A Latent Model for Visual Disambiguation of Keyword-based Image Search

Kong-Wah WAN[1]
kongwah@i2r.a-star.edu.sg
Ah-Hwee TAN[2]
asahtan@ntu.edu.sg
Joo-Hwee LIM[1]
joohwee@i2r.a-star.edu.sg
Liang-Tien CHIA[2]
asltchia@ntu.edu.sg
Sujoy ROY[1]
sujoy@i2r.a-star.edu.sg

[1] Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613

[2] School of Computer Engineering
Nanyang Technological University
Singapore

Abstract

The problem of polysemy in keyword-based image search arises mainly from the inherent ambiguity in user queries. We propose a latent model based approach that resolves user search ambiguity by allowing sense specific diversity in search results. Given a query keyword and the images retrieved by issuing the query to an image search engine, we first learn a latent visual sense model of these polysemous images. Next, we use Wikipedia to disambiguate the word sense of the original query, and issue these Wiki-senses as new queries to retrieve sense specific images. A sense-specific image classifier is then learnt by combining information from the latent visual sense model, and used to cluster and re-rank the polysemous images from the original query keyword into its specific senses. Results on a ground truth of 17K image set returned by 10 keyword searches and their 62 word senses provides empirical indications that our method can improve upon existing keyword based search engines. Our method learns the visual word sense models in a totally unsupervised manner, effectively filters out irrelevant images, and is able to mine the long tail of image search.

1 Introduction

With increasing ease of media creation and the widespread availability of image search engines, there is growing interest on how to tap into the large repository of internet images. For example, in supervised object and scene categorization, a large dataset of labeled images is usually required. However, constructing such databases of high precision images is still challenging because image search engines are limited by poor precision of the returned images. For example, [17] reports a precision of only 32% with Google Image Search. Another factor for the noisy results of image search is the inherent ambiguity in the user query keyword. For example, the keyword *apple* can refer to the fruit, the company or the computer product.

BMVC 2009 doi:10.5244/C.23.67

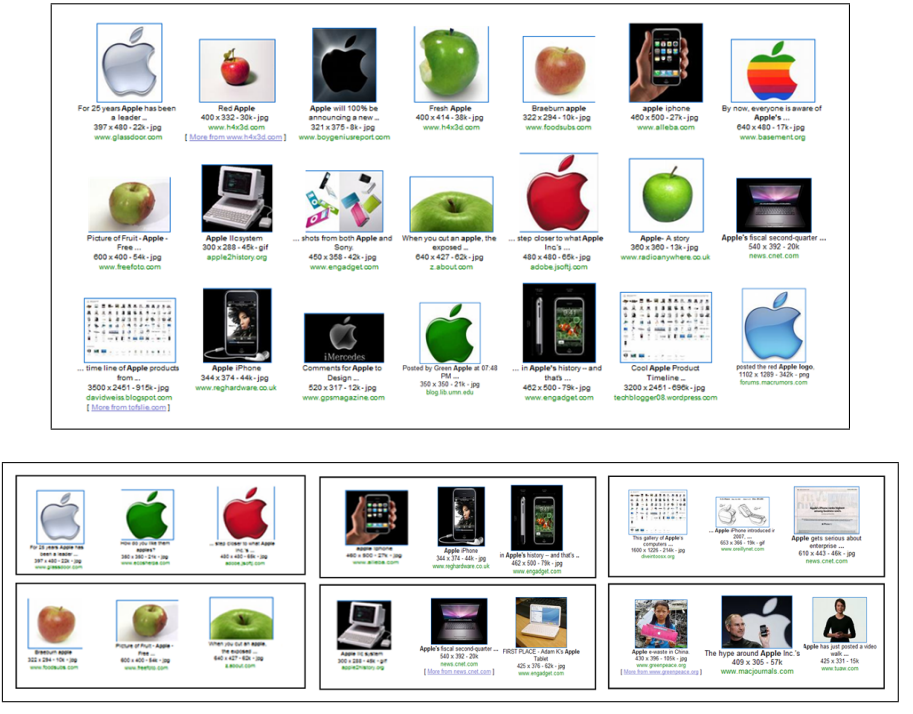

Figure 1: Top: Top-ranked apple images from Google Image Search; Bottom: 6 possible visual senses: *logo, fruit, iphone, iMac, drawings, people*

