Crowdsearching Training Sets for Image Classification

Sami Abduljalil Abdulhak(✉), Walter Riviera, and Marco Cristani

Department of Computer Science, Cá Vignal 2, Verona, Italy
{sami.naji,walter.riviera,marco.cristani}@univr.it

Abstract. The success of an object classifier depends strongly on its training set, but this fact seems to be generally neglected in the computer vision community, which focuses primarily on the construction of descriptive features and the design of fast and effective learning mechanisms. Furthermore, collecting training sets is a very expensive step, which needs a considerable amount of manpower for selecting the most representative samples for an object class. In this paper, we face this problem, following the very recent trend of automatizing the collection of training images for image classification: in particular, here we exploit a source of information never considered so far for this purpose, that is the textual tags. Textual tags are usually attached by the crowd to the images of social platforms like Flickr, associating the visual content to explicit semantics, which unfortunately is noisy in many cases. Our approach leverages this shared knowledge, and collects images spanning the visual variance of an object class, removing at the same time the noise by different filtering query expansion techniques. Comparative results promote our method, which is capable to automatically generate in few minutes a training dataset leading to an 81.41% of average precision on the PASCAL VOC 2012 dataset.

Keywords: Image classification · Training sets · Crowdsearching · CNN · SVM

1 Introduction

Generally underestimated in favor of more appealing themes like feature design and model learning, the challenge of building effective training sets for object recognition is instead very important; an ideal training set should represent the entire visual variation of an object class, capturing all of its facets in an ensemble of representative samples. In image classification, this means to have a set of pictures which portray an object under different poses, illumination conditions, occluded or not, but also spanning all of its semantics variability. Think for example at a classifier aimed at recognizing a dog in an image: a dog could be an husky dog, a fox-terrier, bulldog, which can be running, barking, sitting etc.. All of this should be present in the pool of training images, to ensure a proper generalization.

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part I, LNCS 9279, pp. 192–202, 2015.
DOI: 10.1007/978-3-319-23231-7_18

So far, the issue of building image datasets for training classifiers was committed to scientists [7], or, more recently, to Internet users through crowdfunding platforms like Amazon Mechanical Turk [3]. In the latter case, scientists still have to supervise the image collection, as noisy samples could be erroneously captured. All of these settings require a considerable effort, both in time and monetary terms; for this reasons, datasets are nowadays focused on few classes, depicting generic objects (dogs, cars, aeroplane etc.). A notable exception is Imagenet [5], which mirrors the taxonomy of Wordnet [15], considering circa 21841 different object classes. Anyway, many of them contain very few images ("minibike, motorbike", "tanker plane") making their use as training set unfeasible.

Another research direction looks for automatic strategies to craft image datasets [12], which obviously are faster than the human intervention, but introduce a huge number of noisy elements that cannot be removed by anyone but the human. Therefore, the challenge here is to capture as less as possible false positive images; the idea is to exploit image search engines as Google Image Search, feeding them with keywords [8] or n-grams containing the entity of interest [1], keeping the images they report and in some cases [12] perform some post processing for removing noisy pictures.

All of these automatics approaches rely on indexing methods, as the one of Google Image Search, which are not open source, meaning that one step of the collection consists in a black box, where the cues used to gather images are hidden and changing over time (due to advancements in the search engine). As a consequence, the performance of such approaches could vary considerably, with no repeatability guarantees.

In this paper, we bypass this problem, focusing on an automatic image training set collection strategy which uses social platforms for gathering the pictures (in this work Flickr[1]), exploiting the textual tags usually associated to them by the users. In this way, we produce a genuine crowdsearching algorithm where the cues to extract images are visible and not hidden. In particular, the strategy is based on the interesting mechanism of the *query expansion* [14], which first builds a statistics of the more frequent tags associated to some potentially interesting images (see for example the images resulting from the textual query "dog"); on this statistics, which mirrors the common thinking of the crowd, different filters are applied, pruning away noisy tags. The remaining tags are instead used to retrieve a second set of images, the ones which will finally form the training set.

Our strategy is simpler than the previous approaches in the literature, giving better results on diverse image classification datasets, producing a training set which is absolutely comparable to many cutting edge datasets as the PASCAL VOC and Imagenet. Most notably, the approach exploits uniquely the textual tags remaining agnostic w.r.t the visual content of the images: this shows once more time the tight connection between tags and visual content; using only the textual wisdom of crowd, our approach leads to a 81.41% of average precision on the PASCAL VOC 2012, with a dataset collected in 15 minutes. At the same

[1] www.flickr.com

time, our approach betters also OPTIMOL [12], which instead exploits also the visual content of the images.

