Harvesting Mid-level Visual Concepts from Large-scale Internet Images

Quannan Li[1], Jiajun Wu[2], Zhuowen Tu[1]
[1]Lab of Neuro Imaging and Department of Computer Science, UCLA
[2]Institute for Interdisciplinary Information Sciences, Tsinghua University

Abstract

Obtaining effective mid-level representations has become an increasingly important task in computer vision. In this paper, we propose a fully automatic algorithm which harvests visual concepts from a large number of Internet images (more than a quarter of a million) using text-based queries. Existing approaches to visual concept learning from Internet images either rely on strong supervision with detailed manual annotations or learn image-level classifiers only. Here, we take the advantage of having massive well-organized Google and Bing image data; visual concepts (around $14,000$*) are automatically exploited from images using word-based queries. Using the learned visual concepts, we show state-of-the-art performances on a variety of benchmark datasets, which demonstrate the effectiveness of the learned mid-level representations: being able to generalize well to general natural images. Our method shows significant improvement over the competing systems in image classification, including those with strong supervision.*

1. Introduction

The inventions of robust and informative low-level features such as SIFT [18], HOG [4], and LBP [22] have been considered as one of the main advances/causes for the recent success in computer vision. Yet, one of the most fundamental issues in vision remains to be the problem of "representation", which affects an array of applications, such as image segmentation, matching, reconstruction, retrieval, and object recognition.

Beyond low-level features, obtaining effective mid-level representations has become increasingly important. For example, there have been many recent efforts made along the line of attribute learning [8, 24, 17]. These approaches, however, are mostly focused on supervised or active learning where a considerable amount of human efforts are required to provide detailed manual annotations. The limitations to the previous supervised attribute learning methods are thus three-fold: (1) accurate data labeling is labor-intensive to obtain, (2) the definition of attributes is often intrinsically ambiguous, (3) the number of attributes and training images are hard to scale. Some other methods in which detailed manual annotations are not required (e.g. classemes [11]) however are not designed to build a dictionary of mid-level representations.

In this paper, we propose a scheme to build a path from words to visual concepts; using this scheme, effective mid-level representations are automatically exploited from a large amount of web images. The scheme is inspired by the following observations: (1) search engines like Google and Bing have a massive number of well organized images; (2) using text-based queries, such as "bike", "bird", "tree", allows us to crawl images of high relevance, good quality, and large diversity (at least for the top-ranked ones); (3) the multiple instance learning formulation [1] enables us to exploit common patterns from retrieved images, which have a high degree of relevance to the query words; (4) saliency detection [9] saliency detection helps to reduce the search space by finding potential candidates. The main contributions of this paper thus include the following aspects: (1) we emphasize the importance of automatic visual concept learning from Internet images by turning an unsupervised learning problem into a weakly supervised learning approach; (2) a system is designed to utilize saliency detection to create bags of image patches, from which mixture concepts are learned; (3) consistent and encouraging results are observed by applying the learned concepts on various benchmark datasets.

2. Related Works

Visual attribute learning has recently attracted a lot of attention. However, many existing algorithms were designed as supervised approaches [8, 24, 17, 25, 17], preventing them from scaling up to deal with a large number of images.

A term, "classeme", was introduced in [29] which also explores Internet images using word-based queries; however, only one classeme is learned for each category and the objective of the classeme work is to learn image-level representations. Instead, our goal here is to learn a dictionary

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.115

IEEE computer society

of mid-level visual concepts for the purpose of performing general image understanding, which goes out of the scope of classeme [29] as it is computationally prohibitive for [29] to train on a large scale.

