Vicept: Link Visual Features to Concepts for Large-scale Image Understanding

Zhipeng Wu[1,2], Shuqiang Jiang[2], Liang Li[2], Peng Cui[2], Qingming Huang[1, 2], Wen Gao[3]

[1]Graduate University,
Chinese Academy of Sciences,
Beijing, China

[2]Key Lab of Intell. Info. Process.,
Inst. of Comput. Tech., Chinese
Academy of Sciences, China

[3]Institute for Digital Media,
Peking University, China

{zpwu, sqjiang, lli, pcui, qmhuang, wgao }@jdl.ac.cn

ABSTRACT

On noticing the paradox of visual polysemia and concept polymorphism, this paper proposes a new perspective called "Vicept" to associate elementary visual features and cognitive concepts. Firstly, a carefully prepared large image dataset and associate concepts are established. Secondly, we extract local interest points as the elementary visual features, cluster them into visual words, and use Fuzzy Concept Membership Updating (FCMU) to build the link between codebook and concept membership distributions. This bottommost feature is called "Vicept word". Then, the global level Vicept features are established to correlate concepts with (partial) images. Finally, we validate our Vicept approach and show its effectiveness in concept detection task. Our approach is independent of case-specific training data and thus can be extended to web-scale scenarios.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Vision; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement, Image Representation

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Image Understanding , Concept detection, VPCP paradox Vicept , FCMU.

1. INTRODUCTION

Image processing and analysis is a hot research topic in multimedia domain. While how to extract semantic information from images still remains a big challenge. To deal with this issue, researchers have proposed various approaches from different perspectives, such as image classification [1], image annotation [2], image retrieval [3], object recognition [4], etc. Most existing methods are still far from satisfactory either for limited training data or limited concept types. These problems restrict applications on web-scale environment.

To have a deep insight into this issue, there exists the paradox of visual polysemia and concept polymorphism (VPCP paradox) in image semantic understanding. In Figure 1 (a), v is apt to appear in concept: *zebra.* However, the visual appearance v is actually shared by the elements in concept collection {*zebra, shoes, clothes, cup, bag, camera*}. In Figure 1 (b), concept *"brooch"* shows the appearance variety according to different examples. Every visual appearance contains useful information for *"brooch"* and it is the polymorphism of concept that greatly increases the detection difficulty.

Traditional global features such as image color and texture only capture some aspects of visual characteristics, and they normally cannot be directly correlated with image semantics. Recently, local feature has been studied extensively. Image local feature such as SIFT is robust to affine transformations and illumination changes [5], thus it can generate compact and effective image representations. However, many unrelated local features may degrade the performance of concept detection. On the other hand, according to the generation procedure, one visual word may come from different concepts. This is often neglected when detecting concepts via local descriptors. Therefore, the VPCP paradox should be directly investigated to solve the above problems.

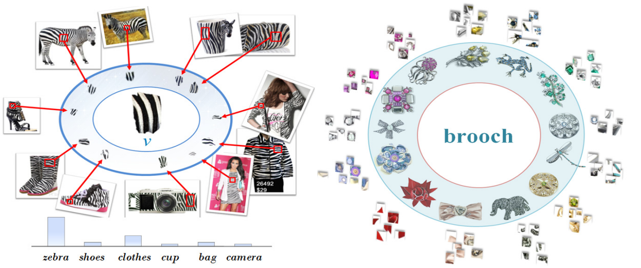


Figure 1 (a) Visual polysemia (b) Concept polymorphism.

Some researches try to understand images from more generous perspective. Liu et al. [6] propose a bi-layer sparse coding formulation operating on the over-segmented images to uncover how these segments reconstruct the semantic region, which can be used for automatic label to region scheme. Sivic et al. [7] discusses the problem of visual polysemia, and proposes an LDA-based method to group codewords in spatial proximity. The same codeword found in two different contexts could be differentiated. These works provide new insights on image analysis, however, most of them are in the primary research stage and they are limited for wide applications on large dataset. The 80 million large image dataset proposed by Torralba et al. [8] is an innovative idea by comparing the concept-unknown images with the established representative concept-known-small-image-set. This method could work well for image global concepts; while many locally existed concepts may be neglected yet.