The rest of the paper is organized as follows: Sec. 2 presents the few approaches for the automatic generation of image classification training set, together with some remainders for the query expansion strategy; Sec. 3 details our approach, while Sec. 4 reports its performance on several, comparative experiments. Finally, Sec. 5 concludes the paper and gives some future perspectives.

2 State of the Art

At the best of our knowledge, the only approaches dealing with the automatic generation of training sets are [1,8,13]. Prior to analyze them, it is worth noting that our goal is different from that of the standard image retrieval, where plentiful of algorithms do exist [4]. In fact, image retrieval, in its more studied form, aims at filling the semantic gap between what the user wants to see (a specific query like "my mother smiling at me when I was child") and what the system provides as output. Conversely, our approach focuses on the task of capturing for a given visual concept (like a dog, a car etc.) the most visual variation in terms of pictures. In other words, our idea is to create a system that ideally is able to select for a concept a *visual synset* [5], in the same way that ImageNet does with its content. The difference w.r.t. ImageNet is that our approach has to be dynamic, 100% customizable (no limits on the kinds of visual concepts the user may want to see, so no preconstructed structure as in ImageNet) and fast.

In [8], the authors propose an unsupervised learning approach exploiting Google Image Search. The innovation consists in the use of the Google's automatic translation tools to translate users keywords into 7 different languages and use the translated keywords for collecting images. Since the Google Image Search engine works by indexing images with the text around the images, and some textual metadata [19] the usage of multiple language enriches the visual variation one may expect. The algorithm is also capable to avoid repeated images. In our case, the adoption of multiple languages could rise linguistic issues that in our case we preferred to avoid (remaining on the English language) and postpone for future work.

The other text-based approach is that of [1], in which the starting search keyword (plus an hyperonymy specifying the context) is used to generate a set of bigrams, each one of them addressing a specific semantic aspect of the visual concept. The bigrams are produced by looking at Google Ngrams[2], individuating those additional terms that are visual adjectives (where the "visual" characteristics is found by WordNet[3]), present participles (found by Natural Language Processing basic maneuvers) and hyponymy. These bigrams, ordered by their original frequency in the Ngrams repository or uniformly weighted are used to create specific sets of images (or classifiers) that once pooled together give the final dataset (or classifier).

[2] See https://books.google.com/ngrams/info.
[3] See http://wordnet.princeton.edu/.

The last approach is content-based [13], and starts from a set of seed images; these images are used to train a set of image classifiers, which subsequently are employed to classify unseen pictures. Positively classified images are successively fed into the classifier as training data and the approach is iterated.

Regarding the usage of the tags associated to the images, its usage for crafting training sets represents a major novelty in this paper. Briefly speaking, textual tags put by the users may help in understanding the content of the social images[4]. In particular, the study in [18] showed that the order of appearance of the tags is related to the visual content of themselves, and in particular, that the first tags are more related with the visually dominant patterns.

3 Proposed Methodology

Our approach exploits uniquely the set of tags given by the users to social images. Unfortunately, tags associated to images are usually noisy [10], especially when single images are taken into account. The idea of the approach is to start with a search of I images in relation to a first input keyword k, where the search operates uniquely on the tags associated to the images, and not on other metadata. In this way, we can collect a statistics on the tags received (that is, each image has associated a tag list), which mirrors the intended semantics shared among the people on a given visual concept, from which the term "crowdsearching". Given the list T of all the tags, we create a dictionary D of N terms, pruning away noisy tags like stepwords and other not relevant expressions (see later). Using the tag list and the dictionary, we perform different filtering operations, which will bring to an updated list of bigrams $T_{\text{filtered}} = \{< x, k >\}$, where x is one of the M filtered tags, $M < N$. With the filtered bigrams, query expansion is performed and the final images are retrieved from the social platform, forming the training set. Naturally, the training sets are then validated using cutting edge classifiers (see the experiments, Sec. 4).

