Visual Categorization with Negative Examples for Free

Xirong Li and Cees G. M. Snoek
Intelligent Systems Lab Amsterdam, University of Amsterdam
Science Park 107, 1098 XG, Amsterdam, The Netherlands
{x.li,cgmsnoek}@uva.nl

ABSTRACT

Automatic visual categorization is critically dependent on labeled examples for supervised learning. As an alternative to traditional expert labeling, social-tagged multimedia is becoming a novel yet subjective and inaccurate source of learning examples. Different from existing work focusing on collecting positive examples, we study in this paper the potential of substituting social tagging for expert labeling for creating negative examples. We present an empirical study using 6.5 million Flickr photos as a source of social tagging. Our experiments on the PASCAL VOC challenge 2008 show that with a relative loss of only 4.3% in terms of mean average precision, expert-labeled negative examples can be completely replaced by social-tagged negative examples for consumer photo categorization.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*; H.2.4 [**Database Management**]: Multimedia databases

General Terms

Algorithms, Measurement, Experimentation

Keywords

Visual categorization, negative examples, social tagging

1. INTRODUCTION

To help people organize and access the increasing amounts of diverse multimedia data, automatic visual categorization is an important prerequisite. Nonetheless, the categorization accuracy is critically dependent on labeled examples used for training classifiers. Traditionally, expert labeling is required to label examples. The recent advent of social multimedia tagging, i.e., assigning tags to images and videos by common users, is creating a considerable amount of loosely labeled visual data on the web. For instance, online photo sharing platforms such as Flickr and Facebook are hosting billions of user-uploaded images. Considering that expert labeling is expensive and time-consuming while social-tagged multimedia is widely accessible for free, an interesting question here is, can social tagging substitute expert labeling for creating training examples?

In a strike towards replacing expert labeling with social tagging, most existing work [2–4, 10, 12] focus on automated approaches to collecting positive examples for a given concept, e.g., cow. A common strategy among these approaches is to take tag-based visual search results as a starting point and refine the results by online learning afterwards. Since social tagging is known to be subjective and inaccurate, the tag-based search results might be unsatisfactory. Automatically obtaining positive examples with a sufficient accuracy for learning classifiers is still an open research problem [5]. As an alternative, some try to encourage common users to label examples by game competitions [11]. In contrast to the intensive effort devoted to the positive examples, the importance of negative examples is overlooked. Since negative examples of a concept belong to many other concepts, they often demand more manual labeling. In the PASCAL VOC challenge 2008 for instance, over 90% of the annotation efforts contributed by 17 experts are spent on labeling the negative examples [1]. In [7, 13] the authors try to automatically generate negative examples by random sampling for video retrieval. They show that adding the pseudo negatives obtains a better retrieval model when compared to solely using query images. To the best of our knowledge, however, the problem of how to leverage negative examples from social tagging for visual categorization remains largely unexplored and its importance underestimated.

In this paper, we study to what extent social tagging substitutes expert labeling for creating *negative* examples. Image classification experiments on the VOC 2008 development set and 6.5 million social-tagged images verify our idea.

2. THREE LEARNING SCENARIOS

According to the availability of negative examples, we divide learning scenarios for visual categorization into three types. That is, one-class learning without negative examples, two-class learning with expert labeling, and two-class learning with expert-labeled positive examples and social-tagged negative examples, as depicted in Fig. 1. Since the

This work was supported by the EC-FP6 VIDI-Video project and the STW SEARCHER project.

MM'09, October 19–24, 2009, Beijing, China.

negative data in the last scenario, see Fig. 1(c), are automatically collected without manual assessment, we term the scenario as learning with negative examples for free. Here we choose the Support Vector Machine (SVM) as a supervised learner, which has proven to be a solid choice [1, 7].

2.1 Scenario 1: One-class learning

Given expert-labeled positive examples only, one-class learning assumes that the examples tend to have similar visual patterns and hence the corresponding data points in the visual feature space stay close to each other. Then, the learning strategy as depicted in Fig. 1(a) is to construct a hypersphere in the feature space to include most of the positive points while at the same time, keeping the hypershpere as compact as possible [8].