A recent approach [28] learns "discriminative patches" in an unsupervised manner. However, [28] learns discriminative patches while we focus on dictionary learning for the mid-level representations; [28] uses an iterative procedure, while our method adopts saliency detection, miSVM and K-means in a novel way; in addition, our method significantly outperforms [28] with a relative 37% improvement on the MIT-Indoor scene dataset, on which both the approaches have been tested. In [15], high-level features are built from large scale Internet images with nine layers of locally connected sparse autoencoder; however, their autoencoder approach is much more complex than the scheme proposed in this paper. In [37], saliency detection is utilized to create bags of image patches, but only one object is assumed in each image for the task of object discovery. Although multiple clusters are learned in [34], its goal is to identify a few cancer patterns for medical image segmentation; in addition, the lack of explicit competition among clusters leads to poor results in our problem. In terms of large-scale natural images, ImageNet [5] is shown to be a great resource. Here, we find it convenient to directly crawl images from the search engines using word-based queries.

3. Automatic Visual Concept Learning

Starting from a pool of words, we crawl a large number of Internet images using the literal words as queries;patches are then sampled and visual concepts are learned in a weakly supervised manner. The flow chart of our scheme is illustrated in Fig. 1. Following our path of harvesting visual concepts from words, many algorithms can be used to learn the visual concepts. In this paper, we adopt a simple scheme, using the max-margin formulation for multiple instance learning in [1] to automatically find positive mid-level patches; we then create visual concepts by performing K-means on the positive patches. The visual concepts learned in this way are the mid-level representations of enormous Internet images with decent diversity, and can be used to encode novel images and to categorize novel categories. In the following sections, we introduce the details of our scheme.

3.1. Word Selection and Image Collection

The literal words are selected from ImageNet [5], which is based on WordNet [19] and Classeme [29]. For the words with similar meanings, *e.g.*, "people", "guest", "worker", and "judge", we keep the most generic one. In all, $M = 716$ words are selected. Most of the words are representative ones of the popular categories in ImageNet such as "animal", "plants", "scenes", "activities", "foods", and "materials". For each word, we crawled the top 400 images from google.com and the top 30 images from bing.com and merged the images by removing the duplicates. For each category (word), around 400 images are retained.

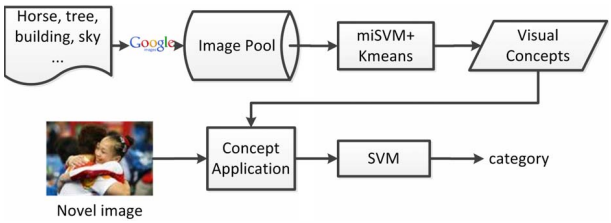


Figure 1. The flow chart of our scheme of creating visual concepts from words.

Fig. 2 shows the top ranked images for 26 words. From Fig. 2, we can see that most of the retrieved images are generally of high relevance to the query word. Also, these images provide sufficient diversity stemming from the intra-category variances. For example, for the word "table", besides the images of dinning tables, images of spreadsheets appear as well. The retrieved images for words such as "video" and "bird" are even more diverse. The diversity in these crawled images makes it inappropriate to train only a single classifier on the images, forcing us to investigate the multiple cluster property. Further, the object of interest usually does not occupy the entire image, making the multiple instance learning formulation a natural fit for this task.

3.2. Saliency Guided Bag Construction

The problem of visual concept learning is firstly unsupervised because we did not manually label or annotate the crawled images. However, if we view the query words as the labels for the images, the problem can be formulated in a weakly supervised setting, making our problem more focused and easier to be tackled.

Firstly, we convert each image to a bag of image patches with size greater than or equal to 64×64 that are more likely to carry semantic meanings. Instead of having randomly or densely sampled patches as th in [28], we adopt a saliency detection technique to reduce the search space. Saliency detection assumes that the object of interest is generally salient in an image. Fig. 3 shows sample saliency detection results (the top 5 saliency windows for each image) by [9], a window based saliency detection method. From Fig. 3, we observe that within the top 5 saliency windows, objects such as airplanes, birds, caterpillars, crosses, dogs, and horses are covered by the saliency windows. In addition, for the airplane and the caterpillar, the salient windows naturally correspond to the parts. This illustrates the benefit of the use of saliency detection: it helps to identify the regions and parts with more significance naturally. In our experiment, the top 50 salient windows are used as the instances of a positive bag directly. For large salient windows with



