Collections for Automatic Image Annotation and Photo Tag Recommendation

Philip J. McParlane[0], Yashar Moshfeghi, and Joemon M. Jose

University of Glasgow,
School of Computing,
Glasgow, G12 8QQ, UK
p.mcparlane.1@research.gla.ac.uk, Yashar.Moshfeghi@glasgow.ac.uk,
Joemon.Jose@glasgow.ac.uk

Abstract. This paper highlights a number of problems which exist in the evaluation of existing image annotation and tag recommendation methods. Crucially, the collections used by these state-of-the-art methods contain a number of biases which may be *exploited* or *detrimental* to their evaluation, resulting in misleading results. In total we highlight *seven* issues for *three* popular annotation evaluation collections, i.e. Corel5k, ESP Game and IAPR, as well as *three* issues with collections used in *two* state-of-the-art photo tag recommendation methods. The result of this paper is two freely available Flickr image collections designed for the fair evaluation of image annotation and tag recommendation methods called Flickr-AIA and Flickr-PTR respectively. We show through experimentation and demonstration that these collection are ultimately fairer benchmarks than existing collections.

Keywords: evaluation, collection, annotation, tag recommendation

1 Introduction

Given the increase in popularity of photo sharing websites, there has been a recent research focus on the indexing and retrieval of such content. A recent study, showed that 65% of images uploaded to popular image sharing website, Flickr[1], contain less than four tags [18], this in turn makes retrieval difficult. Therefore, one of the major challenges in the field involves predicting the objects and concepts present within an image in order to allow for such retrieval. Last decade, a number of research works focussed on the automatic image annotation (AIA) of images and the semi-automatic process of photo tag recommendation (PTR) in order to extract meaning from an image.

Despite the amount of work taken out in these fields, a comparison of approaches is difficult due to the lack of a unified evaluation framework and collection. A review of the 20 most popular automatic image annotation papers[2]

[0] This research was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe)

[1] http://www.flickr.com/

[2] Selected by searching `http://citeseerx.ist.psu.edu/` for "`automatic image annotation`". Order by descending citation count (Dec'12)

showed that at least 15 different collections were tested upon[3]. These collections vary in characteristics and hence introduce biases of their own into the evaluation, highlighting the need for a single test collection which is *representative* of images uploaded to image sharing websites. Additionally, the most prominent works in photo tag recommendation all use their own collections [18,11,5].

Aside from the large number of collections used to benchmark annotation models, we have identified *seven* flaws which may result in misleading performance measures and therefore the incomparability of state-of-the-art (SOTA) models. The problems are as follows: (i) *class ambiguity*, in the form of synonyms e.g. testing for `ocean` vs `sea` (ii) *testing on unnormalised collections*, where SOTA models are able to boost annotation performance by promoting popular tags (iii) *low image quality* (iv) *lack of image meta-data* (v) *lack of image diversity* (vi) *using location tags as ground truth* (vii) *copyright restrictions.*

For photo tag recommendation, we have identified the *three* problems with the collections used in [18] and [5]: (i) *using crowdsourced ground truths*; only the photographer of an image understands the true content and context of an image (ii) *synonyms in the ground truth*; models which promote synonyms in their suggestions are promoted over those models which suggest diverse recommendations, (iii) *lack of distribution;* currently tag recommendation works test on their own private collection.

There are two major contributions in this paper: firstly, the identification and elaboration of the problems in current evaluation test sets, and secondly, the introduction of two new freely available image evaluation collections, namely Flickr-AIA and Flickr-PTR[4], which aim to overcome these discussed issues for the fair evaluation of image annotation and tag recommendation models. The rest of this paper is as follows. Section 2 collates the related work in AIA, PTR and its evaluation. In Section 3 we detail further the problems associated in automatic image annotation evaluation before introducing Flickr-AIA. In Section 4 we detail further the problems associated in photo tag recommendation evaluation before introducing Flickr-PTR. Finally, we conclude in Section 5.

