Harvesting Social Images for Bi-Concept Search

Xirong Li, Cees G.M. Snoek, *Senior Member, IEEE*, Marcel Worring, *Member, IEEE*, Arnold W.M. Smeulders, *Member, IEEE*

Abstract

Searching for the co-occurrence of two visual concepts in unlabeled images is an important step towards answering complex user queries. Traditional visual search methods use combinations of the confidence scores of individual concept detectors to tackle such queries. In this paper we introduce the notion of bi-concepts, a new concept-based retrieval method that is directly learned from social-tagged images. As the number of potential bi-concepts is gigantic, manually collecting training examples is infeasible. Instead, we propose a multimedia framework to collect de-noised positive as well as informative negative training examples from the social web, to learn bi-concept detectors from these examples, and to apply them in a search engine for retrieving bi-concepts in unlabeled images. We study the behavior of our bi-concept search engine using 1.2M social-tagged images as a data source. Our experiments indicate that harvesting examples for bi-concepts differs from traditional single-concept methods, yet the examples can be collected with high accuracy using a multi-modal approach. We find that directly learning bi-concepts is better than oracle linear fusion of single-concept detectors, with a relative improvement of 100%. This study reveals the potential of learning high-order semantics from social images, for free, suggesting promising new lines of research.

Index Terms

Semantic index, bi-concept, visual search

This publication was supported by the Dutch national program COMMIT and the STW SEARCHER project.

X. Li is with the MOE key laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China (email: xirong.li@gmail.com).

C.G.M. Snoek, M. Worring, and A.W.M. Smeulders are with the Intelligent Systems Lab Amsterdam, University of Amsterdam, 1098 XH Amsterdam, The Netherlands.

I. INTRODUCTION

Searching pictures on smart phones, PCs, and the Internet for specific visual concepts, such as objects and scenes, is of great importance for users with all sorts of information needs. As the number of images is growing so rapidly, full manual annotation is unfeasible. Therefore, automatically determining the occurrence of visual concepts in the visual content is crucial. Compared to low-level visual features such as color and local descriptors used in traditional content-based image retrieval, the concepts provide direct access to the semantics of the visual content. Thanks to continuous progress in generic visual concept detection [1]–[4], followed by novel exploitation of the individual detection results [5]–[8], an effective approach to unlabeled image search is dawning.

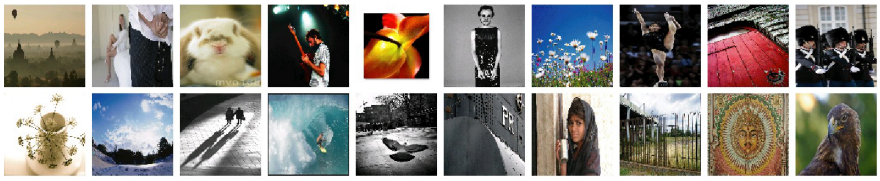
In reality, however, a user's query is often more complex than a single concept can represent [9]. For instance consider the query: "an image showing a horse next to a car". To answer this query, one might expect to employ a 'car' detector and a 'horse' detector, and combine their predictions, which is indeed the mainstream approach in the literature [6]–[8], [10]–[12]. But is this approach effective? We observe that the single concept detectors are trained on typical examples of the corresponding concept, e.g., cars on a street for the 'car' detector', and horses on grass for the 'horse' detector. We hypothesize that images with horses and cars co-occurring also have a characteristic visual appearance, while the individual concepts might not be present in their common form. Hence, combining two reasonably accurate single concept detectors is mostly ineffective for finding images with both concepts visible, as illustrated in Fig. 1.

Ideally, we treat the combination of the concepts as a new concept, which we term *bi-concept*. To be precise, we define a bi-concept as the co-occurrence of two distinct visual concepts, where its full meaning cannot be inferred from one of its component concepts alone. According to this definition, not all combinations of two concepts are bi-concepts. For instance, a combination of a concept and its superclass such as 'horse + animal' is not a bi-concept, because 'horse + animal' bears no more information than 'horse'. Besides, specialized single concepts consisting of multiple tags such as 'white horse' [13], [14] and 'car driver' are not bi-concepts as the two tags refer to the same visual concept. The same holds for events such as "airplane landing" where the tag landing is not a distinct visual concept by itself.

Although not all semantic combinations are bi-concepts, the number of possible bi-concepts is still quadratic in the number of single concepts. Even when we assume that a set of only 5,000 concepts is enough for general purpose search [6], finding sufficient labeled examples for each possible bi-concept already becomes a problem of big proportion. The amount of labeling effort is so considerable that it puts the scalability of expert annotation and the recent Amazon Mechanical Turk service into question. We

(a) Searching for 'car' by a 'car' detector

(b) Searching for 'horse' by a 'horse' detector



(c) Searching for 'car + horse' by combining the results of (a) and (b)

(d) Searching for 'car + horse' using the proposed bi-concept search engine

Fig. 1. **Searching for two visual concepts co-occurring in unlabeled images**. A (green) tick indicates a positive result. Given two single concept detectors with reasonable accuracy, a combination using their individual confidence scores yields a bad retrieval result (c). We propose to answer the complex query using a bi-concept detector optimized in terms of mutual training examples (d).

consider obtaining bi-concept examples without expert labeling as a key problem for bi-concept search in unlabeled images.

A novel source of labeled images for concept detection are user-tagged images on the social web such as those on Flickr and Facebook. However, due to the subjectiveness of social tagging, social tags often do not reflect the actual content of the image. It has been shown in previous studies that directly training on social-tagged images results in suboptimal single concepts detectors [15]–[17]. Learning bi-concept detectors from social-tagged images is, to the best of our knowledge, non-existing in the literature. By definition, the number of images labeled with a bi-concept is less than the number of images labeled with

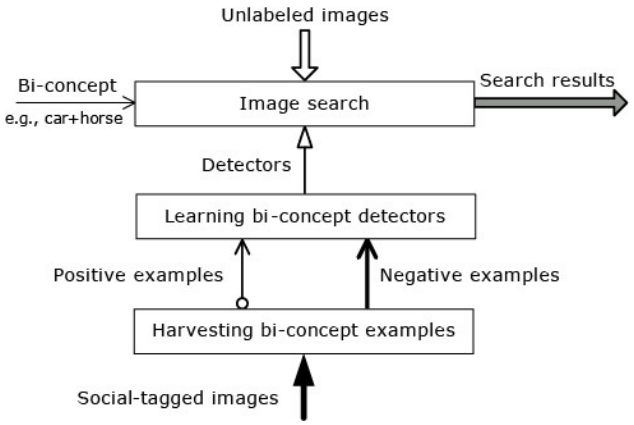


Fig. 2. **A conceptual diagram of the proposed bi-concept image search engine**.

a single concept, meaning bi-concepts have a worse starting point for obtaining examples. In addition, a scene with two concepts present tends to be visually more complex, requiring multi-modal analysis. Given these difficulties, effective bi-concept search demands an approach to harvesting appropriate examples from social-tagged images for learning bi-concept detectors.

In this paper, we introduce the notion of bi-concepts, and propose a multimedia search engine to study their instantiation, development, and applicability. We present a multi-modal approach to collect de-noised positive as well as informative negative training examples from the social web. We learn bi-concept detectors from these examples, and later apply them for retrieving bi-concepts in unlabeled images. A conceptual diagram of the proposed search engine is illustrated in Fig. 2.

II. Related work

We first review related work on combining single concept detectors for visual search, and then discuss recent progress on harvesting training examples from the (social) web. For the ease of consistent description, we use x to represent an image, w for a single concept, and $\mathbf{w}_{1,2}$ for a bi-concept comprised of two single concepts w_1 and w_2. We use $p(w|x)$ to represent a concept detector which produces a posterior probability of observing w given the image. In a similar fashion, we define a bi-concept detector $p(\mathbf{w}_{1,2}|x)$.

A. Visual Search by Combining Single Concepts

Given hundreds of single concept detectors trained on well-labeled examples [18], a considerable amount of papers have been published on how to combine the detectors for visual search. We refer to the

review paper by Snoek and Worring [19] for a comprehensive discussion. Here we discuss two effective, simple and popular combination methods: the product rule [12] and linear fusion [7], [8], [11], [20]–[22].