Figure 2. Sample images collected for 26 words. In the left column, the words from the top row to the bottom row are “abbey”, “airport”, “armchair” “balloon”, “beach”, “bird”, “bride”, “building”, “eagle”, “gun”, “table”, “video”, and “wolf”, respectively; in the right column, the words are “airplane”, “ambulance”, “balcony”, “bar”, “bicycle”, “bookshelf”, “bridge”, “computer_monitor”, “ferry”, “horse”, “tiger”, “window”, and “yard”, respectively.

sizes greater than 192×192, smaller patches within them are sampled, resulting in possible parts of the relevant patterns.

Although the saliency assumption is reasonable, not all category images satisfy this assumption. For example, for the images of “beach”, the salient windows only cover patterns such as birds, trees, and clouds (see the salient windows of the “beach” image in Fig. 3). Although these covered patterns are also related to “beach”, they cannot capture the scene as a whole because an image of “beach” is a mixture of visual concepts including sea, sky, and sands. To avoid missing non-salient regions for a word, besides using the salient windows, we also randomly sample some image patches from non-salient regions. As non-salient regions are often relatively uniform with less variation in the appearance, a smaller number of patches are sampled from the regions. After the patches are sampled, we perform overlap checks between the image patches with similar scale is performed. If two patches are of the similar scale and have high overlap, one patch will be removed.

Each bag constructed in this way thus consists of patches from both salient and non-salient regions. A portion of the patches may be unrelated to the word of interest, e.g., the patches corresponding to the sea in the image of “horse” in

Fig. 3. Such patches are uncommon for the word "horse", and will be naturally filtered under the multiple instance learning framework.

Thus, for a word with N ($N \approx 400$) images, N bags $\{B_i, 1 \leq i \leq N\}$ can be constructed, each bag $B_i = \{\mathbf{x}_{ij}, 1 \leq j \leq m\}$, where m is the number of patches sampled for the image and is about 150 in our paper; $\mathbf{x}_{ij}$ is the descriptor of the patch.

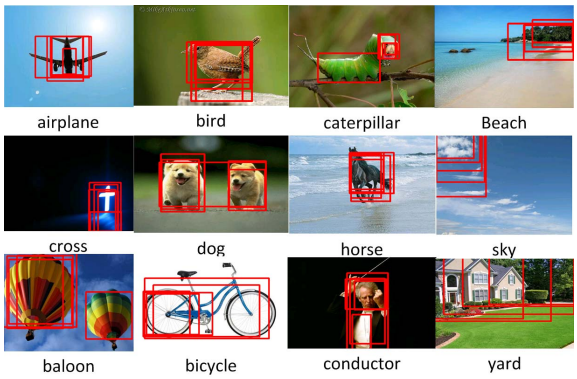


Figure 3. Top five salient windows for images from 12 words. Except for the words "sky", "beach", and "yard", the patterns of interest can be covered by a few top salient windows. For objects such as "caterpillar", "bicycle", and "conductor", parts can be captured by the salient windows.

3.3. Our Formulation

To learn visual concepts from the bags constructed above, there are two basic requirements: 1) the irrelevant image patches should be filtered, and 2) the multiple cluster property of these visual patches should be investigated. Many methods meet these two requirements. In this paper, we simply use the max-margin framework for multiple instance learning (miSVM) in [1] to learn a linear SVM for each word, and then perform clustering on the positive instances labeled by the linear SVM. It is worth mentioning that another formulation for learning the multiple instance multi-classes problem can also be used, but it is not the main focus of this paper.