This is commonly referred to as word sense disambiguation (WSD). Similarly, we can also consider visual sense disambiguation (VSD) of a word. As shown in figure 1, the visual senses of apple correspond to the image clusters of the returned images of apple. It is important to note the subtle differences. WSD is a top-down process arising from ambiguities in our natural language. The word senses of a word are robust and relatively static, and we can easily look them up from a dictionary resource such as WORDNET [2] or Wikipedia [3, 14]. On the other hand, VSD is a data-driven problem that is specific to the image collection. For example, in figure 1, the *drawings* visual sense of apple is an artifact of clustering on apple images. Clearly, *drawings* is not a word sense of apple.

If the user was looking for the Apple Company, the images returned corresponding to other senses, valid as they may be, will be noise to him. In this paper, we address the problem of unsupervised learning of object classifiers for polysemous user keywords. We propose an unsupervised method that resolves user search ambiguity by allowing sense specific diversity in search results. The only input to our algorithm is a list of keywords. We take a three-step approach. First, we determine a list of possible senses of the keyword using the electronic dictionary Wikipedia, to retrieve sense-specific images. Second, we learn a topic model on the visual senses of the keyword, using the images returned from the original polysemous keyword. Thirdly, we learn a visual classifier for each word sense, by incorporating the visual sense topic model. For each Wiki-sense of the keyword, we use the learnt sense-specific model to cluster and re-rank the polysemous images from the original query keyword into its specific senses.

Figure 2: Google search results on "Mouse computing". While results are more homogeneous than that of Mouse, polysemy clearly remains an issue.

Given that we are also retrieving images using sense-specific queries, an obvious approach is to bootstrap sense-specific classifiers from these images. We shall call this method Sense-Specific SVM. While we expect that returned images will be more homogeneous as a result of sense specification, polysemy will still be a problem in learning the sense-specific SVM (figure 2). Nonetheless, it serves as the baseline to our main approach, which overcomes these issues by incorporating a latent model of the visual sense of the original keyword. The key idea is that in these images, there is a rich source of information about the various senses (visual or word) of the word, of which Wikipedia merely provides a subset list denoting the primary senses that are more commonly used. These visual senses capture the salient visual characteristics of images associated with the keyword, and offer a more robust visual model than learning on just the Wiki-sense-specific images. Each Wiki-word sense is then represented in the latent space of hidden visual topics for the polysemous keyword.

2 Method

Figure 3 shows an illustration of our proposed framework. Our method consists of four steps: (1) using Wikipedia to disambiguate the word-sense (Wiki-sense), and issuing sense-specific queries to retrieve sense-specific images, (2) discovering latent visual sense topics in polysemous keyword images, (3) learning probabilistic models of Wiki-senses in the visual sense latent space, (4) using the Wiki-sense models to construct sense-specific image classifiers. We now describe each step in detail.

2.1 Wikipedia based WSD

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteers. Because of the open and collaborative environment the quality and quantity is well trusted. As a large-scale repository of structured knowledge, Wikipedia is a valuable resource for a diverse array of research activities [14]. One structure of particular interest to this paper is the *disambiguation* page. It gives a detailed list of possible senses (meanings) of ambiguous words by attaching the expression *(disambiguation)* to the name of the ambiguous entity, e.g., *bar_(disambiguation)*, which identifies the disambiguation page of the entity bar. The advantage of this disambiguation page is that it not only gives the word senses in a structured categorized way, but also links up pages that have further details. All these advantages motivate us to use it for disambiguating

