

# Cross-Dataset Learning of Visual Concepts

Christian Hentschel, Harald Sack, and Nadine Steinmetz

Hasso Plattner Institute for Software Systems Engineering,  
Potsdam, Germany

`firstname.lastname@hpi.uni-potsdam.de`

**Abstract.** Visual content classification has become a keystone when opening up digital image archives to semantic search. Content-based explicit metadata often is only sparsely available and automated analysis of the depicted content therefore provides an important source of additional information. While visual content classification has proven beneficial, a major concern, however, is the dependency on large scale training data required to train robust classifiers. In this paper, we analyze the use of cross-dataset training samples to increase the classification performance. We investigate the performance of standardized manually annotated training sets as well automatically mined datasets from potentially unreliable web resources such as Flickr and Google Images. Next to brute force learning using this potentially noisy ground truth data we apply semantic post processing for data cleansing and topic disambiguation. We evaluate our results on standardized datasets by comparing our classification performance with proper ground truth-based classification results.

**Keywords:** Image Classification, Cross-dataset learning

## 1 Introduction

In recent years, automatic classification of visual content such as video and photo data has gained increasing interest from several research communities. The digital era not only made recording and storage cheap and easy but also enabled new distribution channels that made pictorial data available for a larger audience. Authorship of visual content is no longer limited to professionals and Internet community platforms such as *Flickr*<sup>1</sup> are hosting an ever increasing number of private photo collections.

With the growth of visual data came the need to search and retrieve information within these collections. Professional archives and companies such as stock photo agencies have a strong commercial interest in making their content not only accessible but also searchable via Internet. It became apparent that a manual annotation of the depicted content with describing metadata will always be incomplete, subjective and most of all infeasible due to the sheer number of assets even if letting alone these that are added every day.

---

<sup>1</sup> Flickr: <http://www.flickr.com/>

Research efforts in computer vision target the demands of today's image archives for efficient search and retrieval methods in stored content. Typically, the task is to automatically recognize categories of objects and scenes depicted in photos (i.e. visual concepts) – a task that is a fundamental ability of humans but still an elusive goal when assigned to machines. Various approaches have been presented in literature in recent years. Recognition is usually considered as a classification problem of separating positive from negative examples of a given visual concept. Most approaches rely on supervised machine learning techniques that require a set of manually annotated data for training a model of a specific concept.

However, as manual annotation is labor-intensive, most datasets publicly available, despite the undisputed effort that has been accomplished, are rather limited in terms of number of annotated images as well as number of concepts used to describe the depicted content. The breadth of the semantic space covered by the selected visual concepts, however, has important implications for the real-world applicability of automatic visual content classification. Considering the aforementioned application scenarios for automatic search and retrieval in large scale photo collections a sufficiently high (preferably unlimited) number of concepts is a minimum requirement. The fewer concepts an automatic classifier is able to recognize, the less useful it is in solving the problems of today's collections as search is limited to the few covered ones.

Hence, a question that arises from these observations is how to significantly increase the number of training data while at the same time limiting the manual effort required to allocate this data. Preferably, in order to meet the goal of supporting arbitrary user queries, the number of covered concepts should be completely independent of manual annotations. In this paper, we present a first step towards this direction by proposing cross-data model training. We first evaluate the straightforward approach of training on additional manually annotated datasets in order to increase the number of covered concepts. The question we target at is whether concepts provided by the various computer vision benchmarking initiatives can be used interchangeably, i.e. whether a specific concept implicitly defined by the training data of one dataset is congruent with the definition of the same concept of another dataset.

Secondly, we investigate the use of training data that requires no manual effort for ground truth generation by using automatically crawled Flickr photos based on matching a specific concept query with user provided tags. We aim at answering the question whether the automatic allocation of labeled data for ad hoc training of a visual concept model can be successful in order to advance towards unlimited user queries. In order to reduce noise within the training data we propose methods for automatic data cleansing based on statistical and semantic analysis of the associated user tags. This goal, however, could only partially be achieved.

This paper is structured as follows: Section 2 reviews the relevant literature related to the presented approach. In Section 3 the applied bag-of-visual-words method for visual concept recognition is briefly described. The main part of

this work is presented in Section 4 where the proposed approaches for labeled data acquisition and cleansing are presented. Section 5 provides an evaluation of the proposed approach based on the ImageCLEF 2011 evaluation data. Finally, Section 6 concludes the paper and gives a brief outlook to future work.