In multiple instance learning, the labeling information is significantly weakened as the labels are assigned only to the bags with latent instance level lables. In [1], the relationship between the bag level labels and the instance level labels is formulated as a set of linear constraints. With these linear constraints, soft-margin SVM is formulated into a mixed integer programming problem, which can by solved heuristically by iterating two steps: 1) given the instance level label y for an instance $\mathbf{x}$, solving the optimization discriminant function $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ via Quadratic programming, where $\mathbf{w}$ is the weight vector, and b is the bias term and 2) given the discriminant function f, updating the instance level labels y. For more details on miSVM, the readers can refer to [1].

3.3.1 Visual Concept Learning via miSVM

Using miSVM and assigning the literal words as the labels for the Internet images, visual concept learning for each word can be converted from an unsupervised learning problem into a weakly supervised learning problem. For a word k, its bag B_i is assigned with a label $Y_i = 1$. The instance level label y_{ij} for each instance $\mathbf{x}_{ij} \in B_i$ is unknown and will be automatically discovered by miSVM. For negative bags, we create a large negative bag B^- using a large amount of instances (patches) from words other than the word of interest. The number of instances in B^- is generally $5 \sim 10$ times more than the number of all the instances in the positive bags. The purpose of creating the large negative bag is to model the visual world, making the visual concepts learned for a word discriminant enough from the other words. For example, for words such as "horse" and "cow", using a large negative bag B^-, the common backgrounds such as the grassland and the sky can be filtered.

Based on $\{B_i, 1 \leq i \leq N\}$ and B^-, a linear SVM f^k can be learned by miSVM for the k-th word. The positive patches related to the word are also automatically found by miSVM. Given a patch, the linear SVM f^k can output a confidence value indicating the relevance of the patch to the word of interest. Therefore, the linear SVM f^k itself can be treated as a visual concept that models the patches of a word as a whole. We call it a *single-concept classifier*. Due to the embedded multi-cluster nature of diversity in the image concepts, a single classifier is insufficient to capture the diverse visual representations to a word concept. Thus, we apply another step in our algorithm: the positive instances (patches) automatically identified by the single-concept classifier are clustered to form some codes $C^k = \{C_1^k, C_2^k, ..., C_n^k\}$. We call these codes multi-cluster visual concepts. Different from the single-concept classifier, each multi-cluster visual concept corresponds to a compact image concept.

Therefore, for each word k, we learn two types of visual concepts, the single-concept classifier and the multi-cluster visual concepts C^k. From Internet images of the M words, we can learn M single-concept classifiers $F = \{f^1, ..., f^M\}$, and a set of multi-cluster visual concepts $C = \{C^k, 1 \leq k \leq M\}$. The single-concept classifiers and the visual concepts can be applied to novel images as the descriptors for categorization.

In Fig. 4, we illustrate the outputs of the single-concept classifiers on the images, as well as the assignments of patches to the multi-cluster visual concepts. For clarity, for each word we cluster six multi-cluster visual concepts from the positive patches and assign them different colors randomly. For each image, we sample dense mid-level patches