2 Background Work

Automatic Image Annotation: The area of AIA has been a well researched area in the last decade [3,10,12]. Firstly, Duygulu *et al.* [3] used a machine translation approach, between image contents and annotations, which was tested on the Corel5k image collection. Joen *et al.* [10] adopted the Cross-Media Relevance Models (CMRM) to predict the probability of generating a word given blobs in an image in the training set. More recently, Makadia *et al.* [12] showed that five existing models could be outperformed by adopting a K-nearest neighbour approach (KNN) trained on colour and texture image features. Despite the progress made in the field, all of these models are evaluated on *small*, *unrealis-*

[3] The collections were: Corel5k, Corel30k, ESP Game, IAPR, Google Images, LabelMe, Washington Collection, Caltech, TrecVid 2007, Pascal 2007, MiAlbum & 4 other small collections.

[4] Both collections are available for download at `http://dcs.gla.ac.uk/~philip/`

tic and *unnormalised* image collections [2,20,8,1]. In this work we introduce the Flickr-AIA collection which aims to overcome these, and other, issues.

AIA Evaluation: A number of issues associated with the evaluation of AIA models have been identified in a number of previous works: Westerveld *et al.* [20] highlighted a number of problems with the Corel collection, such as the fact that images are grouped into *coherent themes*, resulting in misleadingly high performance measures. Athanasakos *et al.* [1] compared two existing models showing that the high performance reported was more to do with the evaluation scheme and test set instead of the approach itself. Müller *et al.* [16] highlighted issues with using the Corel image collection, in that many models test on a different subset of this collection resulting in different performance measures. In this paper we discuss new biases, resolving them in new evaluation collections.

Recently, AIA models have been tested on much larger image collections. Deng *et al.* [2] introduced the ImageNet collection, consisting of 3.2M images (which is constantly being extended), structured into synonym sets of the lexical databse, WordNet [15]. Huiskes *et al.* [13] introduced two Flickr collections of 25K [8] and 1M images [13]. However, these collections are not setup with defined train/test subsets for annotation or tag recommendation evaluation. Further, these collections fail to address a number of the issues presented in this paper such as tag ambiguity and normalisation. Despite the increase in the availability of computation power, in the forms of clusters and multi-core machines, the computationally intensive task of image annotation on this volume of images is out of the reach of many, and therefore a more manageable collection is desirable for most. Additionally, the large size of these collections increases the amount of *noise* and *synonyms* present, ultimately increasing the potential bias in evaluation, as well as the difficulty in its *distribution*. In the Flickr-AIA set, we reduce a large collection of 2M images, to a much smaller collection of high quality, well tagged images, free of *synonyms*. Thus, maintaining the diversity of these large, online collections, whilst allowing for their easy distribution.

Photo Tag Recommendation: Tag recommendation systems have been proposed in literature, which recommend tags based on those tags already present within an image. Sigurbjornsson *et al.* proposed a tag recommendation strategy to support users annotating photos on Flickr [18]. The relationships between tags were exploited to suggest highly co-occurring tags. Garg *et al.* offered personalised tag recommendations [5] in their approach which looked to combine suggestions made from personalised and global tag co-occurrence matrices. In this work we identify flaws in the evaluation procedure of these works, leading to the introduction of Flickr-PTR, a freely available image collection designed for the fair evaluation of tag recommendation models.

3 Automatic Image Annotation Evaluation

The purpose of an image annotation evaluation collection is to benchmark a given annotation method, for a number of image classes or scenes, based purely on its visual discriminatory power. Therefore, these classes should be distinct (and not ambiguous) and easily identifiable by a human being based purely on

Collection	Images	Tags	Ambiguity	Time/Loc	Free	Size	Train	Test	I/T
Corel	5k	374	9.6%	×	×	160px	4.5k	0.5k	88
ESP	22k	269	9.7%	×	✓	156px	20k	2k	377
IAPR	20k	291	12.7%	✓	✓	417px	18k	2k	386
Flickr-AIA	312k	420	0%	✓	✓	719px	292k	20k	2,304

Table 1: Comparison of the Collections (i) Ambiguity: % of tags where there exist at least one synonym (ii) Size: average dimension in pixels (iii) Time/Loc: whether time taken and location details are included (iv) I/T: average # images per tag

their appearance. The images in this collection should reflect real, user images and should cover a diverse range of images for each class; alternatively, the images should be taken in different locations, by different users, in a number of different lighting conditions, on a range of devices. By doing so, annotation models would be benchmarked for as close to a real world scenario as possible. In the following sections, we first introduce three popular annotation collections and the problems they pose for fair evaluation. We introduce an experimental setup which upholds our hypotheses before detailing our new collection which aims to tackle the issues presented.