This paper proposes a new perspective to process and represent images based on a basic proposition that each local visual appear-

MM' 10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10…$10.00.

ance in an image has various possibilities to be related with a serious of semantic concepts. In this procedure, a purified image dataset is first established based on a concept hierarchy covering frequently used concepts. Then SIFT descriptors are extracted and representative visual words are clustered accordingly. Thus the link between visual word and concept distribution possibilities can be built by a proposed Fuzzy Concept Membership Updating (FCMU) method. In another word, a visual word can be represented by a concept histogram. This local representation is called "Vicept word". We can further compute the global level Vicept description on the images or part of images. The proposed Vicept avoids the traditional "feature extraction-concept learning" schema by directly connecting feature and concepts, and forms a hierarchical representation of image semantic from local to global. Semantic concepts at different areas and scales in an image can be represented by Vicept description and thus can directly deal with VPCP paradox. For any given image, the proposed scheme quickly computes its semantic information without depending on specified training model. It can be potentially used for large scale applications without relying on any outside information.

2. VICEPT GENERATION

2.1 Overview of Image Vicept Feature

The observation of VPCP paradox motivates us to design Vicept with the following details:

1. *Local visual appearance:* We adopt local interest points as image local representation. In our approach, SIFT [5] is detected and quantized into visual words [9] by a modified *k*-means algorithm.
2. *Concept collection:* The concepts in real world are not independent but closely related. Therefore, we simplify the concept modeling with a hierarchical representation [10].

Having an *m*-size feature vocabulary and a hierarchical concept collection, "Vicept" is proposed to associate visual features and cognitive concepts. On each concept layer, there are a series of concept membership distributions. In all, the combination of multi-level representation of these memberships and the original visual word cluster center is called "Vicept word". Figure 2 illustrates a typical Vicept word.

2.2 Image Dataset and Concepts

The concept collection in our dataset is organized by a semantic hierarchy which is used by ImageNet [10]. We use ImageNet as the source of our dataset by searching the concept names and downloading the returned images according to their URLs. However, to ensure enough training images and for the sake of limited computation capacity, we manually filter the concepts with less than 1k images returned by ImageNet and select a frequently-used collection with 217 concepts. In all, there are 267k images in our dataset and we use a simple 3-level concept tree structure: 10 concepts on level-1, 88 on level-2, and 217 on level-3.

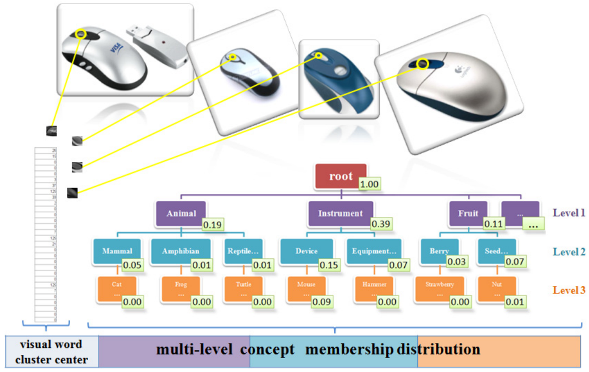


Figure 2. Vicept word.

Although ImageNet offers clean image annotation, the generation of concept labeled local features is also "unclean". This is because the interest points are generated from both the foreground and background. On noticing this fact, we try to "purify" the dataset by manually segment the image and eliminate the irrelevant areas. Balancing the workload for image matting and the data requirement in this task, we prepare a subset of the original 260k image dataset with 120 "purified" images in each concept. This subset contains 120×217 one-concept-labeled images and it is only used for generating Vicept words. Figure 3 illustrates the result for the "purification" of images.

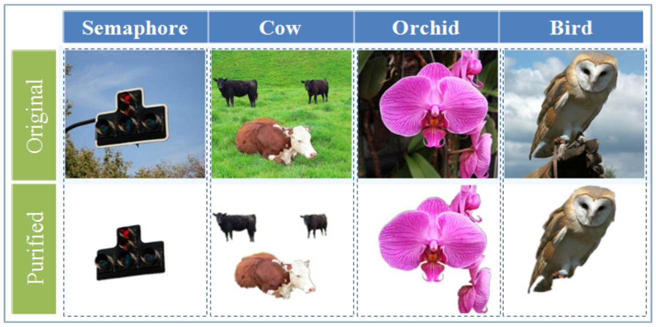


Figure 3. Examples for image matting.