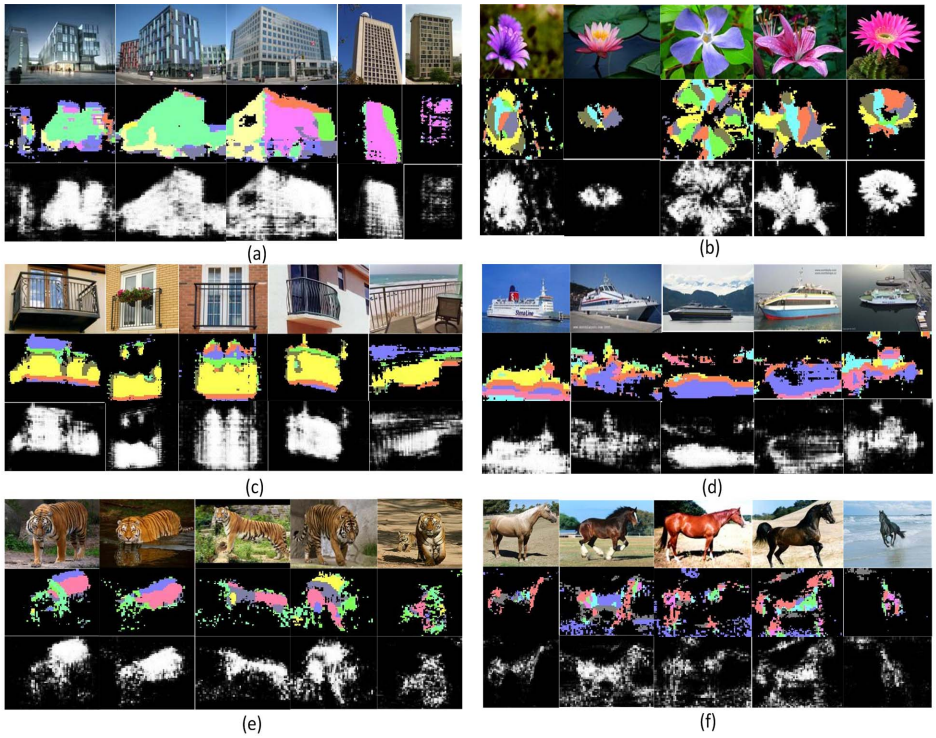


Figure 4. Illustration of the single-concept classifiers and the multi-cluster visual concepts for 6 words. (a) "building", (b) "flower", (c) "balcony", (d) "ferry", (e) "tiger", and (f) "horse". For each word, the first row shows the original images; the second row shows the assignment of the codes, and the third row shows the outputs of the single-concept classifier (the linear SVM for each word). See text for details.

and apply the single-concept classifier to label the patches. We then assign the patches labeled as positive to the six multi-cluster visual concepts in a nearest neighborhood manner and display the colors of the assigned visual concepts in the centers of the patches.

The third row in Fig. 4 shows the outputs of the single-concept classifiers. Though learned in a weakly supervised manner, the single-concept classifiers can predict rather well. However, it cannot capture the diverse patterns of the patches, e.g., for the word "building", the walls of the left two images are different from the walls of the right three images. On the contrary, the multi-cluster visual concepts can capture such differences. The walls in the left two images of the word "building" have the same patten, and they are assigned to the same multi-cluster visual concept that has relatively sparse and rectangle windows (indicated in green). The walls on the right three images have a different pattern and they are assigned to another visual concept that has square and denser windows (indicated in magenta). For the word "balcony", the columns are assigned to a multi-cluster visual concept indicated in yellow. For the other four "words", the objects of interest are generally a combination of several multi-cluster visual concepts. This illustrates that the single-concept classifiers and the multi-cluster visual concepts correspond to different aspects of images and complement each other.

3.4. Application for Image Classification

As our visual concept representation has two components, the single-concept classifiers $F = \{f^1, ..., f^M\}$ and the multi-cluster visual concepts $C = \{C^k, 1 \leq k \leq M\}$, we apply the two components separately on novel images. Each novel image is divided into grids of a three-level spatial pyramid [14]. The single-concept classifier f^k is applied to the densely sampled patches from the grids, and the responses of the classifiers are pooled in a max-pooling manner. For natural images, the objects are generally varying in different scales, and we run the classifiers on the novel images on these scales. Since our method works on the patch level and the visual concepts are learned with image patches of different scales, two or three scales are enough for testing images. The pooled responses across different scales are concatenated, leading to a feature vector with dimension $M \times 21 \times$ number of scales.

We use the multi-cluster visual concepts $C = \{C^k, 1 \leq k \leq M\}$ in a simple way as a single codebook in a spatial pyramid matching manner (SPM) [14]: multi-scale mid-level patches are assigned to the multi-cluster visual concepts via hard assignment; a histogram is constructed for each grid of the three-level spatial pyramid and the fea-

ture is the concatenated version of the histograms of all the multi-cluster visual concepts. In this way, for each novel image, we obtain a feature vector of dimension $M \times n \times 21$, where n is the number of multi-cluster visual concepts for each word.

