the query image.

5 Experiments and discussion

5.1 Testing data set and experimental setup

We employ the Caltech-256 data set [16] to evaluate the proposed system. The Caltech-256 data set contains 257 image categories and a total of 30607 images. We randomly select 5 images from each class or a total of 257 × 5 = 1285 images as query images. In the phase of region sampling, each image gives 1k to 3k sampled regions. For each region, the SIFT [32] features are computed as region descriptor. We then perform k-means clustering to obtain 2000 primitive visual words in total. To discover visual phrase, we perform FIM on the database $\mathbf{G}$ of approximately 3 million visual word groups of size 8. We construct the visual lexicon codebook of size N_p by selecting the 2000 visual words and top $(N_p - 2000)$ visual phrases with highest scores, based on the significance score in (4). In the experiments, N_p is set to 2100, 2300, 2500, 2600, 2700, 2800, 3000 and 3200, respectively.

5.2 Evaluation criteria: MAP score

The evaluation criteria here is the mean average precision (MAP), which is the mean of average precision (AP) of each query. The AP is the sum of the precisions at each relevant hit in the retrieval list, divided by the total number of relevant images in the collection. AP is defined as:

$$AP = \frac{\sum_{r=1}^{R} \text{Prec}(r) \times \text{rel}(r)}{T} \qquad (9)$$

where r is image rank, R is the total number of images retrieved, $\text{Prec}(r)$ is the precision of retrieval list cut-off at rank r, $\text{rel}(r)$ is an indicator (0 or 1) of the relevance of rank r, and T is the total number of relevant images in the corpus. The average precision is an ideal measure of retrieval quality, which is determined by the overall ranking of relevant images. Intuitively, MAP gives higher penalties to fault retrievals if they have higher position in the ranking list. This is rational, as in practice, searchers are more concerned with the retrieved results in the top.

5.3 Experiments

Performance of visual lexicons We first perform the object-based image retrieval, based on 2000 visual words. This yields a mean average precision (MAP) of 0.026. This retrieval is used as the baseline of our experiments. Next, we perform the image retrieval, based on 2100, 2300, 2500, 2600, 2700, 2800, 3000 and 3200 visual lexicons (visual words and phrases), respectively. As shown in Fig. 4, the performance increases as more visual lexicons are incorporated, up to 2500. In particular, the codebook with 2500 visual lexicons gives the highest accuracy of 0.036. This demonstrates that by incorporating spatial co-occurrence information, the visual lexicons do carry more distinctiveness than visual words alone. When objects share some local appearance similarity in a large scope, the visual phrase can combine the ambiguous visual words scattered in such area into one more distinctive unit. This can contribute to distinguishing objects of different classes with larger interclass distance.

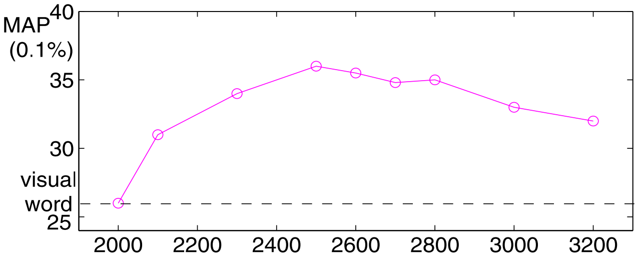


Fig. 4 The mean average precision (MAP) by visual words and visual phrase on Caltech-256 data set

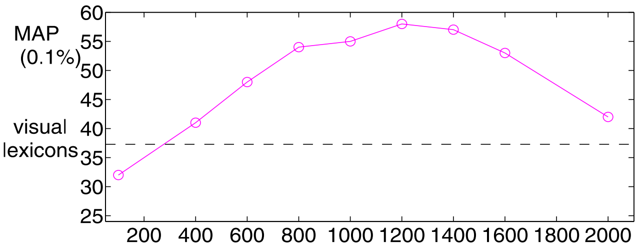


Fig. 5 The mean average precision (MAP) by visual synset on Caltech-256 data set

However, we also observe that when the number of lexicons is above 2600, the performance drops slowly. 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 and brings extra noises to the retrieval.