2.3 Vicept Word with Multi-Level Concept Membership Distribution

We first extract interest points (SIFT) in images and adopt a modified *k*-means algorithm to cluster them into visual words. Then we use a novel iteration algorithm based on fuzzy theory to distribute the visual features into the concepts. Finally, the multi-level concept membership distribution histograms are established and further combined with the cluster center into Vicept words.

2.3.1 Optimizing Initial Cluster Centers by Concept Distribution Density for K-means Clustering

Intuitively, performance and computational costs of the iterative *k*-means clustering depend highly on initial cluster centers. The basic idea of us is to pick the points with large concept label distribution density into the candidate set and select initial cluster centers from this set. In the *k*-means iteration, we employ Extended Partial Distortion Search (EPDS [11]) to distribute each point to its nearest cluster center. EPDS is a fast search algorithm. It increases a little number of comparisons and memory space in order to save the multiplications and additions. Based on the *k*-means clustering, we obtain ***w*** visual words. The visual word codebook and interest points are used to generate Vicept words.

2.3.2 Generating Vicept Word Based on FCMU

The clustering step actually assigns the training interest points into ***w*** visual words. As mentioned above, the process to link the visual features to the concepts requires "soft" classification which resorts to a set of probability distributions. Supposing we have **n** bottom-level concepts, we aim to build the **n**-bin concept membership distribution histograms for each of the visual words. Motivated by the idea of fuzzy objective function [12], we propose Fuzzy Concept Distribution Updating (FCMU) as a novel iterative algorithm to extract the concept membership distribution histogram.

Given a visual word cluster center $a \in \mathbf{R}^u$ and the interest points which are distributed into this visual word, according to the concept labels, these interest points are sorted into **n** subsets. We then define the real distance $rDis_i$ from a to the i th concept subset B_i :

$$rDis_i = Dis(a,b_i)/(Num_i)^z$$

where b_i is the geometrical center for all the interest points in B_i and $Dis(a,b_i)$ computes Euclidean distance. Num_i indicates the number of points in B_i , and z is a decimal ranging from 0 to 1.

After defining the real distance from cluster center a to the concept subsets in the visual word, we will optimize a minimum intra-cluster real distance to obtain the **n**-bin fuzzy concept membership distribution histogram $\mathbf{U} = \{u_1, u_2, ..., u_n\}$. The optimization goal is defined as:

Intra-cluster real distance $J_m(\mathbf{B},\mathbf{U})$ *:*

$$J_m(\mathbf{B},\mathbf{U}) = \sum_{i=1}^{n} (u_i)^m \times rDis_i \qquad s.t. \ \sum_{i=1}^{n} u_i = 1$$

where *m* is the fuzzy smoothing factor, and we usually set it to 2.

$$\min(J_m) = \min[\sum_{i=1}^{n} (u_i)^m \times rDis_i]$$

Using Lagrange multiplier method, we define function **F**:

$$\mathbf{F} = \sum_{i=1}^{n} (u_i)^m \times rDis_i + \lambda \times (\sum_{i=1}^{n} u_i - 1)$$

$$\begin{cases} \partial F / \partial u_i = [\, m \times u_i^{m-1} \times rDis_i + \lambda \,] = 0 \\ \partial F / \partial \lambda = [\, \sum_{i=1}^{n} u_i - 1 \,] = 0 \end{cases}$$

Solution:

$$u_i = \frac{1}{\sum_{k=1}^{n} [\frac{rDis_i}{rDis_k}]^{\frac{1}{m-1}}} \qquad (1)$$

If we regard cluster center a as shiftable and update it to minimize $J_m(\mathbf{B},\mathbf{U})$:

$$\partial J_m(\mathbf{B},\mathbf{U}) / \partial a = 0$$

Solution:

$$a = \frac{\sum \frac{u_i}{(Num_i)^z} \times b_i}{\sum \frac{u_i}{(Num_i)^z}} \qquad (2)$$

Based on the solutions in equation 1 and 2, we can iteratively update the cluster center a and the **n**-bin fuzzy concept membership distribution histogram $\mathbf{U}$.