Frequency Filter: The frequency filter simply sorts the tags in the dictionary by their frequency in the tag lists of the retrieved images. The idea is then to take the first F terms in the ranking, since they presumably indicate a widely shared visual semantics that implicitly prunes away unrelated concepts (or concepts that are occasionally related by the context). For example, while looking at an image of "dog", the tag "Marie" could be present since "Marie" is the owner of the dog. When moving to a large collection of images, most probably the proper name "Marie" will be characterized by a low count of occurrences, moving low in the rank. Viceversa, the proper name "Bernardo" could be high in the rank, as it defines a dog breed. The importance of the tags by their frequency is currently employed in many applications [17,22], for example the tag cloud [9] for information visualization purposes. The output of this step will be a list T_{freq} of bigrams.

[4] With the term "social images" we intend those images uploaded into social platforms like Flickr.

Keyword Filter: This filter exploits the fact that the tags are organized in ordered lists and that the order in a tag list associated to an image carries some meaning [18]. In particular, the filter works by keeping all the tags that occur in the tag list associated to an image before than the keyword k. The underlying principle in that tags that occur earlier in the tag list are more important [18] and that the keyword of search k could be thought as a threshold where all the terms before of that are important. So, each tag list is filtered here, the resulting terms are organized in a dictionary, ordered by frequency and kept the first F terms, appended to the keyword k in a list of bigrams T_{ord}.

Quality Filter: The quality filter exploits a semantic oracle developed in [16], which essentially is a list of 150 terms which are "semantically rich and general". In few words, linguistic researchers in the past century individuated English terms that cover wide variety of descriptions of different entities. The quality filter analyze the N terms of the initial dictionary, keeping only those that are included in the oracle list, ordering them by frequency of usage, keeping the first F terms. This filter essentially is a specialization of the frequency filter, with output a list T_{qual} of bigrams.

Noun Filter: The last filter essentially performs the intersection of the retrieved tags with a set of nouns which are in the hyponym sets of the given keyword, or in the immediate hyperonym set (found by the help of WordNet). Even in this case, the filtered keywords are ordered by frequency and the first F terms are kept, forming the list $T_{\text{N}oun}$ of bigrams with the keyword k.

Once we apply the filter, we basically apply the query expansion step, retrieving from Flickr I images from each of the F bigrams, and pooling together all the images so to obtain the final training dataset.

4 Experiments and Results

For testing our approach we perform two experimental sessions, one comparing against the text-based method of [1], dubbed here *Semantic Trainer* for conciseness, and the second considering the visual feature-based *Optimol* approach [13]. In both the cases we focus on the Flickr social platform. Flickr allows to look for images considering the associated text tags only, fitting perfectly our scenario. It is worth noting that the Semantic Trainer was originally tested with the Google Image Search engine, so here we apply that approach on Flickr, obtaining results which are obviously different from that of [1].

As for the visual concepts to analyze, we focus on 18 classes of the PASCAL VOC 2012 image classification challenge [6] to compare with the Semantic Trainer, and on 7 object classes to compare with Optimol. These choices have been made for the sake of fairness (the same classes have been taken into account in the original papers [1] and [13], respectively).

In both the experimental sessions, our approach is applied by first downloading $I = 500$ images for each visual concept, keeping the related tag lists. On these lists, we perform basic automatic pre-processing such as removing short

strings (e.g., ab,cb,etc), non-English alphabets (e.g.,Japanes character, Chinese characters,etc), meaningless words (e.g., awsm). We then apply our four filtering techniques, keeping $F = 10$ bi-grams. Depending on the cardinality of the image classes to compare with, we divided uniformly the number of images related to each bi-gram.

4.1 Comparison Against the *Semantic Trainer*

The two experiments against the Semantic Trainer of [1] consist in evaluating the capability of a method in creating a training dataset which afterwards is exploited in the VOC 2012 image classification task. Convolutional Networks (ConvNets) are used as a feature extractor [11]; in particular, we use a feedforward multilayer perceptron, adopting publicly available pre-trained ImageNet deep learning model[20], focusing on the FC7 layer for extracting the features. We then feed the 4096-dimension sparse feature vector to learn class model using linear SVM for each class. In the details, for each image class we use 500 images, while for the negative class we consider a uniform sample from the other classes into play (see [1] for further details).