2.2 Scenario 2: Two-class learning

Given both positive and negative examples by expert labeling, two-class learning tries to find a decision boundary which separates most of the positive examples from most of the negative examples while at the same time, keeping the boundary as simple as possible, as shown in Fig. 1(b). Intuitively, this learning scenario will do better visual categorization than scenario 1 as more information, in the form of negative examples, are taken into account.

2.3 Scenario 3: "Free" negative examples

As illustrated in Fig. 1(c), this learning scenario is similar to scenario 2, but with negative examples created by social tagging for free. To select a negative set from a social-tagged image collection, we investigate two strategies, one is random sampling and the other is random sampling after anti-synonym filtering which we will describe in Section 3.1. If this scenario is comparable to scenario 2 in the sense of the categorization accuracy, expert labeling is replaced by social tagging for the construction of negative examples.

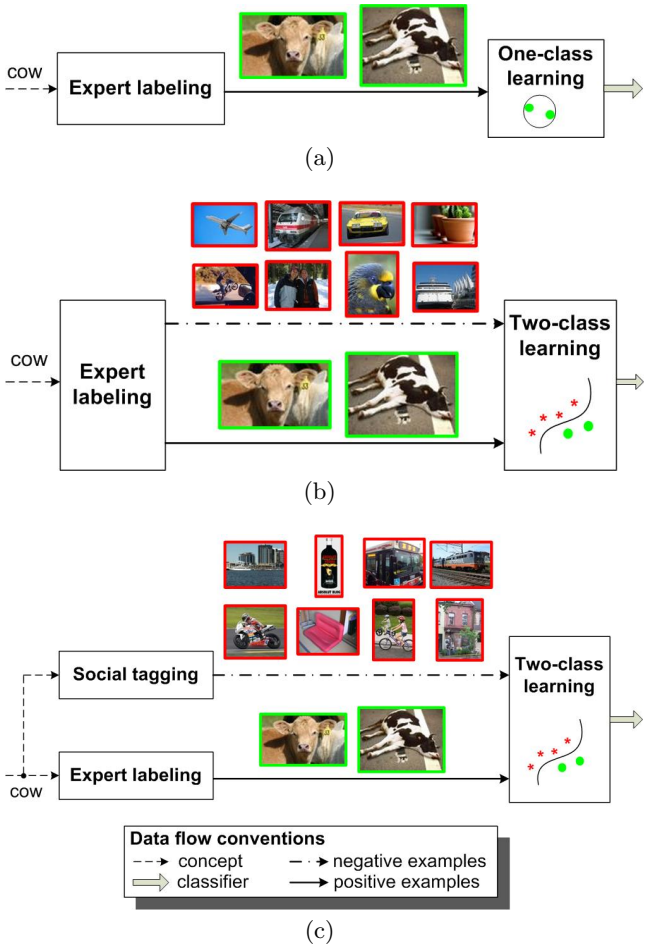

Figure 1: We divide learning scenarios for visual categorization into three types according to the availability of negative examples: a) one-class learning without negative examples, b) two-class learning, and c) two-class learning with negative examples for free.

3. EMPIRICAL STUDY

3.1 Experiments

First, to confirm the intuition that two-class learning is better than one-class learning, we compare scenario 1 and 2. Then, to study to what extent social tagging substitutes expert labeling for creating negative examples, we compare scenario 2 and 3.

• **Experiment 1: Two-class versus one-class**. We compare a one-class SVM, trained only on expert-labeled positive examples, and a two-class SVM, trained on expert-labeled positive and negative examples.