If the assumption would hold that individual concepts are conditionally independent given the image content [12], a bi-concept detector can be approximated by the product of its component concepts, namely

$$p(\mathbf{w}_{1,2}|x) = p(w_1|x) \cdot p(w_2|x). \tag{1}$$

The linear fusion version of the bi-concept detector is commonly expressed as

$$p(\mathbf{w}_{1,2}|x) = \lambda \cdot p(w_1|x) + (1-\lambda) \cdot p(w_2|x), \tag{2}$$

where $\lambda \in [0, 1]$ is a weighting parameter. To automatically determine the weight, many have used a heuristic approach. Natsev et al. [21] suggest average fusion with $\lambda = 0.5$. Chang et al. [22] weight the individual single concept detectors in terms of their training performance. However, Snoek et al. [7] argue that the best performing individual detectors such as 'person' and 'outdoor' are often the least informative for retrieval. So Li et al. [11] set λ to be proportional to the informativeness of the concepts. How to determine the proper combination of the single concept detectors remains a open question [19]. It is the reason why many not only show real results but also resort to an oracle combination using the best possible weights [6], [7], [21].

One may also consider utilizing an object localization system which relies on region-based image analysis to pinpoint regions of the single concepts [23]. Such a system involves image segmentation, a challenging problem in computer vision. Moreover, training the system requires learning examples labeled at the region level, which are more expensive to obtain than global level annotations. In contrast, we are interested in searching for bi-concepts by holistic analysis, based on the hypothesis that examples with two concepts co-occurring are likely to have a characteristic holistic scene. Moreover, we obtain training data without the need of manual annotation other than using social tags.

B. Harvesting Training Data from the (Social) Web

Obtaining training examples from the web with expert annotation for free is receiving much attention recently, with sources ranging from generic web images [24]–[28], professional photo forums [29], to social-tagged images [15], [16], [30]–[32]. Training data consists of positive and negative image examples for a given concept. Therefore, we discuss work on positive examples and on negative examples, respectively.

Harvesting Positive Examples. Given a single concept as a textual query, Yanai and Barnard [24] and Li et al. [26] collect positive examples by re-ranking web image retrieval results using probabilistic

models derived from the initial search results. Since the amount of returned examples is limited by the image search engine used, Schroff et al. [27] propose to directly extract images from web search results. As the images vary in quality and come with noisy annotations, dedicated preprocessing such as filtering of drawings and symbolic images is required. The remaining top ranked images are treated as positive examples together with randomly sampled images as negative examples. Based on these examples an SVM classifier is trained and then applied for image re-ranking. As an alternative, Liu et al. [29] rely on a professional photo forum for harvesting training examples, where image quality is considered higher and the annotations are better [33].

In contrast to web images loosely connected with free text, images on the social web are described by user-contributed tags. Moreover, one has access to social-tagged images without any constraint on the amount, making social-tagged images an appealing source for harvesting positive examples. Kennedy et al. [15] and Ulges et al. [16] consider images labeled with a certain concept as positive examples. However, due to the subjectiveness of social tagging, the accuracy of such social positive examples varies per concept [15]. To improve the social tagging accuracy, a number of methods have been proposed, ranging from semantic analysis [17], visual analysis [34] to multi-modal analysis [35], [36]. Zhu et al. in [17] estimate the relevance of a given single concept with respect to an image by measuring the semantic consistency between the concept and the image's social tags. In our previous work [34], we proposed UniformTagger, which estimates image tag relevance by a uniform fusion of neighbor voting results driven by diverse visual features. As determining the relevance for a bi-concept is more difficult than its single concept counterpart, combining textual and visual analysis seems important for obtaining bi-concept examples. Multi-modal analysis by jointly exploiting image-wise visual similarity and tag-wise semantic similarity is considered by Zhu et al. [35] and Liu et al. [36]. As these methods require matrix analysis on the entire image collection and the whole tag vocabulary, their scalability for exploiting a large amount of social-tagged images is questionable.

Harvesting Negative Examples. Surprisingly, in contrast to extensive research on positive examples, the importance of negative examples is often overlooked. The mainstream approach is to randomly sample a relatively small subset from a large pool of images [15]–[17], [27]. For instance, Kennedy et al. [15] and Ulges et al. [16] construct a negative set of fixed size for a given single concept, by randomly sampling from examples not labeled with the concept. If the pool is sufficiently large, one might end up with a set of reliable negatives, but not necessarily the most informative ones.

For bi-concepts, negative examples are even more important as one not only has to distinguish the bi-concept from 'normal' negative classes, but also from its component single concepts. In a labeled image

re-ranking context, Allan and Verbeek [37] suggest to insert examples of the component concepts into the negative set, from which they train an image re-ranking model. In our previous work [32], we proposed a social negative bootstrapping approach to adaptively and iteratively sample informative examples for single concepts, with the prerequisite that manually labeled positive examples are available. However, the prerequisite is unlikely to exist for the bi-concept case.

Given the related work, we consider the absence of the notion of bi-concepts as a major problem for multi-concept search in unlabeled data. For learning bi-concept detectors, the lack of bi-concept training examples is a bottleneck. Previous work on harvesting single-concept examples from social images including our earlier work [32], [34] yields a partial solution, but needs to be reconsidered for bi-concept learning.

III. Bi-Concept Image Search Engine

To make the new notion of bi-concept explicit, we study its characteristics in a bi-concept image search engine for unlabeled data. To search for a specific bi-concept $\mathbf{w}_{1,2}$ in the unlabeled data, we first harvest bi-concept examples from social-tagged images, namely positive examples in Section III-A and negative examples in Section III-B. Our choice of the bi-concept detector $p(\mathbf{w}_{1,2}|x)$ is explained in Section III-C. Finally, we obtain image search results by sorting the unlabeled collection in descending order by $p(\mathbf{w}_{1,2}|x)$.

A. Harvesting Bi-Concept Positive Examples

In order to obtain accurate positive examples for a bi-concept $\mathbf{w}_{1,2}$, we need a large set of social-tagged images and a means to estimate the relevance of a bi-concept with respect to an image. Let X_{social} indicate such a large set, and let $X_{\mathbf{w}_{1,2}+}$ be images in X_{social} which are simultaneously labeled with w_1 and w_2. To simplify our notation, we also use the symbol w to denote a social tag. We define $g(x, w)$ as a single-concept relevance estimator, and $g(x, \mathbf{w}_{1,2})$ as an estimator for bi-concepts. Finally, we denote $\mathbf{w}_x$ as the set of social tags assigned to an image.

We choose two state-of-the-art methods originally designed for the single-concept problem. One method uses semantic analysis [17], and the other method is our previous work, using multi-feature visual analysis [34]. We adapt them to the bi-concept problem: estimating the co-relevance of two single concepts with respect to an image.

The Semantic Method. Under the assumption that the true semantic interpretation of an image is reflected best by the majority of its social tags, a tag that is semantically more consistent with the majority

is more likely to be relevant to the image [17]. We express the semantic-based relevance estimator for single concepts as

$$g_s(x,w) = \frac{\sum_{w' \in \mathbf{w}_x} sim(w',w)}{|\mathbf{w}_x|}, \tag{3}$$

where $sim(w',w)$ denotes semantic similarity between two tags, and $|\cdot|$ is the cardinality of a set. Zhu et al. [17] interpret $sim(w',w)$ as the likelihood of observing w' given w. To cope with variation in tag-wise semantic divergence, we use

$$sim(w',w) = \exp(-\frac{d^2(w',w)}{2\sigma^2}), \tag{4}$$

where $d(w',w)$ measures a semantic divergence between two tags, and the variable σ is the standard derivation of the divergence. Note that (3) is not directly applicable for bi-concepts. To address the issue, we adopt a similarity measure intended for two short text snippets [38], and derive our semantic-based relevance estimator as

$$g_s(x,\mathbf{w}_{1,2}) = \frac{\sum_{w' \in \mathbf{w}_x} maxSim(w',\mathbf{w}_{1,2}) \cdot idf(w')}{\sum_{w' \in \mathbf{w}_x} idf(w')}, \tag{5}$$

where $maxSim(w',\mathbf{w}_{1,2})$ is the maximum value of $sim(w',w_1)$ and $sim(w',w_2)$, and $idf(w')$ is the inverse image frequency of w', reflecting the tag' informativeness.

The Visual Method. Given an image x represented by visual feature f, we first find k nearest neighbors of the image from X_{social}, and estimate the relevance of every single concept w to x in terms of the concept's occurrence frequency in the neighbor set. To overcome the limitation of single features in describing the visual content, tag relevance estimates based on multiple features are uniformly combined [34]. We express the visual-based single-concept relevance estimator as

$$g_v(x,w) \quad = \frac{1}{|F|} \sum_{f \in F} (\frac{c(w, X_{x,f,k})}{k} - \frac{c(w, X_{social})}{|X_{social}|}), \tag{6}$$

where F is a set of features, $c(w,X)$ is the number of images labeled with w in an image set X, and $X_{x,f,k}$ is the neighbor set of x, with visual similarity measured by f.