## 2 Related work

The allocation of training data sets for content-based image classification has been the mission for various campaigns, challenges and benchmarking initiatives that aim at providing a comparative evaluation of different approaches. One of the largest efforts in this direction has been the “ImageCLEF Visual Concept Detection and Annotation Task” [?]. The contributed dataset today covers 99 different visual concepts that have been used to manually annotate 18,000 photos, i.e. different concept labels have been assigned to images based on human evaluation whether or not a specific concept is depicted within the image. While the task of annotating 18,000 images is huge the potential benefit for real world scenarios still seems negligible since 99 different concepts are by no means satisfactory to cover a reasonable amount of potential user queries emitted to a photo collection.

Only recently, crowd sourcing strategies such as Amazon Mechanical Turk (MTurk)<sup>2</sup> have provided the possibility to substantially increase the number of sample images, the annotation diversity as well as the number of per-image annotations. The labels provided by the ImageNet project [?] – probably today’s largest database of manually annotated photos – are created through a large-scale MTurk process. The ImageNet database is organized according to the WordNet<sup>3</sup> hierarchy: The declared aim of the ImageNet project is to provide on average 1,000 images to illustrate each “synonym set” or “synset” defined in WordNet. In April 2010, ImageNet indexed 14,197,122 images aligned to 21,841 synsets<sup>4</sup>.

Next to these clean, manually labeled training data other initiatives focused on the aggregation of potentially weakly labeled datasets compiled by web photo communities such as Flickr or by querying standard web search engines such as Google. Similar to ImageNet the TinyImages[?] collection is arranged around WordNet. By sending all non-abstract WordNet nouns as search queries to image search engines the authors collected 80 million low resolution ( $32 \times 32$  pixels) images each labeled with the respective noun. On average each noun is described by a set of 1,056 images and the average precision within the sets is estimated to be at 10-25%. The high level of noise makes the dataset less useful for training data acquisition and the lack of additional metadata limits data cleansing to visual-only approaches. Content-based analysis of the images, however, is difficult due to their low resolution.

The Flickr platform provides a public API to query their database in order to retrieve photos that have been tagged by users with a given search term.

<sup>2</sup> Amazon Mechanical Turk: <https://www.mturk.com/mturk/welcome>

<sup>3</sup> WordNet – A lexical database for English: <http://wordnet.princeton.edu/>

<sup>4</sup> ImageNet – Summary and Statistics: <http://image-net.org/about-stats>

By means of this API the MIRFLICKR Retrieval Evaluation[?] automatically assembled the MIRFLICKR-1M collection that provides 1 million Flickr images published under the Creative Commons license. Moreover, next to the plain images, the collection also contains the Flickr user tag data if provided by the Flickr users. However, the images are not manually annotated and the tags submitted by Flickr users cannot be considered of similar quality as the ground truth provided by human annotators in the ImageNet initiative. The dataset used in the ImageCLEF benchmark initiative is a manually annotated subset of MIRFLICKR image data.

To the authors best knowledge no effort to analyze the various datasets for congruency in terms of ground truth data has been made yet. The availability of inexpensive web-image data, however, has created considerable interest in the computer vision community to employ this data for training of visual classifiers. In [?] the author uses co-training in order to improve a classifier trained on a small quantity of labeled data. Unlabeled images, which are confidently classified by one classifier are added to the training set of another classifier. Other work aims at prior cleansing of weakly labeled Internet images by means of visual analysis. For example, the authors in [?] train models for parts and spatial configuration of objects without supervision from cluttered web images. The models are later used to re-rank the output of an image search engine. In [?] an iterative framework for visual classification of downloaded web images retrieved through an image search engine is presented.

Text-based outlier removal is performed in [?] where the 100 words surrounding an image link in its associated web page are used for identification of a set of images to be used as visual exemplars for animal classification. Similarly, in [?] and [?] images returned by a web search engine are re-ranked based on the text surrounding the image and metadata features. The top-ranked images are then used as (noisy) training data. With regard to the use of Flickr tags as resource for training data ground truth the authors in [?] use the MIRFLICKR-25000 dataset to train a multiple kernel learning classifier. Tag data as well as visual features are combined and a semi-supervised approach is applied to remove examples that are likely to be incorrectly tagged. Finally, in [?] the authors propose a method to evaluate the effectiveness of a tag in describing the visual content of its annotated images.