• **Experiment 2: "Free" negative examples**. For both scenarios 2 and 3, we use the same expert-labeled positive examples. While in scenario 3, we fully replace expert-labeled negative examples with examples automatically selected from a social-tagged image collection. We describe a negative example selection strategy as follows. Given a concept, say aeroplane, multiple words may refer to the same concept, e.g., plane and airplane for aeroplane. Moreover, the diversity of social tagging extends the range of synonyms. For example, we observe that the two tags, boeing and airbus, are likely to indicate images of an aeroplane also. Bearing these in mind, we introduce a selection strategy called random sampling after anti-synonym filtering. That is, we first identify a set of possible synonyms of the concept by computing tag co-occurrence within a large social-tagged image collection. Tags whose co-occurrence with the concept exceeds a certain threshold are treated as synonyms of the concept. Further, by removing from the entire collection images labeled with the concept or tags from its synonym set, we obtain a candidate negative set. Here we set the co-occurrence threshold to 1000. Though the rule is strict, we still obtain a considerable amount of negative data, since the collection is very large. Finally, we do random sampling on the candidate set to select a subset of 5000 examples for the subsequent training. We repeat the sampling procedure multiple times and observe that the overall performance is relatively stable among different runs.

• **Experiment 3: Comparing two selection strategies**. Finally, we compare two strategies for negative example selection. One is the proposed random sampling after anti-synonym filtering. The other is fully random sampling, i.e., randomly selecting a number of examples from the social-tagged image collection as negative examples. For both strategies, we choose 1000 and 5000 negative examples

Table 1: Statistics of the 20 concepts in our 6.5 million Flickr collection. We consider a tag as a synonym of a concept if their co-occurrence exceeds 1000. By removing images labeled with the concept or its synonyms, we get pseudo negative examples.

Concept	Frequency	Top 3 co-occurred tags	Pseudo negatives
aeroplane	72,648	plane, airplane, aircraft	4,447,540
bicycle	504,470	bike, cycling, race	784,377
bird	703,781	nature, birds, animal	473,928
boat	685,821	water, sea, river	137,138
bottle	106,094	beer, glass, wine	3,911,373
bus	393,163	london, buses, tour	706,609
car	485,795	auto, cars, show	144,603
cat	899,939	cats, kitten, kitty	81,526
chair	177,483	table, red, furniture	1,113,479
cow	171,919	farm, animal, cows	1,359,959
diningtable	1,144	table, furniture, diningroom	6,575,454
dog	890,403	puppy, dogs, pet	59,524
horse	567,960	horses, cheval, caballo	314,708
motorbike	104,623	motorcycle, bike, moto	3,592,970
person	122,233	people, portrait, woman	1,920,689
pottedplant	2,522	flower, nature, macro	6,314,384
sheep	165,876	farm, animals, lamb	1,452,646
sofa	44,533	cough, furniture, art	3,828,459
train	677,986	railroad, station, railway	340,701
tvmonitor	27	wow, tvmonitorcombinatie, stage	6,576,572

for training, respectively.

3.2 Data preparation

• **Benchmark**. We adopt the PASCAL VOC 2008 development set as our benchmark data [1], which is collected from Flickr with expert verification. The set consists of two predefined subsets, one for training and the other for validation with 2111 and 2221 images, respectively. For all experiments, we learn SVM models on the training set and test the models on the validation set. There are 20 visual concepts in total, as listed in Table 1.

• **Social-tagged image collection**. We collected 6.5 million Flickr images as follows. For each of the 20 concepts, we download images tagged with that concept and uploaded between Jan. 2004 and Dec. 2008, with a maximum of 5000 downloads per week. We remove 1588 images from the downloaded dataset so that the dataset and the benchmark have no overlap. We report the statistics of the 20 concepts in the entire collection in Table 1.

3.3 Bag-of-words based image categorization

• **Image representation**. Since representing images by bag-of-words is well recognized as the state-of-the-art feature for visual categorization [1], we follow this convention. In particular, we adopt dense sampling for point detection and SIFT [6] for point description, using a recently developed fast implementation of the dense-SIFT [9]. We create a codebook of 4000 bins by running K-means clustering on the VOC dataset. By mapping the SIFT features to the bins, each image is represented by a 4000-d bag-of-words feature.