Definitely, there are several other options. One is to train a linear classifier model for each visual concept, and apply the classifiers to the novel images. In this paper, we simply use the basic scheme to illustrate the effectiveness of the visual concepts we learned.

Finally, the features corresponding to the single-concept classifiers and the multi-cluster visual concepts are combined like multiple kernel learning [2, 13]. The kernels K_F for the single-concept classifiers and K_C for the multi-cluster visual concepts are computed respectively and combined linearly: $K = wK_F + (1-w)K_C$. In our paper, as there are only two kernels, instead of using advanced techniques such as the SMO algorithm in [2], we can simply use cross-validation to determine the best w^*. χ^2 kernel is used in the experiments, and it can be computed efficiently using the explicit feature map in [31, 30].

4. Experiments and Results

On the PASCAL VOC 2007 [6], scene-15 [14], MIT indoor scene [27], UIUC-Sport [16] and Inria horse [10] image sets, we evaluate the visual concepts learned from the Internet images. On these image sets, the visual concepts achieve the state-of-the-art performances, demonstrating its good cross-dataset generalization capability. Also, as a type of generic knowledge from Internet images, when used with the specific models learned from specific image sets, the results can be further improved to a large extent.

4.1. Implementations

For each patch, we use HOG [4] (of 2048 dimensions), LBP [36] (of 256 dimensions) and the L*a*b* histogram (of 96 dimensions) as the feature; these features are concatenated, leading to a feature vector of dimension 2400.

The toolbox of LIBLINEAR [7] is adopted for efficient training; for each word, five iterations are used in miSVM. To create the visual concepts, on the patches labeled as positive by miSVM, 20 clusters are found using K-means; Thus, $716 \times 20 = 14320$ multi-cluster visual concepts are created for the 716 words.

We have created another two codebooks of size 14320. The first codebook is created by quantizing the densely sampled multi-scale image patches from images of all the words. The second codebook is created by finding 20 clusters from the images for each word. In the following, we name the first codebook KMS-ALL, and the second codebook KMS-SUB. As the two codebooks are created without using the saliency assumption and the multiple instance learning framework, they serve as two good baselines.

4.2. Quantitative Results

PASCAL VOC 2007 Image Set This dataset contains 20 object classes and 9963 images. It is split into training, validating and testing sets, and the mean average precision (mAP) of the 20 categories on the testing set is reported. The dataset is challenging, with large intra-class variances, cluttered backgrounds, and scale changes. When applying the visual concepts to the dataset, image patches of three scales 64×64, 128×128 and 192×192 are used.

In Table 1, we compare the mAPs. Firstly, we compare the visual concepts with the two baselines KMS-ALL and KMS-SUB. The multi-cluster visual concepts outperform both KMS-ALL and KMS-SUB, indicating that, the multi-cluster visual concepts learned are more effective. Though there are only 716 single-concept classifiers, they perform reasonably well, achieving an mAP 51%. By combining the single-concept classifiers and the multi-cluster visual concepts, the mAP is 57.5%, much higher than that of KMS-ALL and KMS-SUB.

We also compare our visual concepts with the improved Fisher-kernel (FK), locality-constrained linear coding (LLC)[32], and vector quantization (VQ). The fisher kernel starts from a Gaussian Mixture-Model (GMM), and concatenates the average first and second order differences between the patch descriptors and the centers of the GMM, leading to a feature vector of very high dimension. In [26], the Fisher-kernel is improved by reducing the dimensionality of the patch descriptors using PCA. LLC [32] projects the patch descriptors to the local linear subspaces spanned by some visual words closest to the patch descriptors, and the feature vector is obtained by max-pooling the reconstruction weights. The improved Fisher-Kernel and LLC stand for the state-of-the-arts. For FK, LLC and VQ, the results reported here are from the image classification toolbox in [3]. In [3], multi-scale dense SIFT descriptors are used as the local features and the χ^2 kernel is used in SVM when classifying the images. From Table 1, we can observe that even though we do not use images from PASCAL VOC 2007 in the learning stage, the result of our visual concepts approach is comparable to that of the states-of-the-arts.