### 3 Content-based Visual Concept Classification

In this section we briefly present the applied Bag-of-(Visual-)Words (BoW) approach for content-based visual concept classification. As the major focus of this paper is not put on improving the various aspects of the BoW method we restrain the presentation to these details required to ensure repeatability of the conducted experiments. For further information we refer to the related work in concept classification (cf., e.g. [?,?,?]).

In our experiments, we extract SIFT (Scale-Invariant-Feature-Transform, [?]) features at a fixed grid of  $6 \times 6$  pixels on each channel of an image in RGB color

space<sup>5</sup>. By concatenating these features we obtain a 384-dimensional feature vector at each grid point. These features are used to compute a visual vocabulary by running a  $k$ -means clustering that provides us with a set of representative visual words (codewords). We compute  $k = 4,000$  cluster centers on the RGB SIFT features taken from the training images set. By assigning each of the extracted RGB-SIFT feature of an image to its most similar codeword (or cluster center) using a simple approximate nearest neighbor classifier we compute a normalized histogram of codeword frequencies, i.e. a Bag-of-Words, that is used to describe this image. The combination of SIFT for local image description and the BoW model makes the approach invariant to transformations, changes in lighting and rotation, occlusion, and intra-class variations [?].

Once the image feature vectors have been computed the problem of visual concept recognition can be approached by standard machine learning techniques. Kernel-based Support Vector Machines (SVM) have been widely used in image classification scenarios (cf. [?, ?, ?]). We use a Gaussian kernel based on the  $\chi^2$  distance measure, which has proven to provide good results for histogram comparison<sup>6</sup>. Following Zhang et al. [?] we approximate the kernel parameter  $\gamma$  by the average distance between all training image BoW-histograms. Therefore, the only parameter we optimize in a 4-fold cross-validation is the cost parameter  $C$  of the support vector classification. New images can be classified using the aforementioned Bag-of-Words feature vectors and the trained SVM model.

We consider the classification task a one-against-all approach – one SVM per given visual concept is trained to separate the images from this concept from all other given concepts. Hence, the classifier is trained to solve a binary classification problem, i.e., whether or not an image depicts a specific visual concept. This approach provides us with two advantages. First, new concepts can easily be added by simply training a new classifier which is in line with the demand for easy concept extension. Since the features are not adapted to the classification task they can be reused. Second, multiple concepts can be assigned to each image depending on the prediction confidence of each classifier available, again an important property when aiming at preferably unlimited concepts.

## 4 Cross-dataset Training

Benchmarking datasets for visual content classification consist of a set of labeled training images and a set of evaluation images whose ground truth labels are known only to the authors of the benchmarking initiative. Typically training and evaluation sets are obtained by splitting a larger dataset into two smaller ones (e.g. at a rate of 50% training and 50% evaluation data). While this is reasonable in order to provide a certain degree of homogeneity between training and evaluation data, this likewise reduces the already limited and valuable training data.

<sup>5</sup> We use the OpenCV 2.4.1 SIFT descriptor implementation: <http://opencv.org/>

<sup>6</sup> Our implementation is based on the libsvm-3.1 Library for Support Vector Machines: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

We consider cross-dataset training as training a model for a given visual concept on a dataset that is completely disjoint in terms of its history of origins from the dataset used to evaluate the classification accuracy of the trained classifier. By following this approach we evaluate to what extent different datasets are congruent in terms of the concept definition implicitly primed by the provided positive and negative samples.

#### 4.1 Evaluation Dataset

For our experiments in cross-dataset learning we decided to use the evaluation set of the “ImageCLEF 2011 Visual Concept Detection and Annotation Task” as ground truth data to estimate the performance of the models trained on different datasets. The dataset has been used in the 2010 and 2011 benchmarking initiative and provides a training dataset of 8,000 photos manually annotated with 99 different visual concepts while the evaluation set comprises 10,000 likewise annotated photos. The ground truth has been publicly released and therefore can be used by researchers to compare their algorithms with others. The choice of this dataset as ground truth for our tests not only provides traceability of our experiments but also comparability to other research results that were published alongside with the benchmarking initiative.