Performance of visual synset We evaluate the effectiveness of visual synset, by performing IB-based distributional clustering on the codebook of 2500 visual lexicons (best run from previous section). Specifically, we set the cardinality of visual synsets $|\mathbf{S}|$ to 100, 400, 600, 800, 1000, 1200, 1400, 1600, and 2000. As the visual synset is a result of supervised learning, we select 30 images per class as the training set. Figure 5 displays the Mean Average Precision (MAP) of image retrievals based on different number of visual synsets. From Fig. 5, 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 400 visual synsets can achieve a MAP of 0.041, which is superior to the run with 2500 visual lexicons. This representation compactness does not only enable high computational efficiency but also alleviates the curse of dimensionality.

As summarized in Table 1, the best run is the one with 1200 visual synsets and it achieves an accuracy of 0.058. We attribute such improvements to two factors: (1) by fusing semantics-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 retrieval. Figure 7 shows some retrieval examples by words and visual synset representation, respectively. As shown, the relevant images retrieved by visual synset can be visually disparate, while the images retrieved by bag-of-visual-words present local visual similarities with query images (like black texture of motorbike body and black regions of retrieved images). The retrieval via visual similarity can be easily spoiled by large intra-class visual variation, as images of the same class can be fairly distinctive from each other. On the other hand, the visual synset utilizes such distinctiveness to group visual word with consistent relevance to link visually different images of the same class for better retrieval performance.

Table 1 The MAP accuracies of visual word, visual word & phrase, and visual synset, respectively

Visual word	Visual word & phrase	Visual synset
0.026	0.036	0.058

However, after a detailed comparison, we find that 14 classes have visual lexicons delivering better retrieval performance than visual synsets. Figure 6 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 nondistinctive textures and color of objects. This leads to the lack of visual lexicons distinctive to these classes. Consequently, these nondistinctive visual words might be clustered together with visual lexicons indicative of other classes and resulted in nondistinctive 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.

6 Related work

Liu et al. [17] provided a thorough survey on the literature of image retrieval systems, such as QUIB [7], BlobWorld [4], SIMPLcity [28], visualSEEK [23], Virage [10] and Viper [25], etc. The image representation for previous image retrieval systems can be generally classified into 2 types: (1) image-based or grid-based global features like color, color moment, shape or texture histogram over the whole image or grid [7]; and (2) part-based bag-of-words features extracted from segmented image regions, salient keypoints and blobs [4, 12, 23, 24, 33]. The main drawback of global features is their sensitivity to scale, pose and lighting condition changes, clutter and occlusions. On the other hand, the part-based bag-of-visual-words approach is more robust, as it describes an image based on the statistics of local region descriptors (visual words). However, as discussed

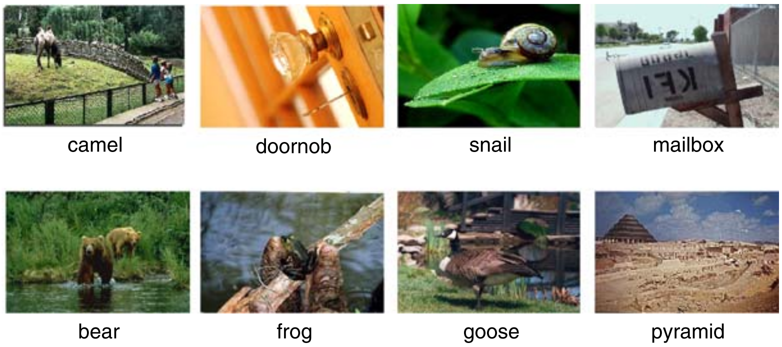


Fig. 6 Example images of classes in which the visual synset yields inferior retrieval performance