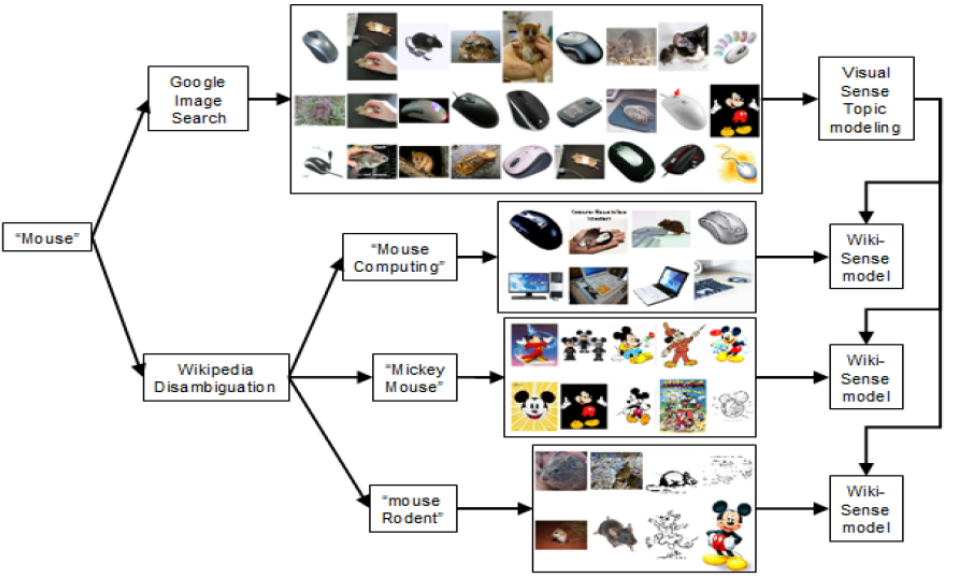

Figure 3: Example process flow. There are 3 Wiki-word senses for mouse: "mouse computing", "Mickey mouse" and "mouse rodent". A visual sense topic model is learnt on the mouse images and incorporated into the wiki-word-sense classifier model.

keyword based image search. Given a query keyword we may send a query to Wikipedia to extract different (senses) meanings of the word automatically. However, for some keywords, Wikipedia provides many spurious and trivial disambiguation. Hence, in our experiments we have handpicked the more salient word senses, limiting them to not more than 10 per keyword. While this may seem contrived, we do anticipate that this step can be automated. We also point out that this manual word sense selection step does not compromise the evaluation on the effectiveness of our approach. Once we have a list of word sense strings, it is used to pull images from the web using Google Image Search.

2.2 Visual Sense Topic Modeling

Drawing parallel from text literature, the idea behind topic modeling is to model a document as a mixture of latent topics where the topics are distributions over words (terms) in the documents. Adapting this model for visual data, we follow the recent trend of using visual words for image representation. We first locate salient keypoints in an image, and a high-dimensional SIFT descriptor [12] is computed for the region around that keypoint. These descriptors are then quantized and clustered into a vocabulary of visual words using the standard K-means algorithm. An image is a document that is a mixture of several visual topics. These visual topics portray the latent semantic content of the image. Latent senses of images can be determined using existing topic modeling techniques. We use Latent Dirichlet Allocation (LDA) [5]. An image d with W visual words is assumed to be generated by the following generative process. For each latent topic z_i $(i = 1..K)$, the parameter ϕ_i of a multinomial distribution over the visual words is sampled from a Dirichlet prior with parameter β. Then, for each image d the parameter θ_d of a multinomial distribution over topics is sampled

from a Dirichlet prior with parameter α. Finally, we choose a topic z_j from θ_d and choose a visual word w_j $(j = 1..W)$ from ϕ_{z_j}. The probability of generating d is given by:

$$P(w_1, w_2, \ldots, w_W | \phi, \theta_d) = \prod_{j=1}^{W} \sum_{i=1}^{K} P(w_j | z = z_i, \phi) P(z = z_i | \theta_d) \quad (1)$$

Because the original query is polysemous, these returned image collection will comprise of images from various senses. We learn a latent space of K visual topics in these polysemous images, and treat each of these visual topics as a visual sense underlying the images. Hence, we uncover a latent visual representation of the various senses for each image.

