Fusing Concept Detection and Geo Context for Visual Search

Xirong Li[†*] Cees G.M. Snoek[‡] Marcel Worring[‡] Arnold W.M. Smeulders[‡§]
[†]MOE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, China
[‡]Intelligent Systems Lab Amsterdam, University of Amsterdam, the Netherlands
[§]Centrum Wiskunde & Informatica, the Netherlands
xirong.li@gmail.com,cgmsnoek@uva.nl,m.worring@uva.nl,arnold.smeulders@cwi.nl

ABSTRACT

Given the proliferation of geo-tagged images, the question of how to exploit geo tags and the underlying geo context for visual search is emerging. Based on the observation that the importance of geo context varies over concepts, we propose a concept-based image search engine which fuses visual concept detection and geo context in a concept-dependent manner. Compared to individual content-based and geo-based concept detectors and their uniform combination, concept-dependent fusion shows improvements. Moreover, since the proposed search engine is trained on social-tagged images alone without the need of human interaction, it is flexible to cope with many concepts. Search experiments on 101 popular visual concepts justify the viability of the proposed solution. In particular, for 79 out of the 101 concepts, the learned weights yield improvements over the uniform weights, with a relative gain of at least 5% in terms of average precision.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

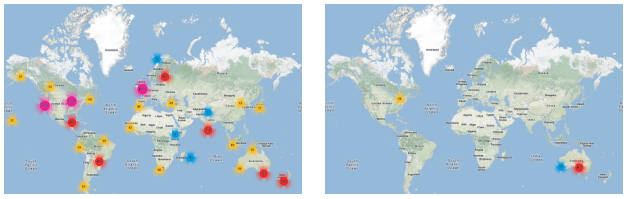
Algorithms, Measurement, Experimentation

Keywords

Visual search, concept detection, geo context

1. INTRODUCTION

Searching for unlabeled images which contain visual concepts such as animals, vehicles, and mountains is a key problem in multimedia retrieval. At the heart of a concept-based visual search engine are a number of concept detectors [5], which score unlabeled images with respect to their relevance to certain concepts. Classical solutions to visual concept detection look into the pixels only [4, 9, 21, 26], ignoring well established facts from world knowledge. As a particular instance of such knowledge, we consider in this paper the geographical position on the planet where a certain image was photographed. Consider the real-world example in Fig. 1, from which we observe that birds are spotted almost everywhere, while the occurrence of kangaroos is typically limited to Australia or a local zoo. Knowing where a picture was taken may reduce the uncertainty in interpreting its content, and thus improve visual concept search.



(a) Concept 'bird' (b) Concept 'kangaroo'

Figure 1: The geographical distribution of visual concepts on planet earth, estimated from one million geo-tagged Flickr images. Large circles indicate dense populations. Note that the distribution varies over concepts, motivating us to exploit geo context for visual concept search.

Thanks to the proliferation of smart phones, GPS enabled cameras and online geo tagging services, the geographical location recorded in the form of latitude and longitude geo tags, are becoming basic metadata for newly generated images. Naturally, this metadata has been exploited to infer geo context, e.g., park, playground, and beach, of the place where an image was taken [3, 6, 10, 13, 17, 20, 22, 27, 28]. The research question arises as *how to include geo context in a visual search engine*?

A number of papers have appeared to exploit geo context for visual content analysis [3, 7, 10, 12, 13, 17, 20, 22, 23, 27, 28]. Most of them target at either image summarization [7, 12] or image annotation [3, 13, 20, 22, 23, 28]. A typical approach as introduced by Moxley et al. [20] and Kleban et al. [13] is to annotate a given image by constrained k nearest neighbor (k-NN) voting, where the visual neighbors are retrieved from the geo region of the given image. Although image

[*]Work performed while the author was with UvA.

ICMR '12, June 5-8, Hong Kong, China

annotation models can be used for concept search, the two tasks have distinct goals. Image annotation strives to rank concepts for a given image, while concept search aims to rank images for a given concept.

Joshi and Luo [10] were among the first to leverage geo context for concept search. They introduce a location related representation for a geo-tagged image by an inverse geo encoding using a geographical information system. Consequently, they build a geo-based detector, and uniformly fuse it with a content-based detector. Recent work by Yaegashi and Yanai [27] studies an early fusion scheme, where they feed both visual features and location related features into a multiple kernel learning framework. As early fusion works at the feature level, it has difficulty in incorporating variants of geo-based detectors which have been introduced [10, 13, 20] or will be introduced in the future. Moreover, for concept-dependent fusion, training data is required for each concept. The use of hand crafted training examples in the system [27] puts its scalability into question.

In this paper, we start from the observation that the strength of content-based and geo-based detectors varies over concepts. For instance, the geo context can be good evidence when looking for beach or zoo scenes, as images taken at these areas tend to have less varied subjects. However, for detecting sky or dog, we have to favor content over geo. We argue that fusing concept detection and geo context in a concept-dependent manner is crucial. We propose a concept-based image search engine working in this manner, as illustrated in Fig. 2. By concept-dependent fusion, the proposed search engine favorably employs content-based and geo-based meta detectors against their uniform combination. We compare three present-day geo-based detectors for visual concept search, which has not been done before. Since all the meta and fused detectors are trained on social-tagged images with no need of extra manual labeling, the proposed search engine can scale up to a large array of concepts.