As for the figures of merit, we employ the PASCAL VOC's interpolated average precision (AP). In the first experiment, 18 classes are taken into account for the original PASCAL VOC image classification task (where an image belongs to the positive set if it contains one (or more) instances of the considered class). As for the competitor, the Semantic Trainer has 6 different versions (a basic version *basic*, a hyponym-based filter *hypo*, a verb-based filter *verb*, a visual adjective-based approach *vadj*), a combination of modules by bi-gram frequency *fcom* and a combination of classifiers *ccom*, which are all reported here. As for our approach, we report the performances of the frequency filter *freq-f*, the quality filter *qual-f*, the keyword filter *keyw-f* and the noun filter *noun-f*. In addition, we report the results using the simple Flickr (*Flickr*) for generating the training images, that is, no filtering + query expansion. The results are shown in Table 1;

As visible, all the filters are comparable with that of the Semantic Trainer, with the difference that our approach is considerably simpler. In addition, the quality filter *qual-f* betters all the other methods, having very good performances on the "tv, sofa, sheep, bird" classes, showing comparable numbers on the other ones.

To understand the quantitative results, we report some qualitative examples of the kind of images (and related bigrams) obtained by our approach and that of the Semantic Trainer (related to the fcom method), considering the classes person and cat. As visible, a higher number of false positives are produced by fcom.

4.2 Generalization Capability

Of significant interest for the practical usefulness of our approach is how well the training datasets generalize beyond the insights gathered on PASCAL VOC. One way to gain an approximate idea is by performing cross-dataset evaluations

Table 1. Comparative classification results on the PASCAL VOC 2012 dataset.

Classes	Flickr	Semantic Trainer						*our techniques*			
		basic	hypo	verb	vadj	fcom	ccom	freq-f	qual-f	keyw-f	noun-f
A.plane	97.7	97.3	95.2	97.4	97.1	**97.9**	97.3	97.3	96.4	97.3	93.6
bicycle	**82.8**	82.5	70.4	79.3	83.6	82.2	81.2	83.1	76.6	82.3	76.5
bird	90.7	90.4	91.5	89.9	90.2	90.1	91.7	90.7	**92.7**	89.3	92.0
boat	88.7	88.2	88.8	87.8	86.9	89.5	89.2	88.9	87.5	**89.3**	89.0
bottle	57.3	56.7	57.5	55.7	55.8	57.6	**58.3**	57.3	54.9	56.8	55.6
bus	**93.8**	93.7	87.3	94.3	93.0	94.1	93.0	93.4	91.6	93.1	92.8
car	72.6	75.6	69.8	71.9	**75.9**	71.6	74.7	73.2	74.6	73.9	73.2
cat	91.5	89.1	92.9	90.6	90.9	91.4	**93.1**	92.9	89.6	90.0	91.5
chair	70.3	69.9	73.3	71.1	72.3	67.8	**74.3**	68.9	66.5	71.0	59.5
cow	**79.0**	73.9	73.6	71.8	75.1	75.7	77.7	76.6	78.8	76.1	64.9
dog	88.9	87.3	89.5	87.3	87.1	86.1	**89.7**	88.8	88.7	86.6	87.5
horse	**85.1**	76.8	80.1	76.9	81.7	80.5	83.0	84.8	83.8	82.2	80.7
M.bike	89.1	89.4	4.7	88.9	91.0	90.7	**91.3**	89.1	89.8	85.5	79.2
person	60.4	61.5	**72.8**	60.6	58.1	63.9	68.4	57.8	71.8	66.8	58.1
sheep	84.9	84.0	85.6	82.9	85.2	84.9	87.2	84.9	**86.3**	86.2	79.5
sofa	58.0	59.6	45.7	52.7	58.7	58.2	59.0	10.6	**62.7**	49.8	39.1
train	92.8	92.4	90.6	93.1	92.2	**93.6**	93.2	89.1	92.7	91.8	91.0
tv	25.0	74.1	55.4	26.2	45.0	46.8	53.1	73.4	**77.1**	31.5	69.3
AP	78.3	80.15	73.6	76.6	78.9	79.1	80.9	77.8	**81.3**	77.8	76.3

between different benchmark datasets, and comparing the relative performance of our training sets.

Following [1], we set out to explore cross-dataset generalization, meaning to perform cross-dataset evaluations between different benchmark datasets, and comparing the relative performance of our training sets. In particular, we analyze the behavior on the "person" class, which is of particular interest for many reasons, spanning from multimedia to social robotics, from surveillance to human computer interaction. For each class, we perform 10 randomized experiments with 200 positive and 400 negative samples split into 50% for training, 25% for cross-validation and 25% for testing. As source of the negative samples, we use the "other" classes of PASCAL VOC. Results are in Table 3.