2.3 Wiki-Dictionary Sense Model

Given a query keyword P, we treat the images retrieved by each Wiki-sense-specific image queries as the (primary) senses S_i of P, $i = 1, 2, \ldots, N_P$, N_P is the number of Wiki-senses of P. For example, in figure 3, the "mouse" keyword has 3 wiki-word-senses: "mouse computing", "Mickey mouse" and "mouse rodent". We define the likelihood of the i^{th} sense S_i given the topic $z = z_j$ as:

$$P(S_i | z = z_j) = \frac{1}{|S_i|} \sum_{a \in S_i} P(a | z = z_j) = \frac{1}{|S_i|} \sum_{a \in S_i} KL(W_a, Z_j) \quad (2)$$

where W_a is the visual word distribution of image a, Z_j is the visual word distribution of topic z_j, and $KL(\cdot)$ is the Kullback Leibler divergence between the two. For an image d, the model computes the probability of d belonging to the i^{th} sense S_i as:

$$P(S_i | d) = \sum_{j=1}^{K} P(S_i | z = z_j) P(z = z_j | d) \quad (3)$$

2.4 Re-ranking

Equation 3 assigns visual sense probabilities to an image according to how similar it is to the sense-specific images. $P(S_i|d)$ provides a way to re-rank the images in the original polysemous order. Images belonging to some sibling sense are given lower probabilities and pushed to the back of the rank list. We call this method VSD-LDA. VSD-LDA extends the method in [16]. The main difference is that [16] is a text-based method, where latent topic discovery is performed on a collection of web text crawled using the polysemous keyword. In contrast, we propose here a VSD-based visual domain topic modelling framework.

3 Dataset and Evaluation

To train and evaluate the algorithms, we define a set of 10 polysemous keywords. We focus on object keywords. For each keyword, we manually select a few senses from Wikipedia. Table 1 shows the 10 keywords and their respective senses. There are a total of 62 senses. We create a total of 72 image datasets by issuing queries to Google Image Search: (a) 10 sets of image search results by separately issuing each of the 10 keywords as search query, and (b) 62 image search results by issuing each of the 62 sense-specific search terms as search query. Hence, for each keyword, we have a total of $S+1$ datasets, where S is the number of senses

of the keyword. All images were automatically downloaded by following the image URLs on the Google image result index page. For each keyword, we retrieve about 500 images, totalling 5013 images. All images were labelled by 3 human annotators. For each image of a particular keyword, the labellers were given the list of the word senses, and they were asked to choose only one (dominant) sense. An extra "None" label is defined on images where the object was too small or occluded. For the sense-specific image datasets, we retrieve about 200 images for each sense. This makes a total of 11951 images for all 62 senses.

Keyword	Wikipedia word senses
bank	Bank finance, Bank building, River Bank, Bank sea floor, Blood bank, Gene bank, Piggy bank
bar	Bar rod, Bar pole, Dessert Bar, Bar Law, Candy Bar, Barbell
bass	Bass Drum, Bass guitar, Bass Flute, Bass Fish, Bass Rock, Bass Strait, Bass Instrument, Acoustic Bass Guitar
mouse	Mouse computing, Mickey Mouse, Mouse Rodent
plant	Tree, chemical plant, implant, herb, bush, grass, vines, ferns, mosses, forest
speaker	Speaker government, loudspeaker, Orator, computer speaker
temple	temple anatomy, hindu temple, mount temple, temple mount
tiger	Bengal Tiger, Tiger Woods, Tiger Shark, Tiger Snake, Tiger Beer, Tiger Mac OS, Tiger Tank, Tony the Tiger, White Tiger, Detroit Tiger
watch	wrist watch, guard, watch tower, wall clock, pocket watch
window	window house, computer window, windows operating system, window snyder, window blind

Table 1: Keywords and their Wiki-senses used in our VSD experiments

We use a local region approach to represent images. Local regions are extracted by Difference of Gaussian (DoG) [12] and and Maximally stable extremal regions (MSER) [13]. These methods can be viewed as complementary to each other, sampling blob-like regions and high contrast image structures [15]. For each region, the SIFT and SURF [4] descriptors are then computed using the Camellia [1] and the VLFeat [18] toolkit. K-means clustering is used to compute the visual codebook.