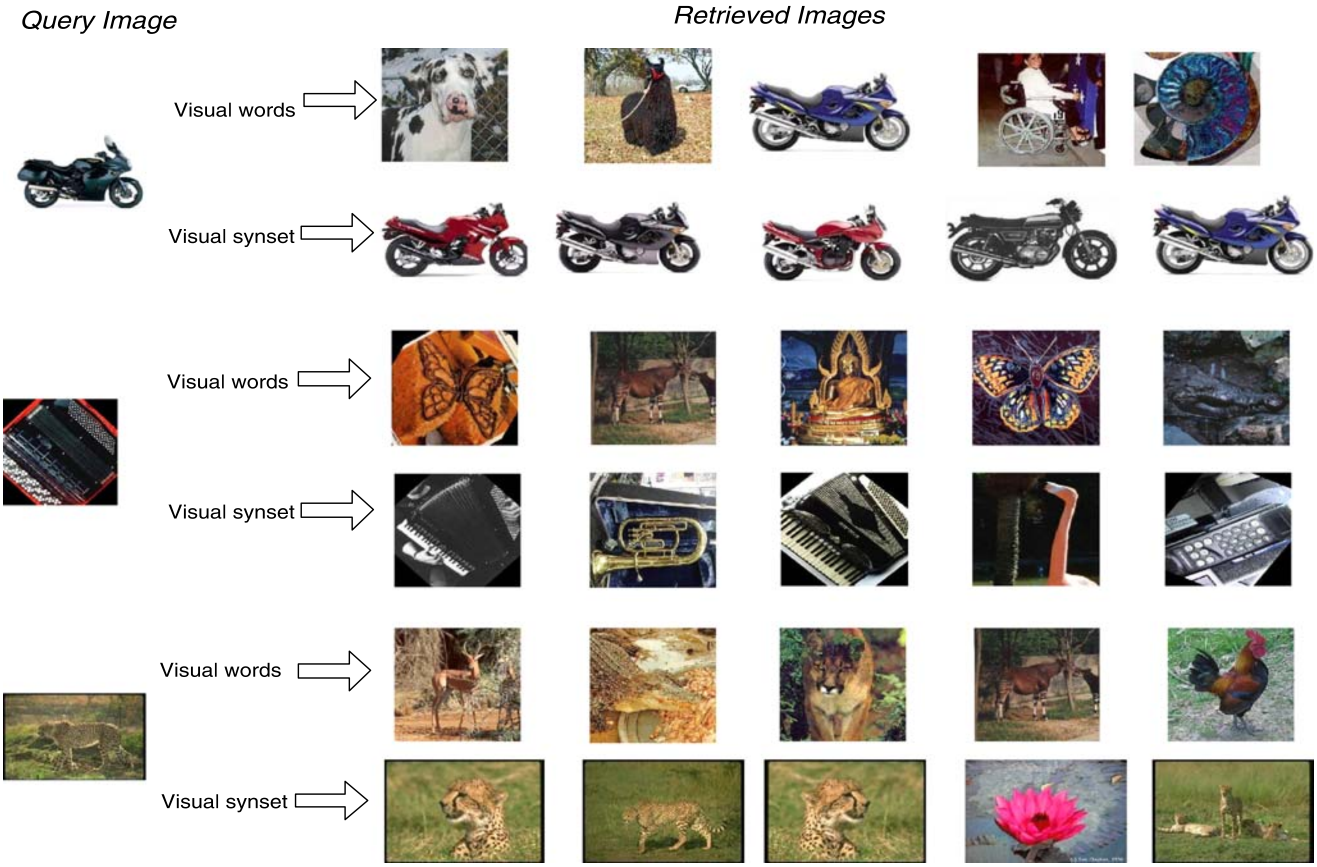


Fig. 7 Example images retrieved based on visual words and visual synsets

in the Introduction, the bag-of-words approach suffers from the discrimination and invariance issues.

To improve the bag-of-words approach, many researchers have proposed various systems. Lazebnik et al. [15] 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 discrimination or polysemy issue of visual words, Juan et al. [31] and Quack et al. [20] proposed visual phrase, i.e. frequently co-occurring visual and spatial configurations.

The performance of primitive visual words and phrases, however, depends highly on visual similarity and regularity. To mitigate such problem, Sivic et al. [21] 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 [1] also demonstrated the effectiveness of LDA in image classification. pLSA and LDA are similar to the proposed visual synset in the way that they are all some kind of intermediate features derived from primitive visual lexicons. However, the proposed visual synset is different from pLSA and LDA in two major aspects.

First, the proposed visual synset is not a latent or hidden semantic variable that connects 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 multimonial distribution over words and the mixture of latent topics per document/image [21]. The Markov condition in pLSA and LDA is to be $\mathbf{V} \leftarrow \mathbf{S} \leftarrow \mathcal{C}$ [26], where $\mathbf{S}$ denotes the latent topic variable. On the contrary, the visual synset is the result 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.

Second, the visual synset is not a generative model. Instead, both pLSA and LDA are unsupervised processes that assume the document/image is a mixture of hidden topics and the word generation in document/image follows some latent topic assignments. The performance of such generative model, however, relies highly on the co-occurrence of latent topic and image semantics observation. Rather than assuming any topic mixture generative model, our proposed visual synset is the result of supervised data-mining on image class probability distributions. It is therefore expected to be more robust than pLSA and LDA.