A straightforward solution to compute (6) for a bi-concept $\mathbf{w}_{1,2}$ is to view the bi-concept as a new tag. This solution boils down to counting the number of images labeled with both w_1 and w_2 in the neighbor set $X_{x,f,k}$. These images are relatively sparse when compared to images labeled with either of w_1 and w_2. The estimator is accurate, but unreliable because $c(\mathbf{w}_{1,2}, X_{x,f,k})$ is often zero or very small. Combining relevance estimates of the two single concepts by linear fusion as described in Section II-A is also problematic, because determining a proper weight is difficult. Simply averaging $g(x,w_1)$ and $g(x,w_2)$ is reliable, yet less accurate. Hence, we need an estimator which accurately reflects the

co-relevance of a bi-concept, and can be computed in a more reliable manner than the straightforward solution. Note the following inequality:

$$c(\mathbf{w}_{1,2}, X) \leq \min\{c(w_1, X), c(w_2, X)\} \leq \frac{1}{2}(c(w_1, X) + c(w_2, X)). \quad (7)$$

In practice the inequality is mostly strict. This means that if we compute $g_v(x, \mathbf{w}_{1,2})$ as the minimum value of $g_v(x, w_1)$ and $g_v(x, w_2)$, the value will be larger than the output of the straightforward solution, and smaller than the output of averaging $g_v(x, w_1)$ and $g_v(x, w_2)$. Moreover, the genuine occurrence of a bi-concept is always lower than any of the two concepts making up the bi-concept. Based on the above discussion, we choose the min function to balance the reliability and the accuracy for bi-concept relevance estimation. Consequently we define our visual-based bi-concept relevance estimator as

$$g_v(x, \mathbf{w}_{1,2}) = \min\{g_v(x, w_1), g_v(x, w_2)\}. \quad (8)$$

An advantage of (8) is that once we have single-concept relevance pre-computed, bi-concept relevance can be rapidly calculated.

Multi-modal: Semantics + Visual. As the *Semantic* method and the *Visual* method are orthogonal to each other, it is sensible to combine the two methods for obtaining bi-concept examples with higher accuracy. As the outputs of $g_s(x, \mathbf{w}_{1,2})$ and $g_v(x, \mathbf{w}_{1,2})$ reside at different scales, normalization is necessary before combining the two functions. Since the Borda count is well recognized as a solid choice for combining rankings generated by multiple sources of evidence [39], [40], we adopt this method for our bi-concept search engine. Given a bi-concept $\mathbf{w}_{1,2}$, we first sort $X_{\mathbf{w}_{1,2}+}$ in descending order by $g_s(x, \mathbf{w}_{1,2})$ and $g_v(x, \mathbf{w}_{1,2})$, respectively. We then aggregate the two rankings by the Borda method to obtain a final ranking. We preserve the top ranked images as positive training examples for the bi-concept detector, denoted as $B_{\mathbf{w}_{1,2}+}$.

Next, we will use $B_{\mathbf{w}_{1,2}+}$ in combination with adaptive sampling to harvest informative negative examples from social-tagged images.

B. Harvesting Bi-Concept Negative Examples

Due to the relatively sparse occurrence of a bi-concept, random sampling already yields a set of accurate negatives. Harvesting negative examples for bi-concepts seems trivial. However, to create an accurate bi-concept detector, we need informative negatives which give the detector better discrimination ability than the random negatives can contribute. We hypothesize that for a given bi-concept, its informative negatives have visual patterns overlapping the patterns of its positive instances. Following this thought,

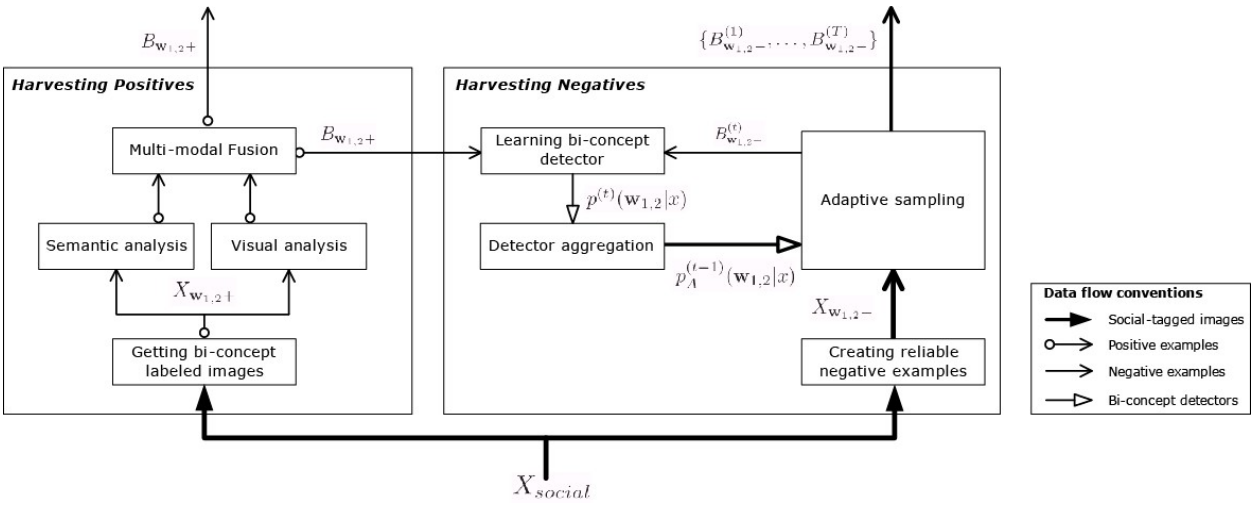


Fig. 3. **Harvesting bi-concept training examples from social-tagged images**: A zoom-in view of the "Harvesting bi-concept examples" component in Fig. 2. We exploit multi-modal analysis to harvest accurate positive examples (Section III-A), and adaptive sampling for informative negative examples (Section III-B). Other than widely available social-tagged images, the process requires no manual annotation.

one might consider positive examples of the individual concepts informative. However, how a bi-concept detector actually works in visual feature spaces with thousands of dimensions is not necessarily consistent with what a human might expect. Given the considerable amount of bi-concepts, it is also impossible to prescribe proper informative negative classes for each bi-concept, say by intensive domain knowledge. Therefore, we leverage the Social Negative Bootstrapping approach proposed in our earlier work [32], and adapt it to the bi-concept search problem. The approach, as detailed next, selects informative negatives from the viewpoint of a detector, but without the need of any human interaction.

Creating Reliable Negative Examples. For a given bi-concept $\mathbf{w}_{1,2}$, we first create a set of reliable negative examples, denoted as $X_{\mathbf{w}_{1,2}-}$, by simple tag reasoning. To describe the procedure, let V_w be a tag set comprised of synonyms and child nodes of w in WordNet [41]. For each image in X_{social}, if the image is not labeled with any tags from $V_{w_1} \cup V_{w_2}$, we add it to $X_{\mathbf{w}_{1,2}-}$.

Adaptive Sampling. Informative negatives are iteratively selected from $X_{\mathbf{w}_{1,2}-}$ by a multiple-round adaptive sampling strategy. Let T be the number of sampling rounds, and $t = 1, \ldots, T$ the index of a specific round. We denote with $p_A^{(t)}(\mathbf{w}_{1,2}|x)$ a bi-concept detector obtained after t rounds. In round t, we first randomly sample n_u samples from $X_{\mathbf{w}_{1,2}-}$ to form a candidate set U_t,

$$U_t \leftarrow \texttt{random-sampling}(X_{\mathbf{w}_{1,2}-}, n_u). \tag{9}$$

Then, we use $p_A^{(t-1)}(\mathbf{w}_{1,2}|x)$ to score each example in U_t, and obtain $\widetilde{U}_t$ in which each example is associated with a confidence score of being positive to the bi-concept,

$$\widetilde{U}_t \leftarrow \texttt{prediction}(U_t, p_A^{(t-1)}(\mathbf{w}_{1,2}|x)). \tag{10}$$

We consider examples which are *most misclassified*, i.e., wrongly predicted as positive with the largest confidence scores, the *most informative* negatives. So we rank examples in $\widetilde{U}_t$ by their scores in descending order and preserve the top ranked examples as the informative negative set found in round t. We denote this new negative set as $B^{(t)}_{\mathbf{w}_{1,2}-}$. By repeating the adaptive sampling procedure, we incrementally select informative negatives from social-tagged images in an adaptive manner.