The rest of this paper is organized as follows. We review related work in Section 2. We describe the proposed search engine in Section 3. Visual concept search experiments are setup in Section 4, followed by result analysis in Section 5. Our conclusions are given in Section 6.

2. RELATED WORK

Visual Concept Detection. Since the seminal paper by Csurka et al. [4], bag of visualwords features plus Support Vector Machines classification are now established as a solid choice for visual concept detection [9, 25, 26]. In order to release many detectors from the limited availability of well labeled training examples, methods aiming to learn detectors from social-tagged examples have been studied [2, 11, 15, 29]. Given the rapid growth of new images, computational efficiency of concept detection is becoming a practical concern for a search engine. Recently, Uijlings et al. [25] propose a real-time algorithm for bag of visualwords feature extraction. To accelerate SVM prediction, Maji et al. [18] introduce fast intersection kernel. We will leverage this good progress for visual concept detection in our system.

Geo Clues for Concept Detection. To describe the geo context of a given image, multiple geo clues have been investigated. For instance, Moxley et al. [20], Quack et al. [22], and Kleban et al. [13] consider consumer photos taken in the geographical region of the given image. Luo et al. [17] manually collect satellite aerial images of the region. Instead of using local images, Joshi and Luo [10] query a GIS database GeoNames[1] to find place entities nearby, and use the associated text, e.g., "Panorama Hotel, a building providing lodging and/or meals for the public", to represent the geo context. As an alterative to GIS, Yaegashi and Yanai [27] employ the Yahoo! Local Search API for constructing geo descriptions. In this paper we choose consumer photos and GIS text as two types of geo clues for their relatively easy accessibility.

Combining Visual and Geo Clues. For combining multiple clues, there are two essentially different approaches, namely early and late fusion, which work on the feature level and the detector level, respectively. Yaegashi and Yanai [27] investigate an early fusion scheme by combining visual and location related feature in the form of multiple kernel learning. As aforementioned, early fusion is not flexible to combine heterogenous meta detectors. Joshi and Luo [10] and Luo et al. [17] train meta detectors separately and combine them in a late fusion scheme with equal weights. We use late fusion as well, but with a noticeable difference from [10, 17] that we employ learning algorithms to let the system automatically figure out the importance of the underlying meta detectors with respect to specific concepts.

3. THE PROPOSED SYSTEM

Our goal is to build a concept-based search engine which jointly exploit visual content and geo context. When using late fusion, the exploitation of the two sources of evidence will naturally yield multiple meta detectors for a specific concept. The meta detectors are on the base of content, geo, or both. For the effective use of these detectors, well performing detectors should be stressed, while poor performing detectors have to suppressed or ignored, depending on the concept at hand.

To make our discussion more formal, we use x to denote an image. As a simplification, we also use x to indicate an m-dimensional visual feature of the corresponding image, where $x(l)$ indicates the value of the l-th dimension. Given two images x and x', we denote their visual distance by $d_{vis}(x, x')$. On the use of geo context, the geographical distance between two images is important, which is denoted by $d_{geo}(x, x')$. For a given visual concept ω, let $H(x, \omega)$ be its detector in general. Shall the concept have r meta detectors, we denote them by $\{H_i(x, \omega)|i = 1, \ldots, r\}$. To evaluate the effectiveness of a specific detector, we introduce an abstract performance metric function $E_{metric}(H(x, \omega))$. For the popular metric Average Precision, the function will be $E_{ap}(H(x, \omega))$.

3.1 Concept-Dependent Fusion of Meta Detectors

Given a set of predefined concepts, we first train the meta detectors for each concept. Towards handling many concepts, we will train all detectors on social-tagged images only [14, 15]. Based on our hypothesis that the relative strength of content-based and geo-based detectors varies over concepts, we consider concept-dependent fusion. Besides their varying performance, the computational efficiency also differs between the individual detectors. For instance, visual detectors naturally conducts content analysis [9, 26], which is bypassed by some geo-based detectors [10, 20]. Fusion is

[1] http://www.geonames.org/

thus to strike a balance between effectiveness and efficiency. We first consider an efficiency oriented strategy which aims to select the "right" detector, rather than use all detectors.

Strategy I: Learn to Switch. This strategy makes the system switch meta detectors in terms of the concept in consideration. For a given concept ω, we switch to the detector $H^*(x, \omega)$,

$$H^*(x, \omega) = \underset{i=1,\ldots,r}{\operatorname{argmax}} E_{metric}(H_i(x, \omega)), \quad (1)$$

which has the best performance on a given validation set.

We now consider the full use of the meta detectors.