2.3.3 Establishing Multi-Level Concept Membership Distribution

Vicept words employ hierarchical structure for concept organization. Based on FCMU, we obtain the bottom-level concept membership distribution histogram. To establish the high-level concept memberships, rather than simply sum the bins belonging to the same bottom-level concept, we assign large weights to the bins with large bin value. In fact, we implement this by adopting sigmoid function and normalize the sum of high-level histogram into 1. Equation 3 is the sigmoid function where the weight for the *i*th bin $w(i)$ is calculated by its bin value $v(i)$.

$$w(i) = 1/(1+\exp\{-v(i)\}), \qquad v(i) \in [0,1] \qquad (3)$$

3. IMAGE VICEPT REPRESENTATION AND SIMILARITY MEASUREMENT

3.1 Image Vicept Representation

We aim to represent image (or part of the image for tasks such as partial annotation) into a concept membership distribution histogram in which the larger bin value denotes the higher probability of concept existence. This can be implemented by:

$$\Pr(C_i \mid I) = \sum_{j=1}^{n} \Pr(C_i \mid Vw_j) \times \Pr(Vw_j \mid I) \qquad (4)$$

Vw_j *(j=1,2,...,n)* denotes the n Vicept words generated by algorithm 1 and 2. *I* is the input image (or partial image) and C_i is the *i*th concept. The multiplicand $\Pr(C_i \mid Vw_j)$ at the right side can be directly accessed according to the Vicept word *j*; the multiplicator $\Pr(Vw_j \mid I)$ can also be solved by counting Vicept words in the image. To a certain image, firstly we detect its interest points (SIFT), then quantize them into Vicept words according to the Vicept vocabulary generated in section 2. Finally, the image is represented as multi-level concept membership distribution histograms (equation 4). Note that our approach already embeds semantic information into the Vicept feature. After detecting the interest points, the multi-level histograms directly depict the existence probability for the concepts.

3.2 Image to Image Distance Based on Vicept

According to Vicept, Image is represented as multi-level histograms. Intuitively, we cannot concatenate the histograms into one and calculate the classical histogram distances (e.g. intersection, chi-square, Minkowski-form) because the concepts at different levels are incomparable and with different discriminative powers. Therefore, we set different weights for histogram distances at different levels.

Suppose there are *l* levels histogram (level = 1,2,…,*l*). The weight for level *i* is defined as:

$$w_i = \frac{(m_i^{\mathrm{intra}} - m_i^{\mathrm{inter}})^2}{(v_i^{\mathrm{intra}} + v_i^{\mathrm{inter}})} \qquad (5)$$

where $m_i^{intra}, m_i^{inter}, v_i^{intra}, v_i^{inter}$ stand for *Intra-Concept Mean*, *Inter-Concept Mean, Intra-Concept Variance, Inter-Concept Variance* for the *i*th level respectively. And we find the best projection direction to fuse multi-level histogram distances.

4. EXPERIMENTS

4.1 Dataset

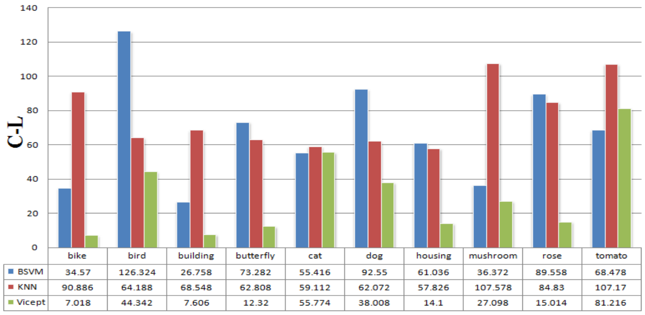
Although there are a lot of preeminent publicly available datasets such as NUS-WIDE [13], Pascal VOC [14], we do not use them simply because of the differences in concept coverage. To be more specific, we choose 10 bottom-level concepts which are intersected with Flickr's all time most popular tags [15]. Then the noises in the returned searching results (Google Images) are removed and we prepare 500 images for each concept. This provides a 5000-image dataset categorized into 10 concepts: “bike”, “bird”, “building”, “butterfly”, “cat”, “dog”, “housing”, “mushroom”, “rose”, and “tomato”.