Existing Collections: We consider the following collections: *Corel* [3], *ESP Game* [19] and *IAPR* [6]. These collections are selected as they have been used to benchmark many AIA models of recent years [12]. We use the same *methods*, *training* and *test* subsets as used in [12]. These collections, along with the collection introduced in this paper (Flickr-AIA), summarized in Table 1.

Although this list does not cover all evaluation collections, they are amongst some of the most popular collections [12,10,1]. One popular collection which has been omitted and is related to this work is the MIR-Flickr 25k [8] and 1M [13] collections. We have not considered these collections as they are not setup with annotation evaluation in mind; they contain user tags rather than high level, visual, classes. However, these collections have been used in the ImageCLEF 2009 annotation task, where the referred 25k collection was annotated using a crowdsourced experiment. Despite this, the collection was only made available for the participants in this task and is no longer publicly available. Therefore, researchers are unable to compare new annotation approaches on this testbed. Additionally, a collection of 25k images, is too small by modern standards. In this work we introduce a larger collection for AIA evalation which is freely available.

3.1 Annotation Model

To demonstrate the issues with the given collections, we conduct a number of experiments using the annotation model described in [12]. The method models the problem of image annotation as that of image retrieval using a KNN ($K = 10$, as used in [12]) approach. Seven features are extracted from images, three colour histograms in three channels (RGB, HSV and LAB), two texture descriptors (HAAR and Gabor filters) and two quantized versions of the texture features. Each feature vectors is normalised, with visual similarity between images computed using the average of the seven distances (for each feature pair).

Each distance is scaled by its *maximum distance*, for the given feature, within the training set. The L_1 distance is used for all features, apart from the LAB descriptor, where the K-L divergence measure is used. N tags ($N = 5$, as used in [12]) are transferred from the nearest neighbour (ordered by frequency in the training set). If the number of tags in the nearest neighbour is $< N$, tags are transferred from the surrounding neighbourhood. The top tags, ranked by the *product* of tag occurrence in the neighbourhood and co-occurrence with the nearest neighbour, are selected. This model is used to highlight problems with testing on unnormalised collections. Firstly we introduce the problems with existing collections in the following section.

3.2 Problems

(1) Tag Ambiguity: One of the major problems with these collections concerns the classes they use as ground truth. All three collections contain synonyms (e.g. `america`/`usa`) or visually identical classes (e.g. `sea`/`ocean`). For the purposes of generic image annotation, a model should not have to differentiate between synonyms, as often (from analysing the visual contents), this is impossible e.g. consider, as a human, differentiating between an image of the `sea` or the `ocean`. To illustrate this problem, we use WordNet [15] to classify keyword pairs as synonyms i.e. those keywords which contain a common synonym set. After a list of potential synonyms is generated, pairs which are seen to be incorrect by an assessor (e.g. `ball`/`globe`) are removed.

Using this approach we identify 36, 26 and 37 *ambiguous tags* (i.e those tags which have at least one synonym) for the Corel, ESP and IAPR collections, respectively. Figure 1 highlights the percentage of *ambiguous tags* present in each collection. Around *one* in *ten* tags in each collection is deemed ambiguous. This equates to 15% of all photo annotations in the IAPR collection meaning a model may under perform by up to 15%, as for each *ambiguous* annotation in the ground truth, the model may predict the synonym. Therefore evaluating on these collections may result in misleading performance measures. For example, if an image's ground truth is [`home`, `sea`] and it is annotated with the tags [`house`, `ocean`] it will achieve precision and recall scores of 0. This is clearly a bias experimental framework as luck plays a *major* role in the scoring of evaluation measures. Table 2 summarises the most occurring synonyms pairs.