Strategy II: Learn to Fuse. We adopt linear fusion due to its widespread use [1], and define the fused detector as

$$H_\Lambda(x, \omega) = \sum_{i=1}^{r} \lambda_i \cdot H_i(x, \omega), \quad (2)$$

where $\Lambda = \{\lambda_i\}$ are the weighting parameters. When the performance of the detectors is unknown, the uniform weights,

$$\Lambda_0 = \frac{1}{r}\mathbf{1}, \quad (3)$$

is a reasonable choice, as has been used in previous work [10, 17]. However, as discussed in Section 1, there are a considerable amount of geo-tagged images online, with their content described by social tags. Though social tags are often noisy [11], techniques have been developed to determine tag relevance to the content [2, 14, 29]. The availability of geo-tagged images with de-noised annotations make supervised parameter learning possible. Thus, we seek Λ which maximizes $E_{metric}(H_\Lambda(x, \omega))$, while at the same time we want to keep it close to the uniform weight to reduce the risk of overfitting. In that regard, we formulate the optimization problem for Λ:

$$\underset{\Lambda}{\operatorname{argmax}} E_{metric}(\sum_{i=1}^{r} \lambda_i \cdot H_i(x, \omega)) - \xi \cdot ||\Lambda - \Lambda_0||, \quad (4)$$

where ξ is a regularization parameter.

Because common metric functions such as $E_{ap}(H(x, \omega))$ are non-differentiable, the objective function (4) cannot be solved by standard gradient ascent algorithms. We seek an optimization technique which maximizes (4) without computing the gradient. To that end, we leverage the coordinate ascent algorithm, introduced by Metzler and Croft [19] for combing multiple sources of ranking evidence in document retrieval. This algorithm iteratively solves (4) by optimizing merely one parameter in Λ per time, with the remaining parameters fixed. Suppose λ is the underling parameter being optimized, we conduct a bi-direction line search with increasing steps to find the optimal value λ^*. If the search succeeds, i.e., Λ^* yields a larger response on the objective function, we update λ with λ^*. Then, the next parameter is activated and the same procedure applies. As only the relative weights of the parameters are important, to reduce the search space, we normalize the parameters by dividing them by the sum of their absolute values. The procedure continues until the objective function stops increasing.

3.2 Meta Detectors

We describe our choices of content-based detectors [15,18] and geo-based detectors [10,13,20], which reflect present-day techniques of their kinds.

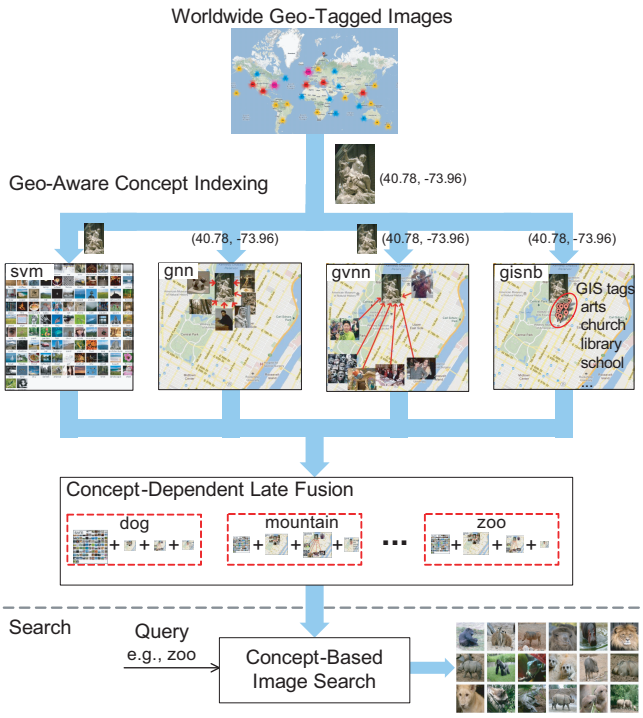


Figure 2: The proposed concept-based retrieval system for geo-tagged images. In the fusion component, the size of the four meta detectors indicates their varying contributions for detecting specific concepts.

Content-Based Detector: SVM. We choose bag of visualwords features plus SVM modeling, which is a solid choice for concept detection [4, 9, 26]. Concerning training data for a given concept ω, a straightforward solution is to treat images labeled with ω by social tagging as positive examples, and those not labeled with ω as negative examples. However, as mentioned above, social tags may be irrelevant to the visual content. Moreover, the positive and negative examples are imbalanced by the nature of social image data. These two difficulties make learning SVM models on social-tagged examples directly problematic.

To improve the quality of social-tagged positive examples, we employ the neighbor voting based algorithm by Li et al. [14]. Given a concept ω and images labeled with ω by social tagging, we apply the algorithm to each image to obtain a score which reflects its positiveness. We sort the images in descending order by their scores and preserve the top N ranked results as positive training examples. To cope with the class imbalance problem, we follow the sampling approach by Li et al. [15]. Different from sampling negative examples at random, this approach adaptively and iteratively selects negative examples which are most misclassified by present classifiers, and thus most informative to improve classification. The combination of neighbor voting [14] and adaptive sampling [15] enables us to obtain better SVM models, compared to models directly trained on social-tagged images.

For large-scale concept detection, efficiency is also important. Maji et al. [18] show that the histogram intersection kernel is as effective as the χ^2 kernel yet much more efficient. Hence, we adopt their algorithm, and express the SVM de-

tector as

$$H_{svm}(x,\omega) = b_\omega + \sum_{j=1}^{n_\omega} \alpha_{j,\omega} \cdot \sum_{l=1}^{m} \min(x(l), x_j(l)), \quad (5)$$

where x_j represents a support vector with $\alpha_{j,\omega}$ as its coefficient, b_ω the intercept, and n_ω the number of support vectors. The additive property of the histogram intersection kernel allows us to exchange the sum operators in (5), and consequently compute the decision function for each dimension as

$$H_{svm,l}(x,\omega) = \sum_{j=1}^{n_\omega} \alpha_{j,\omega} \cdot \min(x(l), x_j(l)). \quad (6)$$

Since (6) can be well approximated by linear interpolation on a fixed number of pre-computed points [18], the computation of (5) becomes independent of the number of support vectors. This trick results in efficient detection, approximately 25 milliseconds for sequentially applying 101 detectors per image on a single computer.