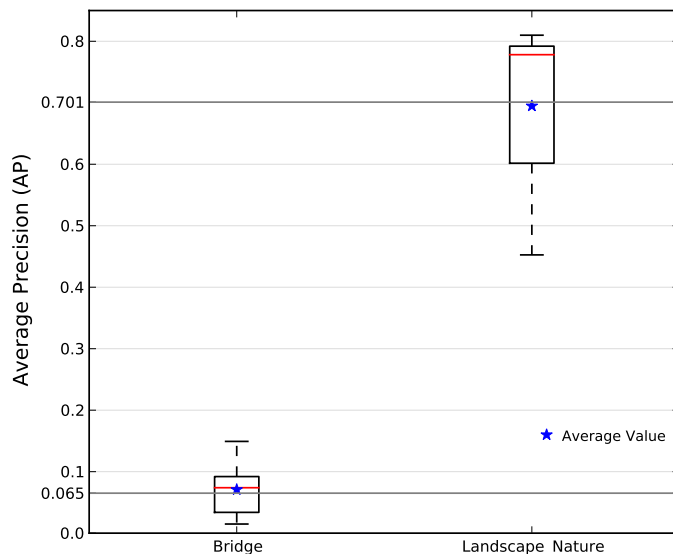
We decided to select two concepts for testing purposes “Bridge” and “Landscape\_Nature”. The reason being that the average classification performance of the concept “Bridge” was one of the worst among all results submitted by different participants of the benchmark despite a reasonably sized training set (i.e. 105 photos labeled with “Bridge” are available in the training set). The decision for the concept “Landscape\_Nature” was based on the observation that most participants in the benchmark performed rather well in classifying photos as members of this concept. A number of 1,362 photos are positively labeled with “Landscape\_Nature” in the training set.

#### 4.2 Training Data Acquisition

We use three different datasets for acquisition of positive and negative sample images:

**ImageCLEF** As a first step in order to provide a baseline to our experiments we choose the ImageCLEF training set for both concepts to train two models that we evaluate by classifying the photos in the evaluation set. Similar to the evaluators of the ImageCLEF task we compute the interpolated average precision (iAP) as an evaluation measure, i.e. the average precision at 11 recall intervals. We train our models using the BoW method described in Section 3. The vocabulary is generated by clustering 800,000 RGB SIFT descriptors randomly sampled from the training set. The classifiers are trained using all photos labeled with the respective concept as positive samples and all other training photos as negatives. The results serve as a baseline (see Section 5). Figure 4.2 plots the distribution

of the results of the various participants of the 2011 ImageCLEF challenge as well as the results obtained by our own default BoW approach. We restrained the results to those obtained by visual-only classifiers (i.e. classifiers that operate on the visual image features only and that do not use any additional textual metadata such as Flickr tags, see [?] for more information). The plot shows that our approach resembles the average result of all participants.



**Fig. 1.** Results obtained by participants of the 2011 ImageCLEF Visual Concept Detection and Annotation Task (in terms of interpolated average precision). The horizontal lines mark the results obtained by our own default BoW approach (i.e. “Bridge”:  $AP = 0.065$ , “Landscape\_Nature”:  $AP = 0.701$ ).

**ImageNet** As a second training corpus we use sample images from the ImageNet project. We map the ImageCLEF visual concept “Bridge” to the ImageNet synset “Bridge, span”, which contains at the time of the experiments 1,598 images. The concept “Landscape\_Nature” is mapped to the synset “Landscape” comprising 76 images. Negative images are randomly sampled from the validation set of the “ImageNet Large Scale Visual Recognition Challenge 2010”, which is available at the ImageNet website after registration. We vary the ratio between positive and negative samples in the training set (see Section 5). The BoW vocabulary is generated by clustering 800,000 RGB SIFT descriptors randomly sampled from all training data. As before, performance is measured using iAP based on the ImageCLEF evaluation set.