Collection	**Top Synonym Pairs*(Instances)***
Corel	field/lawn, field/plain, polar/arctic, ice/frost, ocean/sea
ESP	circle/ring, home/house, rock/stone, baby/child, child/kid
IAPR	woman/adult, building/skyscraper, rock/stone, bush/shrub

Table 2: Top synonyms for each collections

(2) Unnormalised Collections: One of the main issues with the evaluation of existing annotation models lies in the unbalanced nature of collections. By nature, the classes used in image collections follow a long tail distribution i.e. there exist a few *popular* tags and many *unpopular* tags. For the evaluation of annotation models, this leads to a bias experimental setup for two reasons: (i) *Selection*

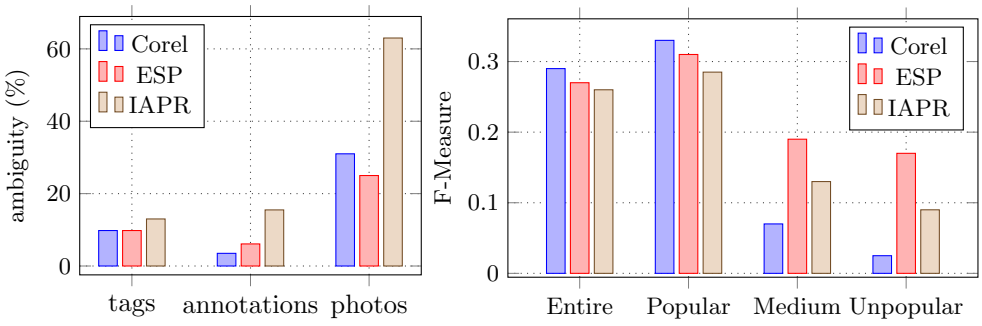

Fig. 1: *(Left)* Ambiguous tags: those tags which have at least one synonym. Ambiguous annotations: those tags assigned to images which have at least one synonym. Ambiguous photos: photos containing at least one ambiguous tag. *(Right)* Normalised annotation for each collection.

Bias: Popular tags exist in more training and test images. Therefore, annotation models are more likely to test their annotation model on these keywords, purely because a popular tag is more likely to exist in a random test image than an unpopular tag. (ii) *Prediction Bias*: Due to the wealth of training data available for popular keywords, annotations models are more likely to annotate images with these tags, as they are more likely to be correct. The unbalanced nature of collections therefore allows for potential *"cheating"* where models promote popular tags over less popular tags. To fairly measure a model's annotation accuracy based *purely on visual content*, models should not be able to exploit attributes of collections, such as tag popularity.

To demonstrate the hypothesis that popular keywords can be exploited to increase annotation accuracy, we split each collection into three vocabulary subsets representing the *popular*, *medium frequency* and *unpopular* tag sets. We denote the full vocabulary as *entire*. We select each subset so that each contains *approximately* the same number of keywords (i.e. one third), from the overall vocabulary. Using the annotation model described in Section 3.1, we annotate the images in each collection three times, annotating only with tags in each tag subset. Precision and recall measures are then computed against the tags in the ground truth, which *exist* in the given subset.

Figure 1 shows the results of this experiment. We observe that *popular* keywords are easier to annotate than *less popular* tags. Additionally, when we annotate the images purely with popular tags, we achieve higher results than the collection as a whole. Therefore, models may exploit this collection characteristic by promoting popular tags, leading to higher than expected measures for precision and recall. This annotation trend is observed across all collections.

It may be argued that by normalising, we are creating an *unrealistic* test set. However, if AIA models are benchmarked *purely* on visual features, we are measuring a model's *true* discriminative visual annotation power, without the bias of promoting popular tags. In our test collection, we propose two ground truths, an unnormalised (real life) and normalised version. We hypothesise that by improving annotation accuracy on the normalised ground truth, we will improve a model's visual discriminatory power, thus increasing accuracy on a real life collection. We encourage researchers to report evaluation metrics on both

ground truths to ensure a model is not exploiting the long tail distribution and is annotating well on visual content.