As visible, even in this case noun filter gives the best result in average among all the approaches of automatic training set generation, having the top scores when considering the Caltech 256 and PASCAL VOC. It is also worth noting that on the PASCAL VOC dataset all our filters give the best performance. Finally, it is encouraging to see that our best score is comparable to what is obtained by ImageNet.

Table 2. Qualitative analysis of our *noun-f* dataset against the 2 *fcom* based, for the classes 'person', and 'cat'.

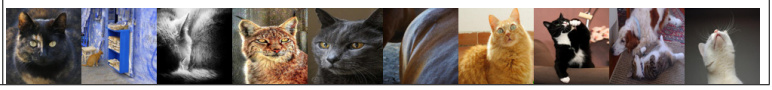
filter	Classes	Top 10 related terms and their corresponding images
noun-f	person	human, people, guy, man, human being, subject, child, lover, artist 
	cat	kitty, feline, meow, kitten, stray, tabby, pussy, lion, tiger, mammal 
fcom	person	black,dead,deceased,dying,good,innocent,living ,religious,white 
	cat	black, blue, domestic, gray, grey, house, orange, playing, sleeping, white 

Table 3. Cross-dataset generalization on the "person" class

Train on :	*Test on:*				Mean others
	ImageNet	Graz	Caltech-256	PASCAL VOC	
PASCAL VOC	95.10	92.22	97.04	94.71	94.77
GRAZ	92.10	99.46	94.32	88.06	93.48
Caltech-256	96.44	90.42	99.33	92.87	94.77
ImageNet	99.14	93.59	97.88	92.39	95.75
ccom [1]	**97.61**	**97.76**	96.07	88.01	94.86
fcom [1]	95.52	94.90	94.79	87.54	93.12
freq-f	95.72	95.72	95.03	88.22	93.68
qual-f	96.48	90.53	96.22	89.85	93.27
keyw-f	96.22	94.58	96.51	90.34	94.41
noun-f	97.21	94.50	**98.00**	**91.28**	**95.25**

4.3 Comparison against OPTIMOL

A more ambitious experiment consists in comparing with the OPTIMOL [13] approach that analyzes the content of the images: in facts, our approach is agnostic with the visual information, working only on textual data. For the sake of comparison, we adopt the same experimental protocol of [13], considering a selection of classes of the Caltech-101 [7], and generating the same number of training images. As for the testing set, we consider all the images provided by the Caltech-101. As for the features, we extract 128-dimensions dense SIFT, quantizing them into a 100-visual word dictionary by applying k-means clustering

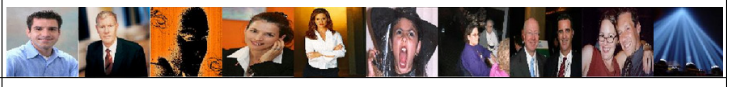
provided by the Vlfeat library [21]. We then use these histograms to train a linear SVM for each class and to perform object classification.

In Table 4 the classification results are reported, showing that surprisingly all of our approaches work better than OPTIMOL. In this case, the frequency filter the best job. Other than the numbers, it is interesting to observe the nature of the images being retrieved, see Table 5. Working on the visual information, OPTIMOL is not capable of distilling basic semantic aspects that strongly penalize its images, for example the fact of having a single face in the image. More in general, it is notable that OPTIMOL retrieve more false positive samples (for more details on OPTIMOL image class sets, please see the link[5]) .

Table 4. Comparison of our techniques performances to OPTIMOL in object classification. The classification performance is better than [13] by 14.11%.

	OPTIMOL	*our techniques*			
		frequency-f	quality-f	keyword-f	noun-f
airplane	76.00	84.07	69.10	79.21	79.87
car	94.50	95.20	94.98	95.11	94.84
face	82.90	83.44	83.32	78.40	90.70
guitar	60.40	97.14	96.99	98.09	97.03
leopard	89.00	92.24	95.49	91.80	92.21
motorbike	67.30	75.83	63.67	71.77	69.03
watch	53.60	94.66	95.98	90.45	89.58
AP	74.81	**88.94**	85.65	86.40	87.61

Table 5. *Face and watch* from our prepared datasets and OPTIMOL. From our database, we show one image for each term given therein.