**MIRFLICKR** Finally, we wanted to know how classification of visual concepts performs when models are trained based on training data crawled from unreliable web resources. We use the MIRFLICKR-1M collection for training data acquisition. Especially, we use only Flickr user tag data as ground truth and thus ignore any manually generated annotations provided by the MIRFLICKR Retrieval Evaluation. All MIRFLICKR-1M tags are preprocessed to be lower case. We select a subset of 100,000 photos to form our training data. Positive sample selection is performed by sub-selecting these photos that are tagged with the term *bridge* as a single word ('bridge') as well as with the term occurring as a substring within a tag ('%bridge%'). Negative sample selection is based on random sub-sampling of all images that were not selected as positive samples. Again, we experiment with different sizes of negative sample sets. Similarly, positive and negative samples are selected for the term *landscape* ('landscape', '%landscape%'). Thus, we obtain two different training sets per concept. The BoW vocabulary generation is based on a random subset of 800,000 RGB SIFT descriptors sampled from all 100,000 training photos. Consistently with the other runs, we measure the performance in terms of iAP computed on the ImageCLEF evaluation set.





### 4.3 Training Data Cleansing

Naturally, as the Flickr user tag data is not intended to provide a reliable ground truth annotation for classification purposes, the training set annotations derived based on these data must be considered as noisy. In Table 1 a few samples for images taken from the MIRFLICKR-1M dataset are presented that have been tagged using one of the selected visual concepts. Clearly, none of the images actually depicts the concept. A mislabeling or misleading labeling with tags can have various reasons. In [?] the authors analyze different functions tags perform in collaborative tagging systems which holds as well for folksonomies such as Flickr. Among others, tags are used as organizational structure, e.g. the photographer wanted to group all photos he or she has taken when visiting the Golden Gate Bridge. As a matter of fact, the pictures do not necessarily depict the bridge at all. Another reason for a misleading tag can be observed when considering the picture in the third row of Table 1: the photographer might have actually stood on a bridge, so the tag rather describes his viewpoint than the actual concept. Moreover, a photo can be tagged with the term 'bridge' when actually showing only a single pylon or a small part of a bridge such as a rivet. Finally, a tag can be assigned with no visible relation to the depicted content as can be seen in the first and last example image in Table 1.

By means of data cleansing we intend to filter these images that actually show the concept from those that have been mislabeled. We use two different strategies for data cleansing: tag co-occurrence analysis and semantic tag analysis.

**Tag Co-occurrence Analysis** In case of the first picture in Table 1 it seems rather unlikely that the tags 'downtown', 'cityscape' and 'skyscrapers' appear



	tag set
	houstontx, houston, downtown, harriscounty, southside, hdr, <b>landscape</b> , cityscape, ben, texas, usa, skyscrapers, canon40d, 40d
	350d, dacha, houseslippers, <b>landscape</b> , sigma1770f2845, uyma, slippers, flickr, explore, interestingness
	stone, stones, river, water, reflection, sky, blue, brown, wet, dry, underwater, ripples, song, it-stonedme, vanmorrison, betws, <b>bridge</b> , amman, fotocyfer, sonyalpha350
	<b>bridge</b> , motel, seattle, sign, toddbates

**Table 1.** Misabeled or misleadingly labeled images taken from the MIRFLICKR-1M dataset. Tags that have been used as query tags marked in **bold**.

frequently together with the tag 'landscape'. We analyze the tag co-occurrences for each of the two query tags and filter these tags that appear most frequently together with the query tag. These ten most frequent tags are listed in Table 2.

Query tag	Most frequent co-occurring tags
'bridge'	'river', 'water', 'night', 'sky', ' <del>nikon</del> ', ' <del>hdr</del> ', 'city', 'reflection', 'blue', 'clouds'
'landscape'	'nature', 'sky', 'clouds', 'water', 'sunset', ' <del>nikon</del> ', 'trees', 'paisaje', 'blue', ' <del>canon</del> '

**Table 2.** Flickr user tags with the most frequent co-occurring tags in the training set.

By manual blacklisting, we drop these tags that are technical tags (e.g. the maker of the camera or the tag 'hdr') and further reduce the list to the 5 most frequent remaining tags ( ['river', 'water', 'night', 'sky', 'city'] and ['nature', 'sky', 'clouds', 'water', 'sunset']). Finally, we sub-select these images from the training set whose tag sets contain at least one of these most frequent non-technical tags co-occurring with the query tag.