We investigate the complementariness of our visual concepts with the model learned from the images of the PASCAL VOC 2007 image set with advanced Fisher-kernels. The kernel matrices of the visual concepts and the improved Fisher-Kernels are combined linearly. The combination weight is learned on the validating set. For the improved Fisher-kernel, its result reported in [3] is 61.69%, but when we run the toolbox with the suggested experimental settings, we get the mAP 59.6%. By combining the improved Fisher-kernel and our visual concepts, the result is boosted to 62.9%. This illustrates that our visual concepts do add extra information useful to the models learned from specific data sets.

FK	LLC-25k	VQ-25K	MCIL	KMS-SUB
59.6%	57.66%	56.07%	43.3%	53.9%
KMS-ALL	SCCs	MVC	VC	FK+VC
53.3	51%	55.6%	57.5%	62.9%

Table 1. The mean average precisions on PASCAL VOC 2007 image set. FK: the improved Fisher-kernel with 256 components; LLC-25k: LLC with $25,000$ codes; VQ-25k: vector quantization with $25,000$ codes; MCIL: multiple clustered instance learning[34]; KMS-SUB: the codebook created by clustering on each word; KMS-ALL: the codebook created by clustering on the image data from all the words; SCCs: the single-concept classifiers; MVC: the multi-cluster visual concepts; VC: the visual concepts, combination of the single-concept classifiers and multi-cluster visual concepts; FK+VC: combining the improved fisher kernel with our visual concepts.

Multiple clustered instance learning (MCIL) [34] investigates the multiple cluster property at the instance level in the MIL-Boost framework. We applied MCIL to learn a mixture of 20 cluster classifiers for each word, and used the outputs of the cluster classifiers as the features to encode the novel images. The result of MCIL is much worse than that of the visual concepts. The reason is that, in MCIL, as the number of weak classifiers increases, the number of positive instances decreases dramatically and the cluster classifiers in MCIL learn little knowledge about the image set because of the lack of positive instances. Also, there is no competition between the cluster classifiers in MCIL, making the multiple cluster property of the image data not fully investigated.

Scene Classification We evaluate the visual concepts in the task of scene classification on three scene image sets, Scene-15 [14], MIT indoor scene [27], and UIUC-Sport event [17]. Scene-15 has 15 natural scene classes; 100 images from each class are randomly selected for training and the remaining images are used for testing. UIUC-Sport has 8 complex event classes; 70 images from each class are randomly sampled for training and 60 images are sampled for testing. On both Scene-15 and UIUC-Sport, we run the experiments for 10 rounds, and report the average classification accuracy. MIT Indoor scene consists of 67 clustered indoor scene categories and has fixed training/testing splits.

On the scene image sets, image patches of two scales 64×64, 128×128 are used. The results are reported in Table 2. On the three datasets, our visual concepts approach outperforms KMS-ALL and KMS-SUB significantly. Object bank learns detection models for 200 objects from supervised data. Even though our visual concepts are learned in a weakly supervised manner, the visual concepts still outperform the detection models of object bank. The main reason for the superiority of our visual concepts is that, while object bank tries to capture an object using a single detection model, our method can capture the multiple cluster property with $14,200$ visual concepts and can model the diversity of the Internet images. We also test vector quantization (VQ) with $10,000$ codes on the three image sets using the toolbox [3]. With such a large amount of codes, VQ performs surprisingly well on the UIUC-Sport and the MIT Indoor scene sets. On all the three scene image sets, our visual concepts perform comparably to VQ though we do not use the images from those image sets. By combining the VQ with our visual concepts, the performance can be boosted significantly. Relatively, the improvement is about 3% on the Scene-15 and UIUC-Sport image sets, and 10% on the MIT indoor scene set. For VQ, with the number of codes increased, the performance will saturate: we have tested VQ with $24,000$ codes on the MIT indoor scene image set, and the accuracy is 47.1%, even a slight decrease. From Table 2, we can see that, our method also outperforms recent methods such as [33], [21],[35], [28] and [12].