(3) Quality of Images: The small size and poor quality of images in many collections often make it difficult to extract semantics from the visual contents of images, due to the lack of resolution and visual artefacts present. Despite this, the images contained in modern evaluation collections are often very small (see Table 1). The quality and size of images used in evaluation collections must increase to reflect those images taken on modern smart-phones and digital cameras.

(4) Lack of Meta-data: AIA is being more recently viewed from an information retrieval perspective, rather than that of content analysis, where time and location [14] are being exploited in the image annotation process. Despite this, all the collections used fail to include time, location and user meta-data. Therefore to allow deeper contextual analysis of images in the annotation process, every detail of an image's meta-data should be made available.

(5) Lack of Diversity: Images in the described collections are often taken by the same user, in the same place, of the same scene/object, using the same camera [20]. This leads to natural clustering in image collections, making annotation easier due to high inter-cluster visual similarity. This also causes problems such as duplicate images in the test and train set, making annotation easier.

(6) Identifying Location: As highlighted by Huiskes *et al.*, identifying a location from an image is often impossible [8]. Despite this, two of the three image collections contain ground truth classes which are locations (e.g. `scotland`).

(7) Copyright: The most popular baseline collection, Corel, is not freely available and is bound by copyright. To allow for the easy comparison of annotation models, a collection should be at least *free* and *distributable*.

3.3 Flickr-AIA

In this following section we detail the process used to build the Flickr-AIA collection, which aims to resolve these problems. In total, we present two test collection ground truths for 20k images, one with a normalised ground truth (i.e. where the image classes contain roughly the same number of test images), and one without (i.e. a real life scenario). We refer to how we address each problem by referencing the problem number in parenthesis e.g. (1).

Building the collection: The dataset is collected by querying Flickr for 2k popular nouns extracted from WordNet [15] (categorised as `animal, artifact, body, food, plant, substance`). The top 2k images, which contain the creative commons license, (7) location, user and time meta-data (4) and at least one tag, for each search are then considered for use in our collection. Using this approach, we collect images covering a wide range of topics (5). We download the "largest" available size version (not the original) for each image (3), ensuring high resolution and small file size.

Initially we collect 2M images before a number of pre-processing stages are taken out to resolve the discussed issues. As ground truth we use the tags assigned by the Flickr users; this has a number of advantages and disadvantages. By using user annotations, we are able to collect a *large* number of images, in comparison to the manually collated ground truths used in the Corel and IAPR

collections. However, user tagging is often *noisy*, where tags do not refer to the visual contents of an image. In order to remove these tags deemed irrelevant for image annotation we use the following approach:

Removing Noise: Firstly, we manually removed (using three assessors) those tags which fell into the following categories: camera meta data (e.g. `d60`), Flickr awards (e.g. `excellentphotograph`) and Flickr groups (e.g. `5photosaday`), from the top 1,000 most frequently occurring tags. After removal of these redundant keywords we consider only the top 500 tags, ranked by descending number of users, for use in the collection. This removes tags which are used by only a few users (i.e. noise) and keeps popular classes which are more likely to be well known objects/concepts (i.e. potential image classes). We use WordNet to classify the remaining tags. Only *nouns* which are *not* categorised as the `noun.time` or `noun.location` sub-categories are used in the collection (6). By selecting nouns, we consider only visual objects, ignoring concepts difficult to identify e.g. verbs such as `talk`. Time and location tags are omitted as they are also difficult or impossible to annotate based purely on visual content [13] e.g. `Romania` or `2010`.

Promoting Diversity: As identified by [20], previous collections, such as Corel, often cluster images into coherent themes, where image similarity is high. This makes it easier for AIA models as, for every test image, there are likely to be many images in the training set which are almost visually identical. We therefore limit the number of images taken by a user to 20 to promote visual diversity (5).