Filter	Classes	Top 10 related terms and their corresponding images
Frequency	face	black, city, light, mono, monochrome, people, street, vienna, white, woman 
	watch	beautiful, curvy, demure, lovely, pretty, scanner, shapely, time, timepiece, woman 
OPTIMOL	face	
	watch	

[5] Face and watch class images: http://www.cs.stanford.edu/groups/vision/projects/

5 Conclusion

The automatic generation of training sets for image classification will be for sure a hot topic in the next years, where object classifiers will be embedded into portable devices like smartphone. Recently, a 30M dollars funding to the popular applet Shazam for object recognition purposes is a valid proof of our thoughts[6]. The message of this paper is that textual tags usually associated to social images, even if noisy, represent an important source of information that taken alone may bring to expressive image datasets, not so distant from man-made repositories such as PASCAL VOC and ImageNet. In the previous work [1] we show how it is possible to use image metadata to crawl image datasets: associating metadata with textual tags will be therefore a straightforward strategy we aim to investigate, after that we will move to analyze the genuine visual content of the images, trying to understand the relation among visual features and textual features. Connecting visual, textual and metadata should be then the final move for creating visual knowledge for feeding visually intelligent systems which see and understand the world around us.

Acknowledgments. This work was supported by a Research Grant 2013, Prog. FSE cod. 1695/1/24/ 1148/2013 titled "SEMANTIC CLOUD COMPUTING" DGR 1148/2013 as part of the program "Obiettivo Competitivit Regionale e Occupazione - Asse Capitale Umano - Sviluppo del potenziale umano nella ricerca e nell'innovazione" . Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect of the FSE Research.

References

1. Cheng, D.S., Setti, F., Zeni, N., Ferrario, R., Cristani, M.: Semantically-driven automatic creation of training sets for object recognition. Computer Vision and Image Understanding **131**, 56–71 (2015)
2. Carolyn, C.J.: An approach to the automatic construction of global thesauri. Information Processing & Management **26**(5), 629–640 (1990)
3. Crowston, K.: Amazon mechanical turk: a research tool for organizations and information systems scholars. In: Bhattacherjee, A., Fitzgerald, B. (eds.) Shaping the Future of ICT Research. IFIP AICT, vol. 389, pp. 210–221. Springer, Heidelberg (2012)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 5:1–5:60 (2008)
5. Deng, J.,Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)

[6] See http://goo.gl/2jNkFC.

7. Fei-fei, L., Fergus, R., Perona, P.: A bayesian approach to unsupervised one-shot learning of object categories. In: Proceedings of the 9th International Conference on Computer Vision, pp. 1134–1141 (2003)
8. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Proceedings of the 10th International Conference on Computer Vision, Beijing, China, vol. 2, pp. 1816–1823, October 2005
9. Hassan-Montero, Y., Herrero-Solana, V., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces (2006)
10. Kennedy, L.S., Chang, S.-F., Kozintsev, I.V.: To search or to label?: predicting the performance of search-based automatic image classifiers. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR 2006, New York, NY, USA, pp. 249–258. ACM (2006)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates Inc (2012)
12. Li, L.-J., Fei-Fei, L.: Optimol: automatic online picture collection via incremental model learning. International journal of computer vision **88**(2), 147–168 (2010)
13. Li, L.-J., Wang, G., Fei-Fei, L.: Optimol: automatic online picture collection via incremental model learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8, June 2007
14. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 191–197. ACM (1999)
15. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
16. Ogden, C.K.R.: Qualities - descriptive words. Linguistic **3**(x), x (1930)
17. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, New York, NY, USA, pp. 259–266. ACM (2008)
18. Spain, M., Perona, P.: Some objects are more equal than others: measuring and predicting importance. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 523–536. Springer, Heidelberg (2008)
19. Sun, F., Wang, M., Wang, D., Wang, X.: Optimizing social image search with multiple criteria: Relevance, diversity, and typicality. Neurocomputing **95**, 40–47 (2012). Learning from Social Media Network
20. Vedaldi, A., Lenc, K.: Matconvnet - convolutional neural networks for matlab. CoRR, abs/1412.4564 (2014)
21. Vedaldi, A., Fulkerson, B.: Vlfeat: an open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia, MM 2010, New York, NY, USA, pp. 1469–1472. ACM (2010)
22. Weinberger, K.Q., Slaney, M., Van Zwol, R.: Resolving tag ambiguity. In Proceeding of the 16th ACM international conference on Multimedia, MM 2008, New York, NY, USA, pp. 111–120. ACM (2008)