7 Conclusion

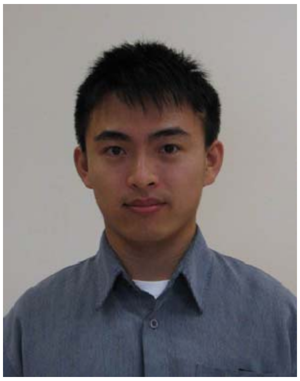
In order to retrieve images beyond their visual appearances, we proposed a higher level image feature, *visual synsets*, for object-based image retrieval. First, we exploit the spatial co-occurrence information of visual words to generate a more distinctive visual configuration, i.e. visual phrase. This improves the discrimination power of visual word representation with better interclass distance. Second, we proposed to group the visual words and phrases with similar 'semantic' into a visual synset. Rather than in a conceptual manner, the 'semantics' of a visual phrase is probabilistically defined as its image class probability distraction. The visual synset is therefore a probabilistic relevance-consistent cluster of visual phrases, which is learned by Information Bottleneck based distributional clustering. The effect of visual synset is to reduce the intra-class variations. The testing on Caltech-256 data set demonstrated that the proposed image representation can achieve good accuracies for object-based image retrieval.

Several open issues remain. First, the generation of visual phrase is a time-consuming task. A more efficient algorithm is demanded. Second, the questions as 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. Agarwal, A., Triggs, W.: Hyperfeatures—multilevel local coding for visual recognition. In: ECCV International Workshop on Statistical Learning in Computer Vision (2006). http://lear.inrialpes.fr/pubs/2006/AT06b/Agarwal-Triggs-eccv06.pdf
2. Baker, L., McCallum, A.: Distributional clustering of words for text classification. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (eds.) Proceedings of ACM SIGIR, pp. 96–103. Melbourne, AU (1998). citeseer.ist.psu.edu/baker98distributional.html
3. Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y.: Distributional word clusters vs. words for text categorization. J. Mach. Learn. Res. G **3**, 1183–1208 (2003)
4. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
5. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: Proceedings of ECCV Workshop on Statistical Learning in Computer Vision (2004)
6. Donoser, M., Bischof, H.: Efficient maximally stable extremal region (MSER) tracking. In: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 553–560 (2006)
7. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. J. Intell. Inf. Syst. **3**(3/4), 231–262 (1994)
8. Griffin, A.H., Perona, P.: Caltech-256 object category dataset. Tech. rep., California Institute of Technology (2007)
9. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: Proceedings of International Conference on Computer Vision, pp. 1458–1465. IEEE Computer Society, USA (2005). http://dx.doi.org/10.1109/ICCV.2005.239
10. Gupta, A., Jain, R.: Visual information retrieval. Commun. ACM **40**(5), 70–79 (1997). http://doi.acm.org/10.1145/253769.253798
11. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: Current status and future directions. Data Min. Know. Discov. **14**(1) (2007)
12. Jing, F., Li, M., Zhang, L., Zhang, H., Zhang, B.: Learning in region-based image retrieval. In: CIVR, pp. 206–215 (2003)
13. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Proceedings of International Conference on Computer Vision. Washington, DC, USA (2005). http://dx.doi.org/10.1109/ICCV.2005.66
14. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001). http://dx.doi.org/10.1023/A:1012460413855
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 2169–2178. Washington, DC, USA (2006)
16. Li, F.F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach based on 101 object categories. In: Proceedings of CVPR Workshop. Washington, DC, USA (2004)
17. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognit. **40**(1), 262–282 (2007). doi:10.1016/j.patcog.2006.04.045. http://dx.doi.org/10.1016/j.patcog.2006.04.045
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **20**, 91–110 (2003).
19. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proceedings of ACL, pp. 183–190. Morristown, NJ, USA (1993). http://portal.acm.org/citation.cfm?id=981598