We now evaluate how well the 2 algorithms (sense-specific SVM, VSD-LDA) can re-rank the polysemous keyword image dataset. For each keyword, the sense-specific SVMs are trained on the sense-specific image datasets. Images in the keyword dataset are then re-ranked by moving the negatively-classified images down to the last rank. For VSD-LDA, we train a separate LDA model for on the keyword image dataset, setting the number of topics K to twice the number of keyword senses. This number is based on the intuition that we expect that there are more visual topics spanning the polysemous image data-sets that that specified by Wikipedia. We also expect that some of these visual topics will align with each of the Wikipedia senses. In fact, multiple topics can represent the same sense. We also note that this number can also be set automatically by cross-validation. We then compute $P(S_i|d)$ for image d in the keyword dataset, using Equation 3, and rank the corresponding images according to the probability of each sense S.

Following [16], we evaluate the retrieval performance using receiver operating characteristic (ROC) by thresholding $P(S|d)$ for every sense S of a keyword. Due to space constrain,

figure 4 shows the ROCs for the first primary sense of each keyword. The dark-blue lines are the ROCs for the original Google search ranks. The cyan lines are the ROCs using the sense-specific SVMs to re-rank the Google search image order. The red lines are the ROCs obtained by our VSD-LDA method. Table 2 shows the total Area Under Curve (AUC) for all senses of each keyword. The full ROC plots of all 62 wiki-senses can be accessed at http://goal.i2r.a-star.edu.sg/Image/sense_models/sense.html. The results demonstrate that our VSD visual sense models can retrieve far more positive class images than the original search engine order. This can be used to diversify image search results.

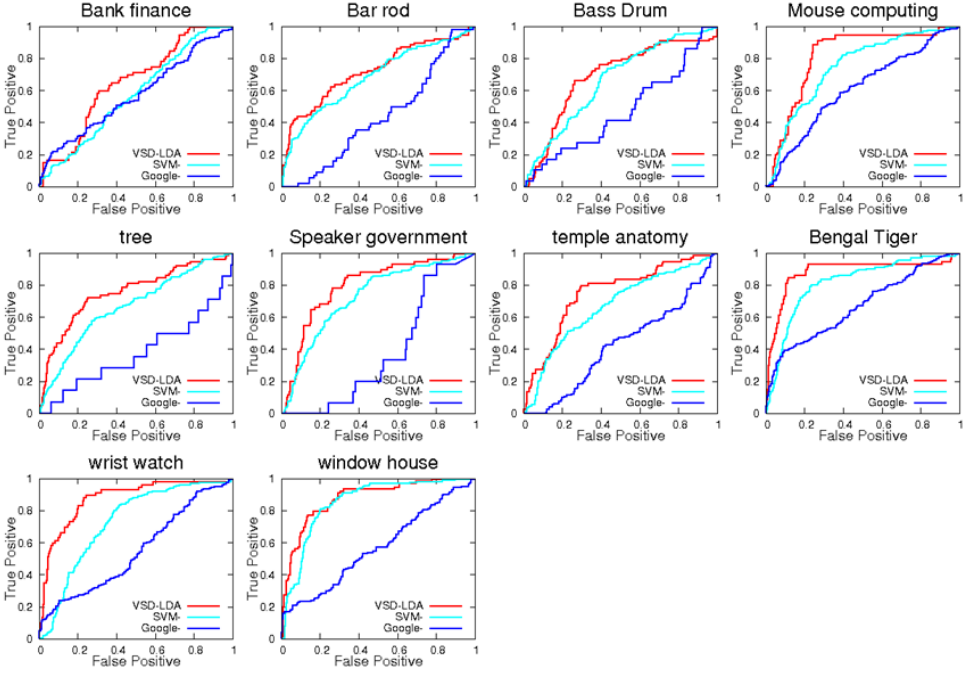

Figure 4: ROC plots of the first primary sense of the 10 polysemous keywords

4 Related Works

Several methods have been proposed to learn object models from web images, e.g. classifier bootstrapping on labeled images [11] and image clustering into coherent components [7] However, they rely on labeled images for initialization or cluster selection. In contrast, Schroff *et al*. [17] describes unsupervised object categorization by learning on the top-ranked images returned by a search engine, by assuming that they are relevant for the category. However, for polysemous keywords, this is a very weak assumption.