In the consideration of forming complementary detectors for fusion, we choose the following three geo-based detectors, exploiting localized tags [20], localized content [13], and GIS knowledge [10].

Geo-Based Detector-I: Geo KNN. Moxley et al. [20] describe a geographical k-NN detector which counts tag frequency on images which are closest to a novel image in terms of $d_{geo}(x,x')$. To suppress the impact of distant images, we define the geographical similarity as

$$\mathcal{K}_{geo}(x,x') = \exp(-\frac{d_{geo}(x,x')}{h_{geo}}), \quad (7)$$

where h_{geo} controls the rate of diffusion. We express the geo knn detector as

$$H_{gnn}(x,\omega) = \frac{\sum_{i=1}^{k} \mathbf{I}(x_i,\omega) \cdot \mathcal{K}_{geo}(x,x_i)}{\sum_{i=1}^{k} \mathcal{K}_{geo}(x,x_i)}, \quad (8)$$

where x_i is the i-th neighbor in terms of d_{geo}, and the indicator $\mathbf{I}(x_i,\omega)$ is 1 if x_i is labeled with ω, 0 otherwise. Notice that in the context of image annotation, the importance of individual tags is considered in [20]. We omit this factor, as it does not affect concept search.

Geo-Based Detector-II: Geo-Constrained Visual KNN. Given a novel image and a geo region $\mathbf{g}$ it belongs to, the detector by Kleban et al. [13] retrieves visually similar pictures which were taken within $\mathbf{g}$. In a similar manner to $\mathcal{K}_{geo}(x,x')$ in (7), we define the visual similarity $\mathcal{K}_{vis}(x,x')$. The geo-constrained visual knn detector is computed as

$$H_{gvnn}(x,\omega) = \frac{\sum_{i=1}^{k} \mathbf{I}(x_i,\omega) \cdot \mathcal{K}_{vis}(x,x_i)}{\sum_{i=1}^{k} \mathcal{K}_{vis}(x,x_i)}, \quad (9)$$

where x_i is the i-th neighbor in terms of d_{vis}. In a similar vein to [13], we find $\mathbf{g}$ by indexing the training data in terms of their latitudes and longitudes using a quadtree. The quadtree partitions the two dimensional geo space by recursively subdividing it into four equal-sized quadrants. Consequently, the region $\mathbf{g}$ is indexed by the leaf node to which the novel image belongs.

Geo-Based Detector-III: GIS Naive Bayes. The last geo-based detector investigated in this paper is by Joshi and Luo [10], a Naive Bayes classifier built upon GIS tags. To construct the GIS tags for a geo-tagged image, we follow [10] to first find 20 entities from the GeoNames which are geographically closest to the image. Each entity in the GeoNames is associated with manually edited text, briefly describing the corresponding geo location. As suggested by [10], we group the 20 entities into two clusters by K-means clustering, and preserve the largest cluster. The descriptions of the entities in the selected cluster are merged to form the tag set for the given image. For a specific GIS tag t, we use $p(t|\omega_+)$ to represent the probability of observing t in the positive training data of a concept ω, and $p(t|\omega_-)$ for the probability of observing t in the negative training data. For a novel image and its automatically assigned GIS tags $\{t_1,\ldots,t_n\}$, the Naive Bayes detector is defined as

$$H_{gisnb}(x,\omega) = \sum_{i=1}^{n} \log\left(p(t_i|\omega_+) - p(t_i|\omega_-)\right). \quad (10)$$

In each model, there are over 18,000 GIS tags and many of them are associated with small values. For reasons of numerical stability, we use the log odds ratio in place of the odds ratio used in [10].

Table 1: The content-based and geo-based concept detectors studied in the paper.

Meta detectors	Descriptions
$H_{svm}(x,\omega)$	Bag of visualwords + SVM (5) [15, 18]
$H_{gnn}(x,\omega)$	Geographical knn (8) [20]
$H_{gvnn}(x,\omega)$	Geo-constrained visual knn (9) [13]
$H_{gisnb}(x,\omega)$	GIS tags + Naive Bayes (10) [10]
Fused Detectors	
$H^*(x,\omega)$	The best meta detector (1)
$H_\Lambda(x,\omega)$	Linear fusion of the meta detectors (2)

We summarize the four meta detectors in Table 1. While the output of $H_{gnn}(x,\omega)$ and $H_{gvnn}(x,\omega)$ is within the interval [0,1], the output of $H_{svm}(x,\omega)$ and $H_{gisnb}(x,\omega)$ can be negative or positive. For the convenience of fusion, we use the following sigmoid function to convert $H_{svm}(x,\omega)$ and $H_{gisnb}(x,\omega)$ to a form of probabilistic output:

$$\frac{1}{1+\exp(A \cdot H(x,\omega) + B)}, \quad (11)$$

where A and B are two real-valued parameters optimized on the training data by regularized maximum likelihood estimation [16].