**Semantic Tag Data Analysis** In this approach, we analyze the semantic relationship between the tags in a tag set of an image and the visual concept that is to be classified. In order to do so, we first manually map the target visual concepts “Landscape\_Nature” and “Bridge” to DBpedia<sup>7</sup> entities (<http://dbpedia.org/resource/Landscape> and <http://dbpedia.org/resource/Bridge>). Next, we automatically map every tag from the tag dataset of a Flickr training image to a semantic entity referenced by a DBpedia URI by means of named entity recognition (NER). NER is the process of annotating textual information with semantic entities. It mainly consists of four steps:

- scanning the text (i.e. tag set) for potential named entity terms
- finding entity candidates for each tag
- defining the context of a tag (all other tags from the same tag set as well as the Flickr title of the image, if available, are used as context)
- ranking the entity candidates and determining the relevant candidate in the given context

The last two steps are only necessary in case of ambiguous tags, i.e. where more than one entity candidate was found for a given tag. The ranking algorithm of the entity candidates calculates two scores for every entity candidate of the ambiguous terms: co-occurrence analysis and link graph analysis.

The co-occurrence analysis uses context terms to determine the co-occurrence of an entity candidate and these terms. Therefore, a textual description of the entities is needed. As we use DBpedia entities, the according Wikipedia articles for the entities are used as descriptive texts. The link graph analysis uses the Wikipedia page link graph to find sub cliques representing the given context within this huge graph. This analysis step is based on the assumption that semantically related entities are linked over their Wikipedia articles. The graph analysis algorithm takes into account paths between the entity candidates of the context with a maximum length of 2. Both analysis algorithms calculate a score for every entity candidate and the weighted sum of both scores is used for entity mapping. The entity candidate with the highest score within the given context will be chosen as mapped entity (cf. [?] for further information on entity recognition).

Based on the successfully disambiguated tags we compute the number of tag entities that have a direct link to the entity of the target visual concept within the DBpedia. This score gives us an indicator of how strongly a tag set of a given image is related semantically to the respective concept and thus, how much related the image is. We use several lower bounds of relatedness and select only these images that have at least  $\tau \in \{1, 3, 4\}$  tags whose entities exhibit a direct link.

## 5 Cross-dataset Classification Results

In this section we present and discuss the results that have been obtained using the aforementioned methods for training data acquisition and visual concept

<sup>7</sup> DBpedia: <http://dbpedia.org/About>

classification. Table 3 summarizes the results for the visual concept “Bridge”. The results obtained by the classifier trained on the ImageNet dataset show an improvement of 20% in interpolated average precision when compared to the baseline classifier. This can be explained by the significantly increased number of positive training samples: 1,598 images falling into the synset “Bridge, span” are used as positive training samples whereas only 105 positively labeled samples are available in the ImageCLEF training set. In fact, when considering this scale ratio, a much higher improvement would have been expected. We assume, however, that the visual variance in the concept “Bridge” is too large to be successfully trained. Second, we can observe that by increasing the number of negative samples better results were obtained. A saturation seems to be reached when using a set of negative samples that is approx. three times as big as the set of positives. To summarize, ImageNet seems to be congruent with ImageCLEF in terms of the implicit definition of the visual concept ‘bridge’. As expected, the results we obtain using the noisy MIRFLICKR training dataset are below the ones obtained using manually labeled training data. The loss in accuracy, however, is rather low: The best result of  $iAP = 0.063$  obtained using the uncleaned training data is only 3% below the result obtained using the baseline classifier. This seems to be a moderate trade-off when comparing the effort needed to provide the different training datasets.

Unfortunately, the proposed methods for training data cleansing do not provide an improvement. While training data refinement using tag co-occurrence analysis (*'bridge'+tag-co-occ*) provides results similar to the results of the best performing uncleaned sample set, semantic data analysis (*'%bridge%'+NER*) does not help to improve classification accuracy. Instead, the performance drops even below the noisy classifier.

Named entity recognition is able to map tags such as ‘PontNeuf’ to the DBpedia ‘bridge’ entity, which should increase diversity. However, in our current prior data selection strategy, images containing ‘PontNeuf’ are only considered when the tag ‘bridge’ is also present, which is not necessarily the case. Thus, in future work, image selection should not be based solely on tags matching the target concept but also on tags matching linked tags provided by NER. Furthermore, since the NER is based on DBpedia and DBpedia itself is based on Wikipedia Infoboxes that are usually not provided for superordinate categories such as ‘landscape’ and ‘bridge’ only few direct links between the target concept and the tags can be identified. Entities of frequently co-occurring and tags (and probably frequently co-appearing concepts) such as ‘bridge’ and ‘river’ do not exhibit direct DBpedia links and are thus much harder to identify than with simple statistical co-occurrence analysis.