	Scene-15	UIUC-Sport	MIT-Indoor
Object Bank [17]	80.9%	76.3%	37.6%
Yang *et al.* [35]	80.4%	-	-
Li *et al.* [16]	-	73.4%	-
Singh *et al.* [28]	-	-	38%
Pandey *et al.* [23]	-	-	43.1%
Quattoni *et al.* [27]	-	-	26%
Niu *et al.* [21]	78%	82.5%	-
Wang *et al.* [33]	80.43%	-	33.7%
Kwitt *et al.* [12]	82.3%	83.0%	44.0%
KMS-ALL	78.7%	81.5%	38.8%
KMS-SUB	80.4%	83.2%	41.9%
VQ	82.1%	85.6%	47.6%
VC	83.4%	84.8%	46.4%
VC+VQ	**85.4%**	**88.4%**	**52.3%**

Table 2. The classification accuracies on the scene datasets.

Inria Horse Image Set INRIA horse dataset contains 170 horse images and 170 background images taken from the Internet. We randomly selected half of the images for training and the remaining images for testing and run the experiments for 10 rounds. On this image set, the accuracy of our visual concepts is 92.47%, better than the accuracy 91.4% of VQ with $10,000$ codes and 85.3% in [20].

5. Conclusion

In this paper, we have introduced a scheme to automatically exploit mid-level representations, called visual concepts, from large-scale Internet images retrieved using word-based queries. From more than a quarter of a million images, over 14,000 visual concepts are automatically learned. These learned visual concepts are generic and have good cross-dataset generalization capability; when combined with the models learned from specific dataset, our algorithm improves the state-of-the-arts to a large extent, demonstrating the complementariness between the visual concepts and the image content in specific datasets.

Acknowledgement This project is supported by NSF CAREER award IIS- 0844566, NSF award IIS-1216528,

and NIH R01 MH094343. We thank the feedback from Andrew Fitzgibbon.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568, 2002.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[9] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, pages 1028–1035, 2011.

[10] V. Ferrari, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *CVPR*, 2007.

[11] J. Heinly, E. Dunn, and J.-M. Frahm. Comparative evaluation of binary features. In *Proc. of ECCV*, 2012.

[12] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV (4)*, pages 359–372, 2012.

[13] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR (2)*, pages 2169–2178, 2006.

[15] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[16] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, pages 1–8, 2007.

[17] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Comp. Vis.*, 60(2):91–110, 2004.

[19] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[20] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1632–1646, 2008.

[21] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR*, pages 2743–2750, 2012.

[22] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[23] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, pages 1307–1314, 2011.

[24] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, pages 1681–1688, 2011.

[25] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510, 2011.

[26] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV (4)*, pages 143–156, 2010.

[27] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.

[28] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV (2)*, pages 73–86, 2012.

[29] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV (1)*, pages 776–789, 2010.

[30] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[31] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492, 2012.

[32] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[33] L. Wang, Y. Li, J. Jia, J. Sun, D. P. Wipf, and J. M. Rehg. Learning sparse covariance patterns for natural scenes. In *CVPR*, pages 2767–2774, 2012.

[34] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *CVPR*, pages 964–971, 2012.

[35] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009.

[36] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007.

[37] J.-Y. Zhu, J. Wu, Y. Wei, E. I.-C. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *CVPR*, pages 3218–3225, 2012.