The entire system is illustrated in Fig. 2. Given a query concept, the system uses the corresponding fused detector to sort images and returns the top ranked results.

4. EXPERIMENTAL SETUP

4.1 Data Collections

To gather geo-tagged images, we use over 25,000 WordNet tags which have correspondences to visual concepts as queries to uniformly sample Flickr images uploaded between 2005 and 2010. After removing batch-tagged images and images whose geo-tagging accuracy (as defined by Flickr) is lower than 10, we obtained 5 million geo-tagged images.

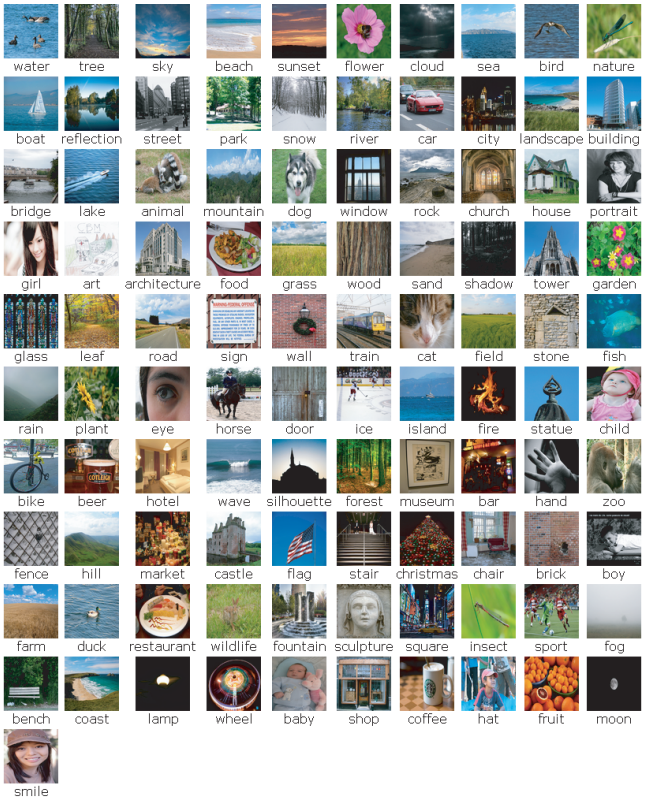


Figure 3: Examples of the 101 visual concepts used in our experiments. The concepts are sorted in descending order by their frequency. Reading order: left to right, top to bottom.

We use this set as our data source to define concepts and datasets for training, validation, and testing.

101 Concepts. We detect visual concepts which are of common users' interest. To construct such a concept set, we sort tags in the 5M set in descending order by the number of distinct users who have used them for image tagging. We manually went trough the top ranked results, and filter out tags which are not depicting objects, scenes, or events. Visual examples of the concepts are given in Fig. 3.

1M Training Data. We randomly sample 1M images from the 5M set as our training data. The data set consists of images taken in 109 countries by 145,029 users. We use the full set for $H_{gnn}(x,\omega)$ (8) and $H_{gvnn}(x,\omega)$ (9). For the feasibility of computation, we train $H_{svm}(x,\omega)$ (5) and $H_{gisnb}(x,\omega)$ (10) on subsets of the training data.

10K Test Data. To the best of our knowledge, there is no geo-tagged ground truth available for the 101 concepts. Hence, we create a test set by sampling the 5M set at random, with user tags as an approximation of genuine labels. For the purpose of de-noising, we remove over-tagged images which have more than 10 tags. We exclude an image, if for all of its tags, their tag relevance scores [14] are smaller than given thresholds. Moreover, to assure independence between the training and testing data, we exclude an image if its user already appears in the training data. After the above preprocess, we take a random subset of 10K images as the test data.

10K Validation Data. In order to train $H^*(x,\omega)$ and $H_\Lambda(x,\omega)$, we select another random subset of 10K images from the 5M set, with the same preprocess applied. The statistics of our experimental data are given in Table 2.

Table 2: Statistics of our experimental data. Note that users (and images) from each set do not overlap users (and images) from the other two sets.

	Training	Validation	Testing
#images	1M	10K	10K
#users	145,029	8,068	7,411
#countries	109	86	86

4.2 Experiments

To evaluate the proposed visual search engine in terms of its ingredients, the fusion strategies, and its efficiency, we conduct the following three experiments.

Experiment 1. Comparing meta detectors. We compare the four meta detectors, denoted by svm, gnn, gvnn, and gisnb respectively.

Experiment 2. Detector Fusion. We compare the two supervised fusion schemes and fusion with uniform weights, abbreviated as learn2switch, learn2fusion, and uniform-fusion, respectively.

Experiment 3. Efficiency Analysis. We evaluate the efficiency of the system by measuring its time cost for indexing novel images with the 101 concepts.