Learning a New Bi-Concept Detector. In each round t, we learn a new detector $p^{(t)}(\mathbf{w}_{1,2}|x)$ from $B_{\mathbf{w}_{1,2}+}$ and $B^{(t)}_{\mathbf{w}_{1,2}-}$. To prevent class imbalance which often hampers classifier learning [42], we enforce the size of $B^{(t)}_{\mathbf{w}_{1,2}-}$ to be equal to $|B_{\mathbf{w}_{1,2}+}|$.

Detector Aggregation. As $B^{(t)}_{\mathbf{w}_{1,2}-}$ is composed of negatives which are most misclassified by the previous classifier, we consider the new detector $p^{(t)}(\mathbf{w}_{1,2}|x)$ to be complementary to $p_A^{(t-1)}(\mathbf{w}_{1,2}|x)$. Therefore, we use model average to aggregate the two detectors to obtain the final detector:

$$p_A^{(t)}(\mathbf{w}_{1,2}|x) = \frac{t-1}{t} p_A^{(t-1)}(\mathbf{w}_{1,2}|x) + \frac{1}{t} p^{(t)}(\mathbf{w}_{1,2}|x). \tag{11}$$

We illustrate the proposed framework for harvesting bi-concept training data in Fig. 3. We first collect positive examples, and then start the social negative bootstrapping process for obtaining informative negative examples. To trigger the bootstrapping process, we train an initial detector $p^{(1)}(\mathbf{w}_{1,2}|x))$ on $B_{\mathbf{w}_{1,2}+}$ and $B^{(1)}_{\mathbf{w}_{1,2}-}$, which consists of examples randomly sampled from $X_{\mathbf{w}_{1,2}-}$. We cache bi-concept detectors trained in the bootstrapping process so that we do not have to re-train a detector after training data is collected. We use the aggregated detector $p_A^{(T)}(\mathbf{w}_{1,2}|x)$ as the input detector for the "Image search" component in Fig. 2. Given a collection of unlabeled images, our search engine sorts the collection in descending order by $p_A^{(T)}(\mathbf{w}_{1,2}|x)$, and returns the top-ranked results.

C. Learning Bi-Concept Detectors

To learn a concept detector $p^{(t)}(\mathbf{w}_{1,2}|x)$ using the positive set $B_{\mathbf{w}_{1,2}+}$ and the informative negative set $B^{(t)}_{\mathbf{w}_{1,2}-}$, we follow the standard procedure from the literature, namely bag-of-keypoints features [43] plus SVM classifiers [44]. We extract Dense-SIFT features, i.e., dense sampling to localize keypoints and SIFT as a keypoint descriptor, using the state-of-the-art [43]. With the SIFT descriptors quantized by a precomputed codebook, each image is represented by a histogram with its length equal to the size of the

codebook. Each bin of the histogram corresponds to a certain code, and its value is the l_1-normalized frequency of the code extracted from the image. Let $h^{(t)}(x, \mathbf{w}_{1,2})$ be an SVM decision function trained on $B_{\mathbf{w}_{1,2}+}$ and $B^{(t)}_{\mathbf{w}_{1,2}-}$. To convert SVM decision values into posterior probabilities, we adopt a sigmoid function

$$p^{(t)}(\mathbf{w}_{1,2}|x) = \frac{1}{1 + \exp(a \cdot h^{(t)}(x, \mathbf{w}_{1,2}) + b)}, \tag{12}$$

where a and b are two real-valued parameters optimized by solving a regularized maximum likelihood problem as described in [45].

IV. Experimental setup

A. Dataset Construction

Bi-Concepts. In order to evaluate the proposed bi-concept image search engine, we need to specify a list of bi-concepts for our experiments. Since searching for single concepts in unlabeled images remains challenging, the single concepts in a prospective bi-concept shall be detected with reasonable accuracy, otherwise searching for the bi-concept is very likely to be futile. Also, there shall be a reasonable amount of social-tagged training images, say thousands, labeled with the bi-concept. Bearing these considerations in mind, we choose three daily concepts commonly used in the literature [18], [19], [46]–[48], namely: 'beach', 'car', and 'flower'. We obtain bi-concepts by combining the concepts with other objects and scenes, resulting in the following 15 bi-concepts: 'beach + bird', 'beach + boat', 'beach + car', 'beach + girl', 'beach + horse', 'bird + flower', 'bird + snow', 'car + flower', 'car + horse', 'car + showroom', 'car + street', 'car + snow', 'cat + flower', 'cat + snow', and 'girl + horse'. While the list of potential bi-concepts is exhaustive, the selection serves as a nontrivial illustration of bi-concept possibilities.

Social Source for Harvesting Training Examples. We use the 15 bi-concepts as well as the 11 single concepts from which they are composed as queries to randomly sample images uploaded on Flickr between 2006 and 2010. We remove batch-tagged images due to their low tagging accuracy, and obtain 300K images in total. To harness a large data set for multi-modal analysis, we further gather 900K social-tagged images from Flickr in a random fashion. Our total training collection thus contains 1.2 million images. We list the single concept and bi-concept statistics in Table I and Table II.

Test Data. For each bi-concept, we create a ground truth positive set by manually checking images labeled with the bi-concept in the 1.2M set, and randomly selecting 50 positively labeled examples. Although the test images are associated with social tags, we ignore the tags and treat the images as unlabeled. Note that these selected examples are held-out from the training process. We supplement the

TABLE I

EXPERIMENT 1. COMPARING METHODS FOR HARVESTING POSITIVE TRAINING EXAMPLES OF SINGLE CONCEPTS, MEASURED IN TERMS OF PRECISION AT 100. WE SORT THE CONCEPTS BY THEIR FREQUENCY IN THE 1.2 MILLION SET IN DESCENDING ORDER. A GRAY CELL INDICATES THE TOP PERFORMER.

Concepts		**Social Tagging Baselines**				**Proposed Search Engine**		
w	*Frequency*	*Random*	*DateUploaded*	*Views*	*TagNum*	*Semantics*	*Visual*	*Borda*
car	71,367	0.69	0.75	0.87	0.61	0.85	**1.00**	0.99
flower	64,233	0.79	0.69	0.64	0.94	0.95	**1.00**	**1.00**
street	61,877	0.52	0.55	0.66	0.42	0.47	**1.00**	0.96
beach	47,636	0.53	0.53	0.69	0.59	0.63	**0.97**	0.95
snow	42,327	0.82	0.85	0.77	0.73	0.90	**1.00**	0.99
bird	33,841	0.79	0.80	0.67	0.94	0.92	**1.00**	0.99
girl	32,983	0.75	0.75	0.91	0.79	0.85	**0.97**	0.94
horse	28,724	0.70	0.60	0.74	0.79	0.85	**1.00**	**1.00**
cat	19,712	0.67	0.68	0.56	0.82	0.96	0.99	**1.00**
boat	15,239	0.75	0.75	0.74	0.76	0.85	0.94	**0.97**
showroom	4,947	0.43	0.43	0.61	0.34	0.34	**0.95**	0.78
MEAN	38,444	0.68	0.67	0.71	0.70	0.78	**0.98**	0.96

collection of bi-concept images with distracter images from the publicly available NUS-WIDE set [46], which is also from Flickr but independent of our training data. Since this set was constructed by single-concept queries, it rarely contains genuine positives of the 15 bi-concepts. For reasons of efficiency, we randomly sample a subset of 10K images from NUS-WIDE as our negative test data. For each bi-concept, we examine how its 50 positive examples are ranked within the 10K negative set.

B. Experiments

In order to provide a step-by-step analysis on the entire bi-concept search framework, we evaluate in the following two experiments: the accuracy of harvested positive training examples and the various mechanisms for bi-concept search.

Experiment 1. Harvesting Bi-Concept Positive Examples. For each concept, we take all images labeled with the concept in our 1.2M set as candidate positives. We sort the candidate set by each of the three methods described in Section 3, namely Semantics, Visual, and Multi-modal Borda. In addition to the *Borda* method, we also consider multi-kernel learning plus SVM [49], which directly combines multi-modal similarities. For each bi-concept, we train a multi-kernel SVM on images labeled with the

TABLE II

EXPERIMENT 1. COMPARING METHODS FOR HARVESTING POSITIVE TRAINING EXAMPLES OF BI-CONCEPTS, MEASURED IN TERMS OF PRECISION AT 100. WE SORT THE BI-CONCEPTS BY THEIR FREQUENCY IN THE 1.2 MILLION SET IN DESCENDING ORDER.