20. Quack, T., Ferrari, V., Leibe, B., Van-Gool, L.: Efficient mining of frequent and distinctive feature configurations. In: ICCV (2007). http://lear.inrialpes.fr/pubs/2006/AT06b/Agarwal-Triggs-eccv06.pdf
21. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2005)
22. Slonim, N., Friedman, N., Tishby, N.: Agglomerative multivariate information bottleneck. In: Advances in Neural Information Processing Systems (NIPS) (2001). citeseer.ist.psu.edu/article/slonim01agglomerative.html
23. Smith, J.R., Chang, S.F.: Visualseek: a fully automated content-based image query system. In: Proceedings of the ACM International Conference on Multimedia, pp. 87–98. ACM, New York, NY, USA (1996). http://doi.acm.org/10.1145/244130.244151
24. Squire, D., Muller, W., Muller, H., Raki, J.: Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback (1999). citeseer.ist.psu.edu/squire98contentbased.html
25. Squire, D., Muller, W., Muller, H., Raki, J.: Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback (1999). citeseer.ist.psu.edu/squire98contentbased.html
26. Tishby, N., Pereira, F., Bialek, W.: The information bottleneck method. In: Proceedings of Allerton Conference on Communication, Control and Computing, pp. 368–377 (1999). citeseer.ist.psu.edu/tishby99information.html
27. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: Proceedings of International Conference on Computer Vision, p. 257. IEEE Computer Society, Nice, France (2003)
28. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive integrated matching for picture LIbraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001). citeseer.ist.psu.edu/wang01simplicity.html
29. Willamowski, J., Arregui, D., Csurka, G., Dance, C., Fan, L.: Categorizing nine visual classes using local appearance descriptors. In: Proceedings of ICPR Workshop on Learning for Adaptable Visual Systems (2004)
30. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco (1999). citeseer.ist.psu.edu/witten96managing.html
31. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (2007)
32. Zhang, J., Marsza, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vis. **73**(2), 213–238 (2007). http://dx.doi.org/10.1007/s11263-006-9794-4
33. Zheng, Q.F., Wang, W.Q., Gao, W.: Effective and efficient object-based image retrieval using visual phrases. In: Proceedings of ACM International Conference on Multimedia, pp. 77–80. Santa Barbara, CA, USA (2006). http://doi.acm.org/10.1145/1180639.1180664
34. Zheng, Y.T., Zhao, M., Neo, S.Y., Chua, T.S.: Visual synset: towards a higher-level visual representation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA (2008)



Yan-Tao Zheng is a PhD candidate at NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore. He received his bachelor degree from Nanyang Technological University, Singapore, in 2004. He has been awarded with MOE scholarship to fully finance his undergraduate studies and A*STAR scholarship for his PhD studies. He also received Tan Kah Kee Yong Inventors Award in 2008. He served as a conference program committee member of ACM Multimedia 2008 SP Content Track, CIVR 2008 Special session, PCM 2008 and MMM 2009. His research interest is in video and image semantic understanding, which relate to both the field of computer vision and pattern recognition.

Shi-Yong Neo is a PhD candidate at computer science department in the National University of Singapore. His research interests include news video retrieval, interactive video retrieval and mobile multimedia content processing. He is a scholar under Singapore Millennium Foundation since 2005 and has won a number of awards including the Tan Kah Kee Young Inventors Award.

Tat-Seng Chua is the Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School of Computing from 1998 to 2000. He spent three years as a research staff member at the Institute of Systems Science (now I2R) in late 1980s. His main research interest is in multimedia information processing, in particular, on the extraction, retrieval and question answering (QA) of video and text information. He focuses on the use of relations between entities and external information and knowledge sources to enhance information processing. His current projects include: news video retrieval and tracking, question answering (QA), video QA, and information extraction on the web. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations. He obtained his PhD from the University of Leeds, UK.

Dr. Chua is active in the international research community. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the Conference Co-Chair of CIVR'2005, ACM Multimedia 2005, and ACM SIGIR 2008. He serves in the editorial boards of: The Visual Computer (Springer-Verlag) and Multimedia Tools and Applications (Kluwer). He is the member of Steering Committee of Computer Graphics Society (Geneva), and Multimedia Tools and Applications (international), and in Review Panel to a Research

Institute in Europe. In the industry front, Dr. Chua serves as Chair of Board of Assessor of Certified IT Project Management (CITPM) and Certification in Outsourcing Management for IT (COMIT), and as Independent Director of several listed companies in Singapore.

Qi Tian is a principal scientist at Institute for Infocomm Research, Singapore. His main research interests are image/video analysis, indexing and retrieval, computer vision, pattern recognition. He has BS and MS from the Tsinghua University, China, PhD from the University of South Carolina, USA.

He joined the Institute of System Science, National University of Singapore, in 1992, he was the Program Director for the Media Engineering Program at the Kent Ridge Digital Labs, then Laboratories for Information Technology in 2001–2002. He is a senior IEEE member, and has served and serves on editorial boards of professional journals, and as chairs and members of technical committees of the IEEE Pacific-Rim Conference on Multimedia (PCM), the IEEE International Conference on Multimedia and Expo (ICME), etc.