Analysis of the classification results for the concept “Landscape\_Nature” (see Table 4) shows a slightly different picture. First of all, the ImageNet-based “Landscape”-classifier does not show superior performance as in the case of “Bridge, span”. On the one hand, this can be attributed to the fact that the number of positive samples is significantly lower (i.e. 76). Again, better results were obtained when scaling up the number of negative samples. Furthermore,

Training set configuration	iAP	#Pos	#Neg
ImageCLEF (Baseline): 'Bridge'	.065	105	7,895
ImageNet: 'Bridge, span'			
1Pos1Neg	.064	1,598	1,598
1Pos3Neg	.078	1,598	4,794
1Pos5Neg	.078	1,598	7,990
MIRFLICKR: 'bridge'			
1Pos1Neg	.059	874	896
1Pos2Neg	.063	874	1,792
MIRFLICKR: '%bridge%'			
1Pos1Neg	.061	1,289	1,311
MIRFLICKR: 'bridge'+tag-co-occ			
1Pos3Neg, $co - occ \geq 5$	.063	485	1,497
MIRFLICKR: '%bridge%'+NER			
1Pos1Neg, $\tau = 1$	.039	631	643
1Pos1Neg, $\tau = 3$	.048	192	195
1Pos1Neg, $\tau = 4$	.031	93	94

**Table 3.** Results (interpolated average precision, iAP) for classification of the evaluation set using models for the ImageCLEF visual concept “Bridge” trained on different datasets obtained with different sampling strategies.

when looking at example images taken from the ImageNet “Landscape” set (see Figure 5, top) we notice that compared to the ImageCLEF evaluation data (see Figure 5, bottom), the visual variance is significantly lower. This example shows a classical problem in dataset annotations: an annotator typically answers the question of whether or not a concept is *present* in an image (i.e. the ImageCLEF annotations), a user of an image search engine, however, presumably looks for *typical* images visualizing the concept (i.e. the ImageNet approach is to ask whether an image is a typical or wrong example). This discrepancy is only visible when manually analyzing the visual content of the datasets, which makes cross-dataset learning difficult if not infeasible. We state that the ImageNet synset “Landscape” is not congruent with the ImageCLEF concept “Landscape\_Nature”: While the ImageNet set depicts typical landscape scenes, the “Nature” aspect is much stronger within the ImageCLEF data. Future runs therefore should try to enrich the ImageNet data by samples taken from synsets such as “Natural object” or “Geological formation” in order to increase the variance and meet the diversity of the evaluation set.

Similarly to the results obtained for the concept “Bridge”, MIRFLICKR-based ‘landscape’-classifiers achieved lower accuracy when compared to the baseline. Again, this is an expected outcome due to the assumed noise in the dataset. However, in this case, training data cleansing based on tag co-occurrence analysis ( $iAP = 0.662$ ) outperforms the best classifier obtained without data cleansing ( $iAP = 0.647$ ) by 2%. Furthermore, differently from the previous results for the

concept “Bridge”, classifiers based on training data cleansed using semantic tag data analysis outperform those who are based on raw Flickr training data.

The loss in accuracy when comparing the best MIRFLICKR-based ‘landscape’-classifier to the baseline classifier is at 6%. Again, we believe this is an acceptable result, w.r.t. the labeling efforts required.

Training set configuration	iAP	#Pos	#Neg
ImageCLEF (Baseline): ‘Landscape_Nature’	.701	1,362	6,638
ImageNet: ‘Landscape’			
1Pos5Neg	.591	76	380
1Pos20Neg	.607	76	1,520
MIRFLICKR: ‘landscape’			
1Pos1Neg	.638	1,704	1,831
1Pos2Neg	.646	1,704	3,662
1Pos5Neg	.647	1,704	9,155
MIRFLICKR: ‘landscape’+tag-cooc			
1Pos3Neg, $co - occ \geq 5$	.662	1,059	3,441
MIRFLICKR: ‘%landscape%’+NER			
1Pos1Neg, $\tau = 1$	.645	1,528	1,641
1Pos2Neg, $\tau = 1$	.655	1,528	3,282
1Pos3Neg, $\tau = 1$	.649	1,528	4,923
1Pos1Neg, $\tau = 3$	.624	310	337

**Table 4.** Results (interpolated average precision, iAP) for classification of the evaluation set using models for the ImageCLEF visual concept “Landscape\_Nature” trained on different datasets obtained with different sampling strategies.