4.3 Implementation

Meta detectors. To train $H_{svm}(x,\omega)$ for a given concept ω, we rank images labeled with ω in the 1M training data by the multi-feature tag relevance learning algorithm [14], and preserve the top 300 results as the positive training set. We then run the negative bootstrapping algorithm [15] with 10 iterations, resulting in 10 models. We average the models to obtain $H_{svm}(x,\omega)$. We extract a 1,024-dimensional bag of visualwords feature, by quantizing densely sampled SIFT descriptors [26]. For $H_{gnn}(x,\omega)$, we empirically set k and h_{geo} in to be 20 and 2 kilometers. The k in $H_{gvnn}(x,\omega)$ is also set to 20. To train $H_{gisnb}(x,\omega)$, we use the same positive data as for $H_{svm}(x,\omega)$, while randomly sample 3,000 examples from the 1M set as the negative training data.

Detector Fusion. We empirically set the regularization parameter ξ in (2) to be 0.01. We solve (1) and (4) by optimizing average precision.

Evaluation Criteria. We use Average Precision (AP), which is in wide use for evaluating visual search engines [24]. We also report Normalized Discounted Cumulative Gain (NDCG), commonly used to assess the top few ranked results of web search engines [8].

5. RESULTS

5.1 Experiment 1. Comparing meta detectors

Comparing Geo-Based Detectors. As shown in Fig. 4, for the top ranked result, the gvnn method with NDCG@1 of 0.248 performs best among all the three geo-based detectors, followed by gisnb with NDCG@1 of 0.089 and gnn with NDCG@1 of 0.059. When taking the entire ranking into account as shown in Fig. 5, gnn with mean average precision of

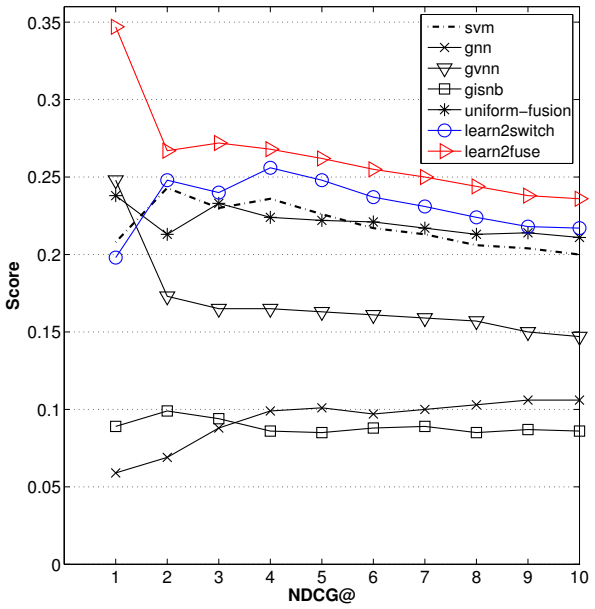


Figure 4: Comparing the top ranked results obtained by different detectors. While the SVM performs best among the four best detectors, learn2fuse is the best strategy for detector fusion.

0.054 beats gvnn and gisnn which have mean average precision of 0.049 and 0.028, respectively. As the gvnn combines both geo and content information, it is more discriminative, favoring precision over recall. With the results, we conclude that for visual concept search, geo context inferred from local photos is more helpful than its counterpart derived from the GeoNames.

Content-Based versus Geo-Based Detectors. As shown in Fig. 4 and Fig. 5, the SVM detector surpasses the geo-based detectors in general, with NDCG@10 of 0.200 and mean average precision of 0.060. This is due to the fact that several concepts such as car, dog, and portrait, may appear almost everywhere. Consequently, adding the geo information is not helpful for detecting these concepts. However, for some concepts such as zoo, beach, island, mountain, and museum, photos taken in the corresponding areas tend to have less varied subjects. As a consequence, we observe clear advantages of the geo-based detectors against the SVM. For the same reason, gvnn outperforms svm in terms of NDCG@1. These results show that for the effective use of geo context, concept-dependent fusion is necessary.

5.2 Experiment 2. Detector Fusion

learn2fuse versus uniform-fusion. As shown in Fig. 4 and Fig. 5, learn2fuse, with NDCG@10 of 0.236 and mean average precision of 0.085, is better than uniform-fusion, with NDCG@10 of 0.211 and mean average precision of 0.064. For 79 out of the 101 concepts, the learned weights yield improvements over the uniform weights, with a relative gain of at least 5% in terms of average precision. To gain insight into the contributions of the individual meta detectors, we visualize their weight distribution in Fig. 6. The SVM detector contributes most due to its good performance, followed by gnn, gvnn, and gisnb. The large degree of dispersion in the boxplot shows that the optimal weights vary over concepts, suggesting the necessity of concept-dependent fusion for deriving geo-aware concept detectors.

learn2fuse versus learn2switch. As shown in Fig. 4, for the first hit, learn2fuse with NDCG@1 of 0.347 is much better than learn2switch with NDCG@1 of 0.198. Notice that learn2switch is also worse than uniform-fusion in terms of NDCG@1. However, when measuring in terms of mean average precision as given in Fig. 5, learn2switch with a score of 0.077 beats uniform-fusion. We attribute this result to the reason that the winner-take-all policy makes the learn2switch strategy less robust than uniform-fusion and learn2fuse. Compared to the best meta detector (svm), learn2fuse successfully ranks positive results at the top for 18 concepts where svm fails, and incorrectly ranks negative results at the top for 4 concepts where svm succeeds. For the second rank, due to the performance drop of gvnn, learn2fuse incorrectly ranks negative results for 14 concepts, while the number of successful corrections is 5. As a consequence, we observe a larger performance gain at NDCG@1 than at NDCG@2. In sum, the above results allow us to conclude that for detector fusion, learn2fuse is the best strategy.