Bi-Concepts			**Social Tagging Baselines**				**Proposed Search Engine**			
w_1	w_2	*Frequency*	*Random*	*DateUploaded*	*Views*	*TagNum*	*Semantics*	*Visual*	*Multi-kernel*	*Borda*
car	*street*	22788	0.64	0.57	0.70	0.53	0.58	**0.97**	0.23	0.86
car	*snow*	7109	0.62	0.62	0.54	0.68	0.66	**0.91**	0.67	0.85
beach	*car*	5432	0.10	0.14	0.11	0.19	0.25	0.27	0.42	**0.43**
car	*flower*	3604	0.13	0.11	0.05	0.39	0.40	0.21	0.37	**0.42**
beach	*girl*	3507	0.29	0.36	0.60	0.53	0.57	0.60	**0.78**	0.76
beach	*bird*	3093	0.26	0.25	0.20	0.51	0.57	0.56	**0.68**	0.63
beach	*boat*	2659	0.42	0.36	0.50	0.62	**0.66**	0.39	0.53	0.58
cat	*flower*	2316	0.07	0.05	0.05	0.55	**0.59**	0.14	0.55	0.42
bird	*flower*	2103	0.11	0.10	0.11	0.38	0.43	0.36	0.41	**0.51**
car	*horse*	1496	0.19	0.15	0.16	0.41	0.29	0.46	0.22	**0.59**
bird	*snow*	1352	0.64	0.44	0.52	0.77	0.75	0.77	**0.94**	0.83
car	*showroom*	1301	0.55	0.70	0.85	0.61	0.72	**0.92**	0.70	0.86
cat	*snow*	788	0.20	0.18	0.07	0.48	0.57	0.46	0.58	**0.65**
girl	*horse*	692	0.45	0.48	0.44	0.69	0.68	0.71	0.66	**0.77**
beach	*horse*	622	0.48	0.57	0.53	0.69	0.71	0.65	**0.81**	0.80
MEAN		3924	0.34	0.34	0.36	0.54	0.56	0.56	0.57	**0.66**

bi-concept, and then use the SVM to predict the positive training examples. For a more comprehensive comparison, we also report the performance of image ranking using three simple metadata features: DateUploaded, TagNum, and Views. Given an image, TagNum is the number of tags contributed by its user, while Views indicates how many times the image has been viewed.

As there is no ground-truth available for the 1.2M set, we manually check for genuine positives in the top ranked images. To reduce the manual annotation effort and (possible) labeling bias towards certain methods, we employ a pooling mechanism similar to the TRECVid benchmark [47]. For each method, we put its top 100 ranked images into a common pool without indicating their origin. For a given query of a single or bi-concept, we label an image as positive if the (bi-)concept is (partially) visible in the image. Artificial correspondences of the (bi-)concepts such as drawings, toys, and statues are labeled as negative. Notice that as the chance of including genuine positives in the negative sets is very small, we do not assess the accuracy of the negatives.

Experiment 2. Bi-Concept Search in Unlabeled Images. To configure a bi-concept search engine, we have to specify the following three choices:

1) detector: building a bi-concept detector versus combining the confidence scores of two single-concept detectors,

2) positive: random sampling versus the multi-modal Borda fusion of *Semantic* and *Visual* selection,

3) negative: random sampling versus adaptive sampling.

In order to study the impact of the individual choices on bi-concept search, we design three setups for a head-to-head comparison, namely *Social*, *Borda*, and *Full*, as listed in Table III. The optimal choice of the amount of positive examples may vary over bi-concepts. For bi-concepts whose positive data can be collected at a higher accuracy, it is sensible to preserve more top ranked examples for training. Note that this study does not aim for the best possible performance, but rather focuses on revealing the advantages of bi-concepts as a retrieval method, in the context of the existing works using single-concept detectors. Hence, for each setup, we simply set the number of positive examples per bi-concept to 100. For harvesting informative negative examples, we set the number of iterations T to 10. Consequently, we also create 10 sets of randomly sampled negatives for the reason of fair comparisons. By comparing *Borda* and *Social*, we study the impact of positive training data. By comparing *Full* and *Borda*, we assess the effectiveness of informative negatives.

For combining single-concept detectors, we investigate the most common methods from the retrieval literature: the product rule and linear fusion. While the product rule helps to make a combined detector more discriminative, averaging the individual detectors helps to improve the robustness of the combined detector. Linear fusion is often used to establish a performance upper bound [6], [7]. We follow this idea, and establish a performance upper bound of linear fusion for bi-concept search by grid search with a step size of 0.05 on λ. We use average precision, a common choice for evaluating visual search engines [47].

C. Implementation

Parameters for Training (Bi-)Concept Detectors. We create a codebook with a size of 1,024 by K-means clustering on a held-out set of random Flickr images. So each image is represented by a vector quantized Dense-SIFT histogram of 1,024 dimensions. For a fair comparison between detectors trained using different setups, we train a two-class SVM using the χ^2 kernel, setting the cost parameter to 1.

Parameters for the Semantic Method. As an instantiation of $d(w', w)$ in (4), we choose the Normalized Google Distance [50], which measures semantic divergence between two tags based on their (co-)occurrence frequency in a large collection of social-tagged images. As our 1.2M set might be relatively

TABLE III

EXPERIMENT 2. CONFIGURING OUR BI-CONCEPT IMAGE SEARCH ENGINE USING THREE SETUPS.

Setup	Positive training data	Negative training data
Social	100 examples randomly sampled from $X_{\mathbf{w}_{1,2}+}$	10 negative sets, each having 100 randomly generated negatives
Borda	The top 100 examples retrieved by Borda count	The same negatives as *Social*
Full	The same positives as *Borda*	Social negative bootstrapping with T=10, n_u=1000

small for computing this distance, we use the full list of LSCOM concepts [18] as queries, and collect up to 10 million Flickr images with social tags. The idf values in (5) are also computed on the 10M set.

Parameters for the Visual Method. We choose the following four visual features which describes image content from different perspectives: COLOR, CSLBP, GIST, and Dense-SIFT. COLOR is a 64-d global feature, combining the 44-d color correlogram [51], the 14-d texture moments [52], and the 6-d RGB color moments. CSLBP is a 80-d center-symmetric local binary pattern histogram [53], capturing local texture distributions. GIST is a 960-d feature describing dominant spatial structures of a scene by a set of perceptual measures such as naturalness, openness, and roughness [54] (using software from [55]). Dense-SIFT [43] is the same bag-of-keypoints feature as we have used for concept detection. We compute (6) with the feature set $F = \{$COLOR, CSLBP, GIST, Dense-SIFT$\}$ on the 1.2M set. We set k=1,000, a good choice for the neighbor voting algorithm [34].

Parameters for the Multi-kernel SVM. We construct multiple kernels as follows. For each of the four visual features, we use the χ^2 kernel. To measure the semantic similarity between two images, we adopt the choice of Guillaumin et al. [56], and define a tag kernel which returns the number of tags shared by two images. To train a multi-kernel SVM, we take the top 100 examples ranked by TagNum as positive training data and 100 examples sampled at random as negative training data. We use the Shogun software [49], with the l2 normalization on the combination weights

V. RESULTS

A. Experiment 1. Harvesting Bi-Concept Positive Examples

Social Tagging Baselines. Comparing single concept harvesting results in Table I and bi-concepts harvesting results in Table II, we observe that the social tagging accuracy of bi-concepts (P@100=0.34) is much lower than its single-concept counterpart (P@100=0.68). Recall that we already removed batch-tagged images beforehand. So a possible explanation could be that when users label images with multiple