## 6 Conclusions

In this paper we have shown that cross-dataset classification of visual concepts is possible and can actually achieve equivalent and even better results than training performed on the same dataset. The performance of the classifier for the “Bridge” concept trained on the manually labeled ImageNet dataset outperformed the classifier trained on the ImageCLEF dataset. Furthermore, the loss in accuracy imposed by the weakly labeled MIRFLICKR training data seems acceptable when considering that no manual effort is required to assemble the data. While weakly labeled datasets such as Flickr do not intend to provide a “clean” ground truth for model training they still represent a very valuable resource for learning useful tag-image relationships simply due to their size.

We furthermore have shown that training data cleansing in fact can help to reduce noise in weakly labeled datasets and thus provides a promising first step towards reliable, inexpensive and unlimited training data. While this is an encouraging result, future research must proof whether our observations hold



**Fig. 2.** Examples for images assigned to the synset “Landscape” in ImageNet (top) and to the concept “Landscape\_Nature” in ImageCLEF (bottom).

for a larger classification scenario with larger training data. Therefore, as a next step we intend to train models for all ImageCLEF concepts in order to provide a broader analysis. Besides, our work on training data cleansing was limited to improve reliability of positive samples sets. Likewise, negative sets need closer attention especially when realizing that the probability of users *not* tagging a specific concept is much higher than the probability of tagging it.

Important questions that need to be answered are the aforementioned lack of visual coherence between different datasets and the problem of understanding what a tag actually means with respect to the tagged image. User tag data can have various functions and not necessarily relates to the visual information in an image. The aim of data cleansing is to assemble visually coherent image content, as e.g. full view pictures of bridges rather than images where bridges are only depicted in parts, barely visible due to clutter, low resolution, and strong viewpoint variations or even not depicted at all. By thorough research of the correlation of different user provided tags, especially under consideration of the folksonomy aspects of communities such as Flickr, we intend to improve our data cleansing strategies by semantic analysis.

**Acknowledgements** This work was supported in part by means of the German National Library of Science and Technology under the project AV-Portal.

## References

1. Berg, T.L., Forsyth, D.A.: Animals on the web. In: Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition - Volume 2. pp. 1463–1470. CVPR '06, IEEE, Washington, DC, USA (2006)

2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., Maupertuis, D.: Visual Categorization with Bags of Keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
4. Fergus, R., Perona, P., Zisserman, A.: A visual category filter for google images. In: Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic. pp. 242–256 (May 2004)
5. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. *J. Inf. Sci.* 32, 198–208 (April 2006)
6. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: IEEE Conference on Computer Vision & Pattern Recognition, CVPR 2010. pp. 902–909. IEEE, San Francisco, CA, USA (2010)
7. Leung, T., Malik, J.: Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision* 43(1), 29–44 (2001)
8. Levin, A.: Unsupervised improvement of visual detectors using co-training. In: ICCV. pp. 626–633 (2003)
9. Li, L.J., Wang, G., Fei-fei, L.: Optimol: automatic online picture collection via incremental model learning. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
10. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (Nov 2004)
11. Mark J. Huiskes, B.T., Lew, M.S.: New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval. pp. 527–536. ACM, New York, NY, USA (2010)
12. Mbanya, E., Hentschel, C., Gerke, S., Liu, M., Ndjiki-nya, P.: Augmenting Bag-of-Words – Category Specific Features and Concept Reasoning. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
13. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
14. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. In: Proceedings of the 11th International Conference on Computer Vision (2007)
15. Snoek, C., Worring, M.: Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval* 2(4), 215–322 (2009)
16. Steinmetz, N., Sack, H.: Named entity recognition for user-generated tags. In: Proc. of the 8th Int. WS. on Text-based Information Retrieval. IEEE CS Press (2011)
17. Sun, A., Bhowmick, S.S.: Image tag clarity: in search of visual-representative tags for social images. In: WSM '09: Proceedings of the first SIGMM workshop on Social media. pp. 19–26. ACM, New York, NY, USA (2009)
18. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(11), 1958–1970 (Nov 2008)
19. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
20. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. J. of Computer Vision* 73(2), 213–238 (Sep 2006)