5.3 Experiment 3. Efficiency Analysis

To make the system more compact, we preserve the svm and the gnn detectors only, as they contribute most to detection fusion. This simplification results in a performance loss of 0.003 in terms of mean average precision, but makes the system more efficient. Given a novel image with a size of 500×500, extracting the visual feature by the real-time bag-of-words algorithm [25] takes about 35 milliseconds. Sequentially applying the 101 svm detectors costs 25 milliseconds, and 2 milliseconds for executing the gnn detector on a single computer with 2.4 GHz multi-core cpu and 24 GB memory. Scoring an image takes 62 milliseconds in total, meaning the proposed system can index 16 images per second, approximately.

6. CONCLUSIONS

In this paper we study how to fuse generic visual concept detection and geo context for improving visual search. To that end, we propose a concept-based image search engine which fuses a content-based and three geo-based meta detectors in a concept-dependent manner. We conduct search experiments for 101 visual concepts popular in social image tagging. Our major findings are as follows. Comparing the three geo-based detectors, we find that for visual concept search, geo context inferred from local images is more helpful than geo context derived from a geographical information system. Comparing the visual and geo-based detectors, while the visual detectors perform best in general, the geo-based detectors have a noticeable advantage on concepts associated with strong geo characteristics. Hence, the exploration of visual and geo clues has to be concept-dependent. Concerning the fusion strategies, the concept-dependent learn2fuse surpasses fusion with uniform weights. Moreover, compared to search by visual detectors alone, learn2fuse boosts the retrieval performance, increasing mean average precision from 0.060 to 0.085. Since all the meta and fused detectors are learned from social-tagged images without the need of extra manual labeling, the proposed system can scale up to a large array of concepts. These results verify the viability of the proposed solution for fusing