(a) *Views*

(b) *Semantics*

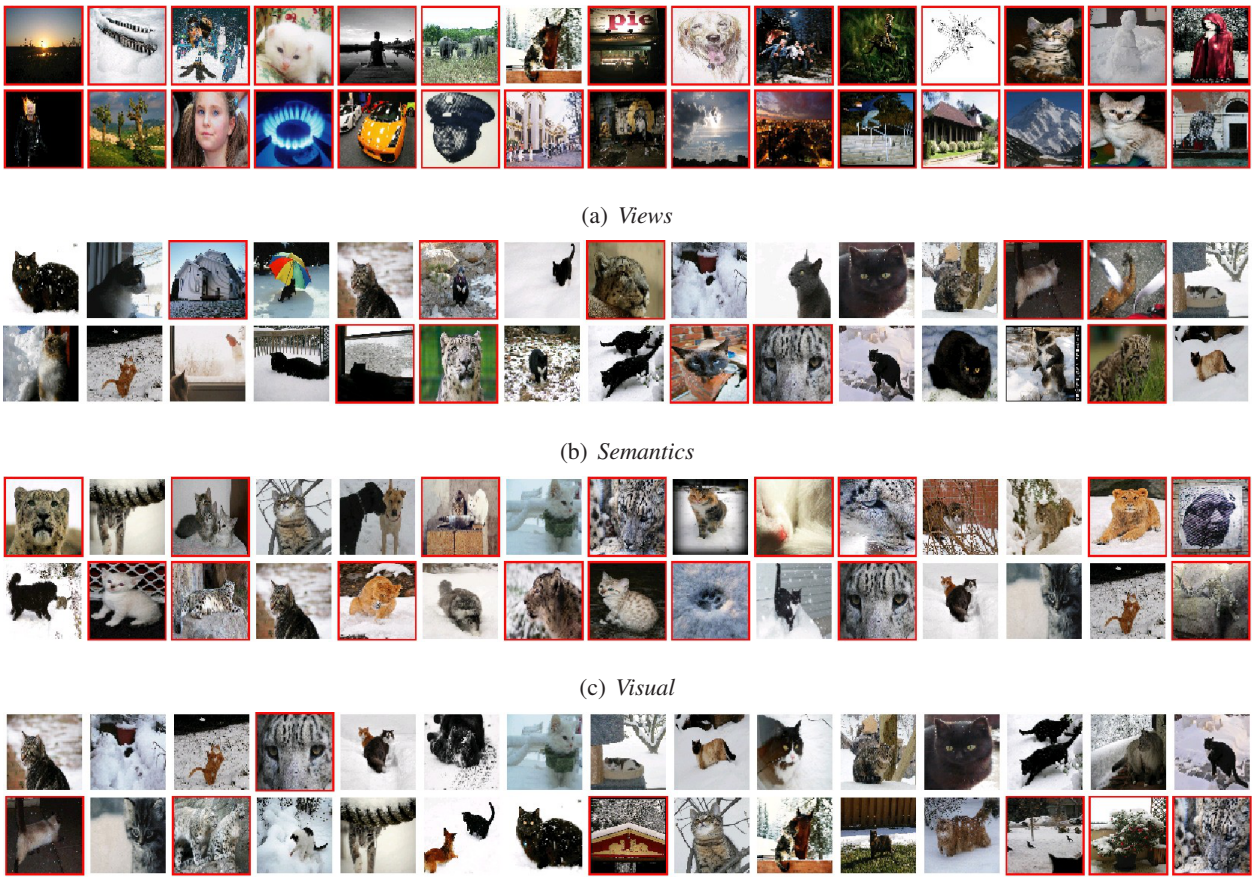
(c) *Visual*

(d) *Multi-modal Borda*

Fig. 4. **Positive training examples of 'car + horse' automatically obtained from social-tagged images by different methods**. The top 30 results of each method are shown. A red border indicates a false positive example.

tags, they tend to add more tags irrelevant to the actual content to improve the retrievability of their images. This explanation is confirmed to some extent by the behavior of the *TagNum* feature. While *TagNum* is slightly better than random sampling for single concepts, it clearly outperforms random sampling for bi-concepts. Simply ranking images by *DateUploaded* does not improve the accuracy at all, indicating that freshness has no impact on relevance. The result of Views (P@100=0.71 for single concepts and P@100=0.36 for bi-concepts) shows popularity has a limited positive impact on relevance.

Bi-Concept Search Engine versus Social Tagging Baselines. As shown in Table II, for bi-concepts, our search engine with multi-modal *Borda* doubles the accuracy, with P@100=0.66, when compared to random sampling from the social web with P@100=0.34. The results show the effectiveness of the proposed search engine for harvesting positive training examples from social-tagged images. We



(a) *Views*

(b) *Semantics*

(c) *Visual*

(d) *Multi-modal Borda*

Fig. 5. **Positive training examples of 'cat + snow' automatically obtained from social-tagged images by different methods**. The top 30 results of each method are shown. A red border indicates a false positive example.

demonstrate some harvested bi-concept training examples in Fig. 4 and Fig. 5. Compared to the *Random* run, the performance of the *Visual* run improves for all bi-concepts except for 'beach + boat'. For that bi-concept, *Visual* incorrectly ranks images of "boats in sea" at the top due to visual ambiguity between sea and beach. Although the multi-kernel method performs well for some bi-concepts such as 'bird + snow', it is not as effective as the *Borda* method in general.

Multi-modal versus Uni-modal. For single concepts, *Visual* reaches the best performance, on average, having 98 genuine positive examples in its top 100 retrieved results. We attribute the success of *Visual* to two reasons. First, while visual appearance of a single concept, e.g, 'bird', may vary significantly, the typical visual context where the concept is observed is relatively consistent, e.g., 'water' and 'sky' for 'bird'. The *Visual* method, empowered by diverse visual features, thus estimates single concept relevance

accurately. Second, rather than re-ranking a small number of image search results [27], [37], we analyze all images labeled with a given single concept. The large candidate pool allows the search engine to find the images for which it is most confident. The results for bi-concepts are different. Neither of the two uni-modal methods is a clear winner. For bi-concepts with 'flower' as a component concept, *Semantics* tend to outperform Visual. Recall that the features used in our experiments are global, meaning they are better at describing visual context than capturing insignificant objects within an image. As a flower is often small in an example of 'car + flower', retrieving a number of flower images in the neighbor set becomes difficult. Region-level analysis could be helpful in this case. As *Semantics* and *Visual* are complementary, combining them with the Borda count method results in a relative improvement of 18%.

In sum, our main findings after experiment 1 are as follows. Since the social tagging accuracy of bi-concepts is much lower than that of single-concepts, harvesting positive bi-concept examples is more difficult than harvesting positive single-concept examples. While visual analysis seems adequate for single-concepts, multi-modal analysis is crucial for bi-concepts. When compared to selecting bi-concept labeled images from the social web in a random fashion, our proposed bi-concept search engine harvests bi-concept examples with doubled accuracy.

B. Experiment 2. Bi-Concept Search in Unlabeled Images

Comparing Methods for Combining Single Concepts. For single concept search, unsurprisingly, single concepts trained using the *Full* setup, with an MAP of 0.120, is better than single concepts trained using the *Social* setup, with an MAP of 0.080. As shown in Fig. 6, compared to single concepts trained on random samples, single concepts learned from informative negatives are more discriminative, favoring precision over recall. As shown in Table IV, the product rule works slightly better than average fusion for combining single concepts. This result implies that searching for bi-concepts demands detectors with better discrimination ability.

Bi-Concept Search Engine versus Combining Single Concepts. As shown in Table IV, our bi-concept search engine, using the *Full* setup, performs best, with an MAP of 0.106. For single concept detectors trained using the *Full* setup, the upper bound on the performance of linear fusion with an oracle is 0.053 , which is unlikely to be approached in practice. Even with this upper bound, we still outperform linear fusion for most bi-concepts, and overall with a relative improvement of 100%.

The Impact of Positive Training Data. Concerning positive training data for learning bi-concept detectors, the *Borda* setup improves the MAP of the search engine from 0.042 to 0.080 when compared to *Social*. The bi-concept comparison shows that for most bi-concepts *Borda* is better than *Social*. Because

TABLE IV

EXPERIMENT 2. COMPARING METHODS FOR BI-CONCEPT SEARCH IN TERMS OF AVERAGE PRECISION. WE COMPARE OUR PROPOSED BI-CONCEPT SEARCH ENGINE WITH APPROACHES COMBINING SINGLE CONCEPTS USING PRODUCT RULE, LINEAR FUSION WITH $\lambda = 0.5$ AND LINEAR FUSION WITH AN ORACLE ESTIMATOR FOR λ. FOR *Borda* AND *Full*, THEIR POSITIVE TRAINING DATA ARE OBTAINED BY THE BORDA COUNT METHOD REPORTED IN TABLE II.