• **Image categorization**. We follow the definition of the VOC categorization task. For each concept we predict the presence of that concept in a test image with a real-valued confidence. All test images are then ranked according to their confidences in descending order. We train a one-class SVM for scenario 1 and a two-class SVM for scenario 2 and 3 by three-fold cross validation. For all experiments, we adopt the χ^2 kernel, one of the best kernels for visual categorization [1].

• **Evaluation criteria**. To assess the classification performance, we use average precision, a good combination of precision and recall [1]. To evaluate the overall performance, we use mean average precision (MAP), the mean value of average precision scores over all concepts.

• **Random baseline**. Since frequent concepts such as people tend to have higher average precision scores than rare concepts like cow, we also report a random baseline for the ease of analysis, which is calculated as follows. For each concept, we randomly rank the validation set and calculate average precision. We run the process 100 times and take as the random baseline an averaged score over the 100 runs.

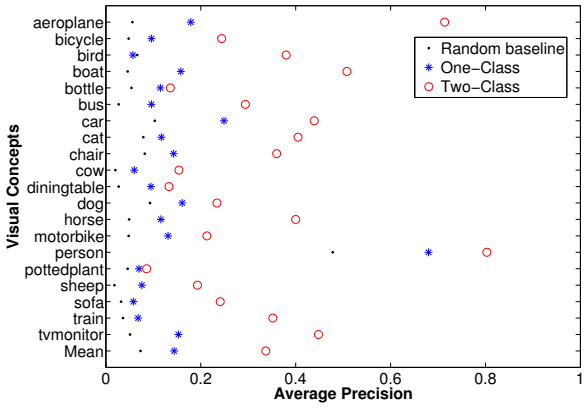

Figure 2: Experiment 1: Two-class versus one-class. The results show that negative examples are helpful for visual categorization.

4. RESULTS

Experiment 1: Two-class versus one-class. As shown in Fig. 2, the two-class learning surpasses the one-class learning, with an MAP of 0.337 and 0.144, respectively. Modeling visual concepts by one-class SVM is difficult due to the fact that examples belonging to different concepts may have visual patterns in common. For instance, the two concepts car and motorbike often have similar visual context such as street. Hence, examples belonging to different classes may also stay close to each other in the feature space, meaning these examples are likely included in the same hypersphere. Negative examples are thus helpful to distinguish such ambiguous cases.

Experiment 2: "Free" negative examples. As shown in Fig. 3, social tagging fully substitutes expert labeling for creating negative examples, with a relative loss of only 4.3% in terms of MAP. Recalling that over 90% of the VOC 2008 expert labeling effort is spent on negative examples, we believe the small loss in the classification accuracy is worthy of the immense annotation effort saved. Note that the expert-labeled negative set is closed in the sense that the number of visual classes is fixed. In contrast, our social-tagged negative set is more diverse as we sample from an open collection. Therefore, on one hand, the performance of social negatives on the closed validation set degrades to some extent for 14 concepts. On the other hand, we expect classifiers trained on a more diverse set to have a better generalization ability. Due to the limit of the validation set, however, this advantage is not obvious in the current experiments. As a future work, it would be interesting to reveal the generalization issue, say by diversifying the validation set.

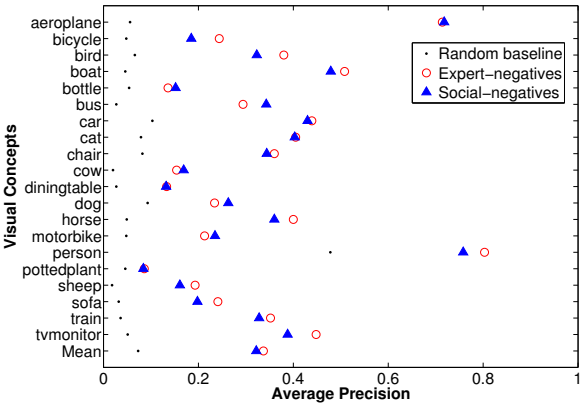

Figure 3: Experiment 2: "Free" negative examples. By fixing the positive examples, we compare two-class learning with expert-labeled negative examples (denoted as Expert-negatives) and learning with 5000 social-tagged negative examples (denoted as Social-negatives). With a relative loss of only 4.3% in terms of mean average precision, social tagging fully substitutes expert labeling for creating negative examples.