4.2 Baseline Approaches

Two basic approaches are implemented as baseline for concept detection. (1) Binary SVM; (2) KNN based voting. Following [6]’s solution, we prepare 217 binary SVM classifiers with an output of classification probability. For the KNN approach, we replicate the experiment described in [8].

4.3 Evaluation Metric

For every query image, three 217-dimension concept membership vectors are obtained based on BSVM, KNN, and the bottom-level concept membership distribution histogram in Vicept approach. Then, Complete-Length (CL, [16]) is adopted as the performance measurement. Complete-Length is defined as the average minimum length of returned concepts which contain the correct label for query. We implement CL by sorting the 217-dimension vectors in descending order and traverse the concept labels. Figure 4 shows the experimental results.



	bike	bird	building	butterfly	cat	dog	housing	mushroom	rose	tomato
🟦 BSVM	34.57	126.324	26.758	73.282	55.416	92.55	61.036	36.372	89.558	68.478
🟥 KNN	90.886	64.188	68.548	62.808	59.112	62.072	57.826	107.578	84.83	107.17
🟩 Vicept	7.018	44.342	7.606	12.32	55.774	38.008	14.1	27.098	15.014	81.216

Figure 4. Complete-Lengths for experiment.

Table 1 illustrates the average CLs for three approaches. According to Figure 4, Vicept provides a comparatively better concept detection result. The proposed Vicept achieves small CLs (less than 20) on half of the total concepts (“bike”, “building”, “butterfly”, “housing”, “rose”), which seems to be satisfactory in this task. Besides, the fluctuations on concept “cat” and “tomato” are likely to be influenced by small number of test data and insufficient available data for Vicept word generation.

Table 1. Average Complete-Lengths

	BSVM	KNN	Vicept
Complete-Length	66.4344	76.5018	30.2496

5. CONCLUSION

There is a saying “a picture is worth a thousand words”. In this paper, we propose a new method to “interpret” an image into its “semantic words” (concept). On noticing the VPCP paradox, the local visual appearances are represented as multi-level concept membership distribution histograms. Then Vicept representation is constructed to correlate concepts with images. The Vicept approach provides fast computation, compact expression and local-to-global description, thus can be implemented for generic large scale web applications. In the future, we will enlarge the initial processing image dataset and concept types to generate more powerful Vicept descriptors. The co-occurrence of concepts in the same image will also be taken into account.

6. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China: 60833006 and 60702035, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042.

7. REFERENCES

[1] J. Wang, J. Yang, K. Yu, F. Lv, T.S. Huang, and Y. Gong. Learning Locality-constrained Linear Coding for Image Classification, Proc. CVPR, 2010.

[2] Y. Xiang, X. Zhou, T. Chua, and C. Ngo. A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. Proc. CVPR, pp. 1153-1160, 2009.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, Vol.40, No.2, pp.51-60, 2008.

[4] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Proc. ECCV, pp.349-354, 2002.

[5] D. G. Lowe. Distinctive image features from scale invariant keypoints. International Journal of Computer Vision, 60 (2): 91-110, 2004.

[6] X. Liu, B. Cheng, S. Yan, J. Tang, T. Chua, and H. Jin. Label to region by bi-layer sparsity priors. Proc. ACM Multimedia, pp.115-124, 2009.

[7] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering Objects and their Location in Images. Proc. ICCV, pp.370-377, 2005.

[8] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (11):1958-1970, 2008.

[9] J. Sivic, and A. Zisserman, Video Google: A text retrieval approach to object matching in videos. Proc. ICCV, pp.1470-1477, 2003.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. Proc. CVPR, pp.248-255, 2009.

[11] S. H. Chen and J. S. Pan. Fast search algorithm for VQ-based recognition of isolated words. Communications, Speech and Vision, IEEE Proceedings I, 136 (6), 391-396.

[12] J. C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, 1981.

[13] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of Singapore. ACM International Conference on Image and Video Retrieval, No.48, 2009.

[14] M. Everingham, L.Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, Vol.88, pp.303-338, 2010.

[15] Flickr all time most popular tags. www.flickr.com/photos/tags/

[16] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma and H.-J. Zhang. A probabilistic semantic model for image annotation and multimodal image retrieval. Proc. ICCV, pp.846–851, 2005.