Bi-Concepts		$p(w_1\|x) * p(w_2\|x)$			$0.5p(w_1\|x) + 0.5p(w_2\|x)$			$\lambda p(w_1\|x) + (1-\lambda)p(w_2\|x)$			**Proposed Search Engine**		
w_1	w_2	*Social*	*Borda*	*Full*	*Social*	*Borda*	*Full*	*Social*	*Borda*	*Full*	*Social*	*Borda*	*Full*
car	*street*	0.033	0.039	0.049	0.033	0.041	**0.051**	0.015	0.019	0.036	0.041	0.040	0.050
car	*snow*	0.014	0.017	0.018	0.014	0.014	0.019	0.034	0.044	0.053	0.026	0.056	**0.109**
beach	*car*	0.040	0.035	0.035	0.040	0.032	0.033	0.042	0.033	0.037	0.037	0.040	**0.068**
car	*flower*	0.006	0.010	0.011	0.007	0.009	0.009	0.010	0.012	**0.013**	0.009	0.010	0.011
beach	*girl*	0.058	0.040	0.027	0.054	0.012	0.019	0.054	0.026	0.026	0.039	0.139	**0.180**
beach	*bird*	0.129	0.125	0.123	0.131	0.122	0.125	0.143	0.129	0.126	0.085	**0.195**	0.188
beach	*boat*	0.064	0.055	0.059	0.063	0.054	0.056	**0.082**	0.056	0.056	0.039	0.067	0.075
cat	*flower*	0.015	0.018	0.020	0.014	0.018	0.021	0.021	0.023	**0.032**	0.006	0.017	0.025
bird	*flower*	0.019	0.015	0.016	0.020	0.015	0.015	0.024	0.020	0.019	0.011	**0.039**	0.022
car	*horse*	0.032	0.029	0.032	0.032	0.017	0.021	0.048	0.037	0.026	0.022	0.038	**0.061**
bird	*snow*	0.012	0.012	0.010	0.012	0.013	0.011	0.013	0.018	0.016	0.047	0.067	**0.079**
car	*showroom*	0.108	0.103	0.118	0.108	0.106	0.130	0.108	0.157	0.197	0.142	0.185	**0.271**
cat	*snow*	0.014	0.015	0.014	0.014	0.017	0.020	0.014	0.020	0.057	0.017	0.091	**0.096**
girl	*horse*	0.021	0.026	0.038	0.020	0.028	0.034	0.024	0.035	0.038	0.028	0.030	**0.049**
beach	*horse*	0.036	0.069	0.086	0.041	0.062	0.041	0.053	0.069	0.061	0.085	0.193	**0.309**
MEAN		0.040	0.041	0.044	0.040	0.037	0.040	0.046	0.047	0.053	0.042	0.080	**0.106**

Borda and *Social* use the same negative training data, the result allows us to conclude that positive examples harvested by our system are better than the original social-tagged positives.

The Impact of Negative Training Data. Comparing the *Full* setup with the *Borda* setup, we observe from Table IV that for most bi-concepts *Full* surpasses *Borda*, with a relative improvement of 32.5%, in terms of MAP. Since the two setups use the same positive training data, the results show the importance of informative negatives for accurate bi-concept search. To see what negative classes are recognized as informative for a given bi-concept, we show in Fig. 7 the most informative negative examples harvested from the 1.2M set by the proposed search engine. Note that negative examples which are visually close to bi-concept examples are automatically identified as informative for optimizing bi-concept detectors. Consider 'car + showroom' for instance. As shown in Fig. 7(a), indoor scenes such as offices and

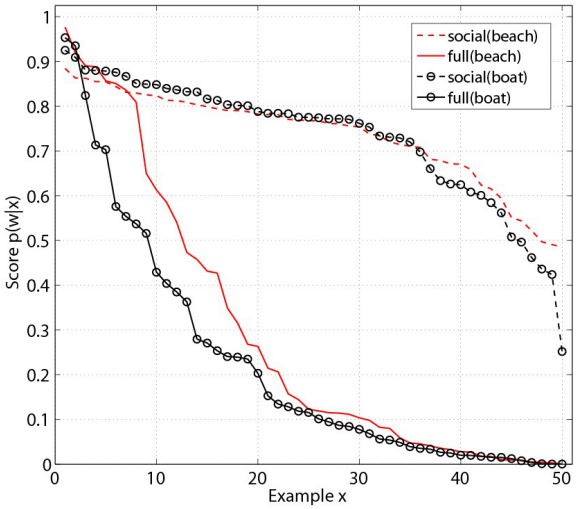


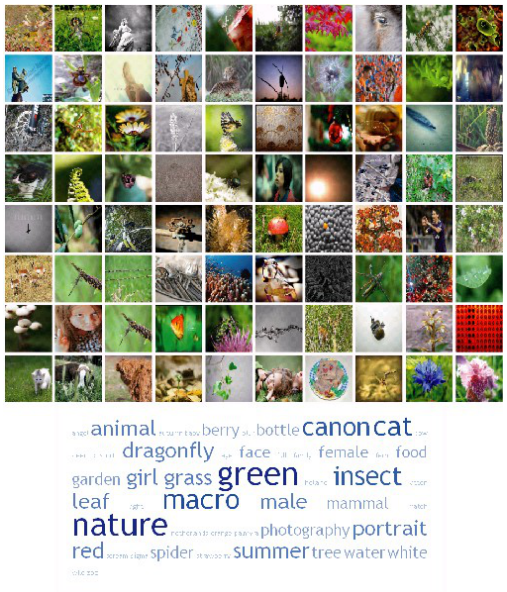
Fig. 6. **Comparing predicted scores of single concept detectors trained using different setups**. social(beach): a 'beach' detector trained using the *Social* setup. full(beach): a 'beach' detector trained using the *Full* setup. social(boat) and full(boat) are defined in a similar fashion. Each curve is obtained by running a detector on 50 positive examples of 'beach + boat', and sorting by predicted scores in descending order. Steeper slopes indicate that detectors trained using the *Full* setup are more discriminative, favoring precision over recall.

restaurants and outdoor scenes such as streets are selected by our system. Images such as close-ups of electronic devices, as one often see in a showrooms, are also found by our system. These negative examples are helpful for the search engine to distinguish genuine examples of 'car + showroom' from examples where only one of the two single concepts is present, resulting in an absolute improvement of 0.086. Further, by visualizing social tag frequency in the informative negative set with a tag cloud, we see which negative classes are most informative with respect to a specific bi-concept. Both the quantitative and qualitative results demonstrate the viability of the proposed bi-concept image search engine.

Concerning generalization of the proposed method for more complex queries such as "finding an image showing a girl and a horse on a beach", a straightforward extension is to harvest examples for the tri-concept 'beach + girl + horse'. Though images with the three tags are relatively sparse, positive examples of the tri-concept may have a more characteristic visual appearance. Hence less training data is required. Using the same protocol as used for the bi-concepts, we have conducted an additional experiment for searching for 'beach + girl + horse'. Compared to average fusion of the three single concepts with an AP of 0.058, the proposed search engine obtains a much better performance with an AP of 0.290. Fig. 8 shows the top 20 search results by different methods. The results demonstrate the potential of our method

(a) Informative negative training examples of 'car + showroom'

(b) Informative negative training examples of 'beach + girl'



(c) Informative negative training examples of 'bird + flower'

(d) Informative negative training examples of 'car + horse'

Fig. 7. **The 80 most informative negative examples for specific bi-concepts, harvested from social-tagged images by the proposed bi-concept image search engine**. By visualizing tag frequency in the selected negatives as a tag cloud, we see which negative classes are most informative to a given bi-concept.

(a) Searching for 'beach + girl + horse' by average fusion of beach, girl, and horse detectors

(b) Searching for 'beach + girl + horse' by the proposed search engine

Fig. 8. **Finding unlabeled images with three visual concepts co-occurring**. Compared to average fusion of single concepts, the proposed search engine obtains better search results for tri-concept 'beach + girl + horse'.

for tri-concept search.

VI. Discussion and Conclusions

This paper establishes *bi-concepts* as a new method for searching for the co-occurrence of two visual concepts in unlabeled images. To materialize, we propose a bi-concept image search engine. This engine is equipped with bi-concept detectors directly, rather than artificial combinations of individual single-concept detectors. Since the cost of manually labeling bi-concept training examples is prohibitive, harvesting social images is one – if not the – main enabler to learn bi-concept semantics.

The core of our search engine is a multimedia data-driven framework which collects from the social web 1) de-noised positive training examples by multi-modal analysis and 2) informative negative training examples by adaptive sampling. We study the behavior of the search engine using 1.2M social-tagged images as a data source.

Obtaining positive training examples for bi-concepts is more difficult than for single concepts, as the social tagging accuracy of bi-concepts is much lower. For single concepts, uni-modal (visual) analysis is often sufficient for de-noising. For bi-concepts, multi-modal analysis is crucial, gaining a relative improvement of 18% over uni-modal. When compared to the social tagging baseline, we obtain positive examples of bi-concepts with doubled accuracy.

The training examples, obtained without the need of any manual annotation other than social tags, are used to train bi-concept detectors. These detectors are applied to 10K unlabeled images. Using the de-

noised positive data allows us to lift the performance of the social baseline from 0.042 to 0.080, in terms of MAP. Substituting informative negative examples for randomly sampled negatives further improves the performance, reaching an MAP of 0.106. Our system even compares favorably to the oracle linear fusion of single concept detectors, with an upper bound MAP of 0.053. The results allow us to conclude that compared to existing methods which combine single concept detectors, the proposed method is more effective for bi-concept search in unlabeled data.