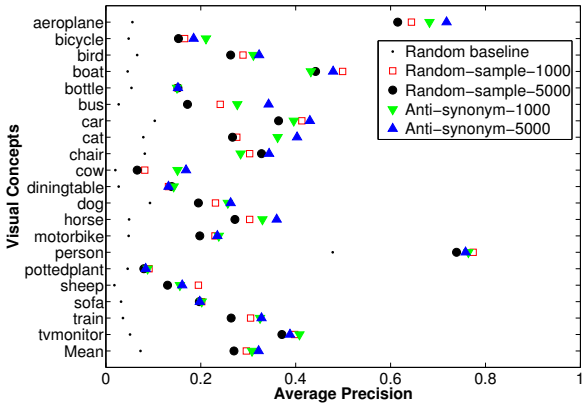

Figure 4: Experiment 3: Comparing the two strategies for negative example selection. For both random sampling and random sampling after anti-synonym filtering, we choose 1000 and 5000 negative examples for training, denoted as Random-sample-1000, Random-sample-5000, and Anti-synonym-1000, Anti-synonym-5000, respectively. Using the anti-synonym filtering, we create a better candidate set for negative example selection.

Experiment 3: Comparing the two strategies for negative example selection. As shown in Fig. 4, with anti-synonym filtering, we achieve better categorization performance. Interestingly, we observe that as more negative examples are selected, from 1000 to 5000, the performance of the random sampling result degenerates from 0.296 to 0.270 in terms of MAP. By contrast, the performance of random sampling after anti-synonym filtering improves by 4.5%, from 0.308 to 0.322 in terms of MAP. This result demonstrates that with the anti-synonym filtering algorithm, we create a better candidate set for negative example selection.

5. CONCLUSIONS

This work is an attempt towards substituting social tagging for expert labeling for deriving visual classifiers. In particular, we focus on replacing expert labeling with social tagging for creating *negative* examples. We discuss a new supervised learning scenario in which the negative examples are automatically collected from social-tagged images for free. As a main contribution of this work, we empirically show that compared to a traditional two-class learning with expert-labeled examples, learning with the "free" negative examples and expert-labeled positive examples achieves a comparable classification accuracy. Using 6.5 million social-tagged images as a source of "free" negative examples, our experiments on the PASCAL VOC 2008 development set demonstrate that with a relative loss of only 4.3% in terms of mean average precision, social tagging can fully substitute expert labeling for creating negative examples for consumer photo categorization.

6. REFERENCES

[1] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008.

[2] X.-S. Hua and G.-J. Qi. Online multi-label active annotation: towards large-scale content-based video search. In *ACM MM*, pages 141–150, 2008.

[3] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers In *ACM MIR*, pages 249–258, 2006.

[4] L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: automatic online picture collection via incremental model learning. In *CVPR*, pages 1–8, 2007.

[5] X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *ACM MIR*, pages 180–187, 2008.

[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[7] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM MM*, pages 991–1000, 2007.

[8] D. Tax and R. Duin. Uniform object generation for optimizing one-class classifiers. *JMLR*, 2(2):155–173, 2002.

[9] J. Uijlings, A. Smeulders, and R. Scha. Real-time bag-of-words, approximately. In *CIVR*, 2009.

[10] A. Ulges, M. Koch, C. Schulze, and T. Breuel. Learning TRECVID'08 high-level features from YouTube. In *TRECVID*, 2008.

[11] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *SIGCHI*, pages 319–326, 2004.

[12] K. Yanai and K. Barnard. Probabilistic web image gathering. In *ACM MIR*, pages 57–64, 2005.

[13] R. Yang, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, pages 343–346, 2003.