The application of topic modeling to the visual domain has also received much attention. LDA was used in [8] to discover visual categories like cars, people, cows, etc. An LDA variant that combines spatial constrains into topic modeling is reported in [9] to better retrieve particular objects. Chum *et al*. [6] described a generative image retrieval model on a large

Keyword	Google rank	SVM	VSD-LDA
bank	3.18402	5.1769	5.34275
bar	2.48154	3.94088	4.20582
bass	3.71719	5.47068	5.84983
mouse	1.54961	2.21711	2.41246
plant	5.03565	7.45235	7.96624
speaker	2.07675	2.82301	3.03133
temple	2.1008	2.90866	3.01708
tiger	3.87138	7.72587	8.11975
watch	2.06972	3.7881	4.06134
window	1.93045	4.05622	4.19338
Total-AUC	28.0172	45.5596	48.1998

Table 2: Area Under Curve (AUC) of all senses of each keyword

database. They adopt bag-of-visual-words indexing with a novel query expansion extension, where a number of the highly ranked documents from the original query are reissued as new queries. To contol term expansion, they apply strong spatial constraints between the query image and each result and learn a latent feature model on the verified images.

A related problem is the selection of images that are highly relevant but yet are diversified. A graph-based approach is used in [10] that iteratively assigns a numerical weight to each image based on its relative importance to other images.

Existing image search engines also allow sense specific search by recommending alternate queries based on query log information. However, this approach is not as effective as dictionary based query formulations, since they are more carefully chosen and collaboratively verified.

5 Conclusions

We introduced a method that combines a dictionary and the visual content of web images to disambiguate keyword-based image search. We introduce the concept of the visual sense of a word. Given a polysemous keyword, we propose learning a latent model of the visual sense of images of this keyword. The key idea is that in these images, there is a rich source of information about the various senses (visual and word) of the word. Compared to dictionary word senses (from Wikipedia), these visual senses capture the salient visual characteristics of images associated with the keyword, and offer a more robust visual model than learning on just the Wiki-sense-specific images. Each Wiki-word sense is then represented in the latent space of hidden visual topics for the polysemous keyword. Our approach is mainly inspired by the work in [16] but extends its text-based topic modeling framework to the visual domain. It capitalizes on the large amount of unlabeled images available through keyword image search in conjunction with the dictionary entries to learn a generative model of sense. We evaluated our model on a large dataset of over 17K images consisting of search results for 10 polysemous words. On the retrieval task, our VSD model improved on both the baseline search engine precision and the sense-specific SVMs by re-ranking the images according to sense probability.

References

[1] Camellia image processing library. http://camellia.sourceforge.net.

[2] Wordnet, a lexical database for the english language. http://wordnet.princeton.edu.

[3] Wikipedia:disambiguation. http://en.wikipedia.org/wiki/Dab_page.

[4] H. Bay, T. Tuytelaars, and L. Gool. Surf: Speeded up robust features. In *Proc. of European Conference on Computer Vision*, pages 404–417, 2006.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

[7] R. Fergus, F. Li, P. Perona, and A. Zisserman. Learning object categories from googles image search. In *Proc. ICCV*, 2005.

[8] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *Proc. CVPR*, 2008.

[9] J. Sivic J. Philbin and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *Proc. BMVC*, 2008.

[10] Y. Jing and S. Baluja. Pagerank for product image search. In *Proc. WWW*, 2008.

[11] J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic object picture collection via incremental model learning. In *Proc. CVPR*, 2007.

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

[13] J. Matas, O. Chum, M. Urba, and T. Pajdla. Robust wide baseline extremal regions. In *Proc. of British Machine Vision Conference*, pages 384–396, 2002.

[14] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proc. NAACL*, 2007.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. In *International Journal of Computer Vision*, volume 65.

[16] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proc. NIPS*, 2008.

[17] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proc. ICCV*, 2007.

[18] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms, 2008. http://www.vlfeat.org/.