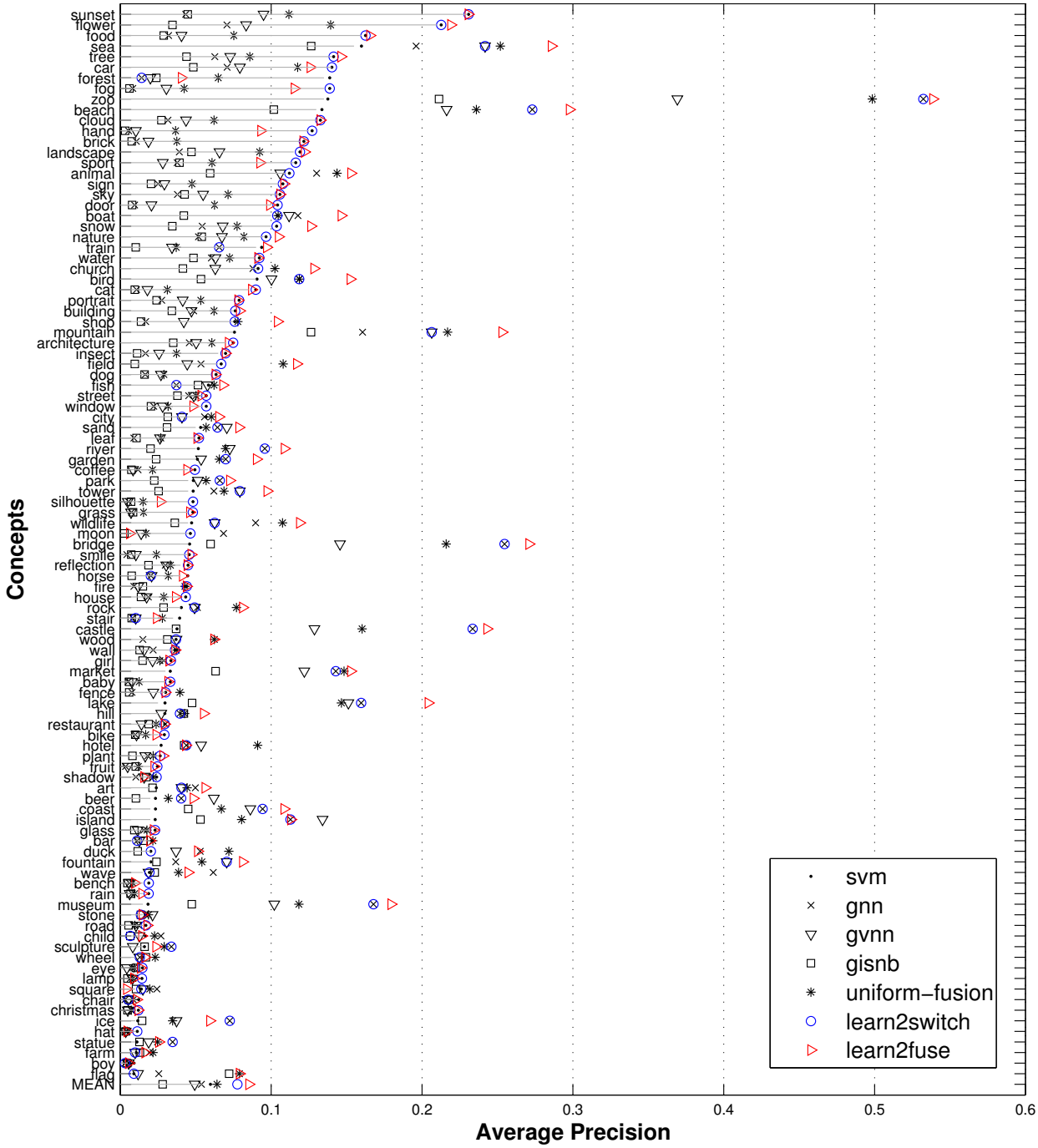


Figure 5: A concept-by-concept comparison between the meta and fused detectors for visual search. The concepts are sorted in descending order by the performance of the SVM detector.

concept detection and geo context.

With a compact configuration of fusing svm and gnn only, the system can index 16 images per second, approximately, meaning indexing one million geo-tagged images within a day. The efficiency shows the potential of the proposed solution for highly demanding applications such as live analytics on worldwide geo-tagged images.

Acknowledgements

This work was partly supported by the STW SEARCHER project, the Dutch national program COMMIT, and the HGJ program (2010ZX01042-002-002-03). The first author thanks Prof. H.T. Shen from the University of Queensland for helpful discussion on the topic.

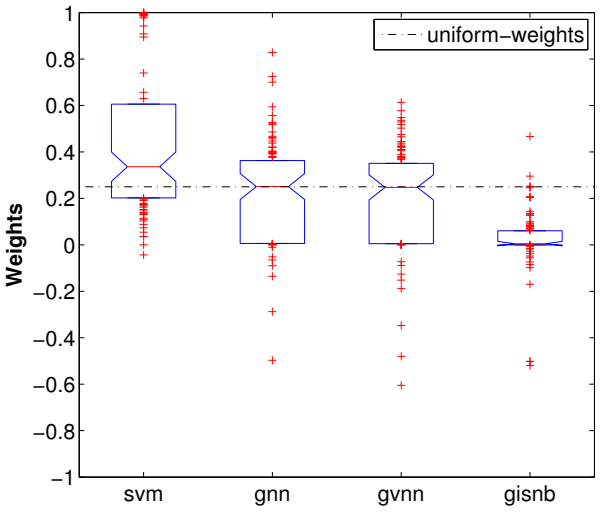


Figure 6: The weight distribution of the four detectors on the 101 concepts, obtained by solving (4). The SVM detector contributes most, followed by gnn, gvnn, and gisnb. The large degree of dispersion indicates the necessity of concept-dependent fusion.

7. REFERENCES

[1] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.*, 16(6):345–379, 2010.

[2] D. Borth, A. Ulges, and T. Breuel. Relevance filtering meets active learning: improving web-based concept detectors. In *MIR*, 2010.

[3] L. Cao, J. Yu, J. Luo, and T. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *ACM MM*, 2009.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[5] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Trans. MM*, 9(5):958–966, 2007.

[6] J. Hays and A. Efros. IM2GPS: Estimating geographic information from a single image. In *CVPR*, 2008.

[7] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *MIR*, 2006.

[8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20:422–446, 2002.

[9] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. MM*, 12:42–53, 2010.

[10] D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *CIVR*, 2008.

[11] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *MIR*, 2006.

[12] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM MM*, 2007.

[13] J. Kleban, E. Moxley, J. Xu, and B. Manjunath. Global annotation on georeferenced phototgraphs. In *CIVR*, 2009.

[14] X. Li, C. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *CIVR*, 2010.

[15] X. Li, C. Snoek, M. Worring, and A. Smeulders. Social negative bootstrapping for visual categorization. In *ICMR*, 2011.

[16] H.-T. Lin, C.-J. Lin, and R. Weng. A note on Platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68:267–276, 2007.

[17] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM MM*, 2008.

[18] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[19] D. Metzler and B. Croft. Linear feature-based models for information retrieval. *Inf. Retr.*, 10(3):257–274, 2007.

[20] E. Moxley, J. Kleban, and B. Manjunath. SpiritTagger: A geo-aware tag suggestion tool mined from Flickr. In *MIR*, 2008.

[21] M. Naphade, C.-Y. Lin, A. Natsev, B. Tseng, and J. Smith. A framework for moderate vocabulary semantic visual concept detection. In *ICME*, 2003.

[22] T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR*, 2008.

[23] A. Silva and B. Martins. Tag recommendation for georeferenced photos. In *LBSN*, 2011.

[24] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *MIR*, 2006.

[25] J. Uijlings, A. Smeulders, and R. Scha. Real-time visual concept classification. *IEEE Trans. MM*, 12(7):665–681, 2010.

[26] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1582–1596, 2010.

[27] K. Yaegashi and K. Yanai. Geotagged image recognition by combining three different kinds of geolocation features. In *ACCV*, 2010.

[28] J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. In *CIVR*, 2008.

[29] S. Zhu, G. Wang, C.-W. Ngo, and Y.-G. Jiang. On the sampling of web images for learning visual concept classifiers. In *CIVR*, 2010.