One concern of the paper might be that the number of bi-concepts in our current evaluation is relatively small, when compared to single concept benchmarks [46]–[48]. Though our framework needs no manual verification for exploiting bi-concept examples, we actually require manually verified ground truth for a head-to-head comparison. Therefore, a novel benchmark dedicated to bi-concepts or even higher-order semantics is urged for.

Our study is orthogonal to work which aims to detail a single concept by describing it's visual attributes [13], [14], e.g., automatically adding the tag 'red' to 'car' to generate a more specific single concept 'red car'. These methods might be incorporated into our bi-concept search engine to answer bi-concept queries with two specified single concepts such as 'red car + white horse'. This would lead to a search engine capable of answering very precise queries.

Our proposed methodology is a first step in deriving semantics from images which goes beyond relatively simple single-concept detectors. We believe that for specific pre-defined bi-concepts, they already have great potential for use in advanced search engines. Moving to on-the-fly trained queries based on bi-concepts opens up promising avenues for future research.

REFERENCES

[1] C. Snoek and A. Smeulders, "Visual-concept search solved?" *IEEE Computer*, vol. 43, no. 6, pp. 76–78, 2010.

[2] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, 2008.

[3] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, pp. 465–476, 2009.

[4] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 19, pp. 733–746, 2009.

[5] R. Datta, W. Ge, J. Li, and J. Wang, "Toward bridging the annotation-retrieval gap in image search," *IEEE Multimedia*, vol. 14, no. 3, pp. 24–35, 2007.

[6] 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. Multimedia*, vol. 9, no. 5, pp. 958–966, 2007.

[7] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.

[8] X.-Y. Wei, Y.-G. Jiang, and C.-W. Ngo, "Concept-driven multi-modality fusion for video search," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 1, pp. 62–73, 2011.

[9] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua, "Utilizing related samples to enhance interactive content-based video search," *IEEE Trans. Multimedia*, 2011, in press.

[10] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *ACM Multimedia*, 2007.

[11] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *ACM CIVR*, 2007.

[12] R. Aly, D. Hiemstra, A. de Vries, and F. de Jong, "A probabilistic ranking framework using unobservable binary events for video search," in *CIVR*, 2008.

[13] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *ICCV*, 2009.

[14] K. Yang, X.-H. Hua, M. Wang, and H.-J. Zhang, "Tag tagging: towards more descriptive keywords of image content," *IEEE Trans. Multimedia*, 2011, in press.

[15] 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.

[16] A. Ulges, C. Schulze, M. Koch, and T. Breuel, "Learning automatic concept detectors from online video," *Comput. Vis. Image Underst.*, vol. 114, no. 4, pp. 429–438, 2010.

[17] 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.

[18] M. Naphade, J. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[19] C. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retr.*, vol. 2, pp. 215–322, 2009.

[20] R. Yan and A. Hauptmann, "The combination limit in multimedia retrieval," in *ACM Multimedia*, 2003.

[21] A. Natsev, M. Naphade, and J. TešiĆ, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *ACM Multimedia*, 2005.

[22] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia university trecvid-2006 video search and high-level feature extraction," in *TRECVID workshop*, 2006.

[23] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR*, 2011.

[24] K. Yanai and K. Barnard, "Probabilistic web image gathering," in *ACM MIR*, 2005.

[25] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, 2008.

[26] L.-J. Li and L. Fei-Fei, "OPTIMOL: Automatic online picture collection via incremental model learning," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 147–168, 2010.

[27] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 754–766, 2011.

[28] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma, "ARISTA - image search to annotation on billions of web photos," in *CVPR*, 2010.

[29] Y. Liu, D. Xu, I. Tsang, and J. Luo, "Textual query of personal photos facilitated by large-scale web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 1022–1036, 2011.

[30] Y. Shen and J. Fan, "Leveraging loosely-tagged images and inter-object correlations for tag recommendation," in *ACM Multimedia*, 2010.

[31] N. Sawant, J. Li, and J. Wang, "Automatic image semantic interpretation using social action and tagging data," *Multimedia Tools Appl.*, vol. 51, pp. 213–246, 2011.

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

[33] L. Zhang, L. Chen, F. Jing, K. Deng, and W.-Y. Ma, "EnjoyPhoto: a vertical image search engine for enjoying high-quality photos," in *ACM Multimedia*, 2006.

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

[35] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *ACM Multimedia*, 2010.

[36] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *ACM Multimedia*, 2010.

[37] M. Allan and J. Verbeek, "Ranking user-annotated images for multiple query terms," in *BMVC*, 2009.

[38] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, 2006.

[39] J. Aslam and M. Montague, "Models for metasearch," in *SIGIR*, 2001.

[40] X. Olivares, M. Ciaramita, and R. van Zwol, "Boosting image retrieval through aggregating search results based on visual annotations," in *ACM Multimedia*, 2008.

[41] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[42] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002.

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

[44] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.

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

[46] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *CIVR*, 2009.

[47] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *ACM MIR*, 2006.

[48] M. Huiskes, B. Thomee, and M. Lew, "New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative," in *ACM MIR*, 2010.

[49] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc, "The shogun machine learning toolbox," *J. Mach. Learn. Res.*, vol. 11, pp. 1799–1802, 2010.

[50] R. Cilibrasi and P. Vitanyi, "The Google similarity distance," in *IEEE Trans. Knowl. and Data Eng.*, 2004.

[51] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *CVPR*, 1997.

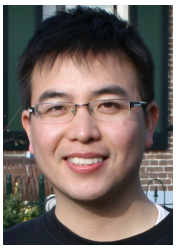
[52] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moment for content-based image retrieval," in *ICIP*, 2002.

[53] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recogn.*, vol. 42, pp. 425–436, 2009.

[54] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[55] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *CIVR*, 2009.

[56] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *CVPR*, 2010.



Xirong Li received the Ph.D. degree (2012) from the University of Amsterdam, and the M.Sc. (2007) and B.Sc. (2005) degrees from the Tsinghua University, all in computer science. He is currently an Assistant Professor in the MOE key laboratory of Data Engineering and Knowledge Engineering, Renmin University of China. His research interest is visual search in varied context. He received the Best Paper Award of the ACM International Conference on Image and Video Retrieval 2010.

Cees G.M. Snoek received the M.Sc. degree in business information systems (2000) and the Ph.D. degree in computer science (2005), both from the University of Amsterdam, Amsterdam, The Netherlands. He is currently an Assistant Professor in the Intelligent Systems Lab at the University of Amsterdam. He was a visiting scientist at Carnegie Mellon University, Pittsburgh, PA (2003) and at the University of California, Berkeley, CA (2010-2011). His research interest is video and image search. Dr. Snoek is the lead researcher of the MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is a co-initiator and co-organizer of the VideOlympics, co-chair of the SPIE Multimedia Content Access conference, associate editor for IEEE MultiMedia, and guest editor for IEEE Transactions on Multimedia, special issue on Socio-Video Semantics. He is recipient of a young talent VENI grant from the Dutch Organization for Scientific Research (2008) and a Fulbright visiting scholar grant (2010).

Marcel Worring received the M.Sc. degree (honors) in computer science from the VU Amsterdam, Amsterdam, The Netherlands, in 1988 and the Ph.D. degree in computer science from the University of Amsterdam in 1993. He is currently an Associate Professor in the Informatics Institute of the University of Amsterdam. His research focus is multimedia analytics, the integration of multimedia analysis, multimedia mining, information visualization, and multimedia interaction into a coherent framework yielding more than its constituent components. He has published over 100 scientific papers covering a broad range of topics from low-level image and video analysis up to multimedia analytics. Dr. Worring was co-chair of the 2007 ACM International Conference on Image and Video Retrieval in Amsterdam, co-initiator and organizer of the VideOlympics, program chair for the ICMR 2013, and the ACM Multimedia 2013. He was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and he is Associate Editor of the Pattern Analysis and Applications journal.

Arnold W.M. Smeulders is at the national research institute CWI in Amsterdam leading COMMIT, a nation-wide, very large public-private research program. He is also chair of IPN, the national policy committee for research in computer science. And he is with the ISIS group at the University of Amsterdam for research in the theory and practice of visual search. He is co-owner of Euvision Technologies BV, a company spun off from the UvA. He is associate editor of the IJCV. He was recipient of a Fulbright fellowship at Yale University, and visiting professor in Hong Kong, Tuskuba, Modena and Cagliari. He is fellow of the International Association of Pattern Recognition, and honorary member of the Dutch Society for Pattern Recognition. He was general chairman of IEEE and ACM conferences on Multimedia.