Removing Synonyms: We remove synonyms in the remaining tag set using the same method as described in Section 3.2, by grouping tags which co-exist in a common WordNet synonym set. 49 synonym pairs are identified and merged (1). The details of the final collection are shown in Table 1.

Test Sets: From this collection, we remove 20k random images for testing purposes, leaving the rest for training. We offer two ground truths to test against for these images (i) *full ground truth* i.e. image contain all the classes remaining after preprocessing (ii) *normalised ground-truth* i.e. only those middle frequency classes are selected (2). Specifically, we select only those tags which occur in the middle third of tags ordered by frequency i.e. tags #140 to #280. By offering this normalised ground truth, we are able to test annotation models based purely on their visual discriminative power, removing the bias from offering popular tags.

4 Photo Tag Recommendation Evaluation

In photo tag recommendation, the typical evaluation approach is to take a small number of tags from an image and attempt to predict the other tags. As predictions are made based on textual features, the range of ground-truth classes can take a larger number of classes than those used in AIA. Differing to that of AIA evaluation, ground truth tags can also refer to both an image's visual content (e.g. an object within the scene) or its context (e.g. its location). In the following sections, we first highlight problems with test collections used by two existing tag recommendation methods. Finally we detail our new collection, Flickr-PTR, which is built for the purposes of tag recommendation evaluation in mind.

Collection	# Training	# Test	Tags	Freely Available	Ground Truth
Sigurbjornsson	52M	331	3.7M	×	Crowdsourced
Garg	50M	9k	-	×	User Tags
Flickr-PTR	2M	1k	1M	✓	Clustered User Tags

Table 3: Comparison of the Collections (i) I/T = average # images per tag (ii) T/I = averages # tags per image

4.1 Existing Collections

In this work we consider the evaluation collections for tag recommendation used by Sigurbjornsson *et al.* [18] and Garg *et al.* [5]. Unfortunately these collections are not freely available making any analysis or comparison with our collection difficult; however, we detail what is described in the respective papers, along with details of our new collection, Flickr-PTR, in Table 3. Firstly, we identify a number of problems with these collections:

4.2 Problems

(1) Crowdsourced ground-truths: The test collection used in [18] compares predictions against a crowdsourced ground truth for 331 images. We agree with [5], that the ground truth of an image can only be identified by the user whom the photograph is taken by. For example, consider an image taken by a father at their son's soccer game: only the father will know the location, team name etc. Therefore, an approach which tags images using a crowdsourced experiment will result in substandard annotations. Garg *et al.* follow this notion by adopting user tags as image ground truth, however, we identify an issue with this approach which may give mis-leading results, as described in the following subsection.

(2) Synonyms: One of the issues with using user tags is that, by nature, users tend to tag images with multiple synonyms in order make their image searchable for the various versions of the same entity. For example, instead of tagging an image solely `newyork`, many images also include a number of synonym tags e.g. `ny`, `nyc` and `newyorkcity`. In our collection, 52%, 43% and 35% of images tagged with `newyork` are also tagged with `nyc`, `ny` and `newyorkcity`, respectively.

This poses evaluation problems where models which simply promote synonyms achieve higher precision/recall scores than a model which promotes tag *novelty* and *diversity* in their rankings. For example, a recommendation model which suggests [`nyc`, `newyorkcity`, `ny`] may achieve a higher recommendation accuracy than a model which suggests [`taxi, street, centralpark`], due to the number of synonyms in the image ground-truth. However, we consider the recommendations made by the later more useful to the user as they offer novel tags instead of synonyms of already defined concepts. In this paper, we address this problem by clustering the tags in user images into related aspects, allowing for intent-aware metrics to be computed (e.g. αnDCG) instead of the traditional precision/recall metrics which ignore diversity.

(3) Free Distribution: One of the largest problems with these collections is that they are not available for distribution, making comparison with new recommendation method difficult. In our work, we download a manageable number of Flickr images which use the creative commons license, allowing for distribution.

4.3 Flickr-PTR

In this following section we detail the process used to build the Flickr-PTR collection, which aims to resolve these problems. As before, we refer to how we address each problem by referencing the problem number in parenthesis e.g. (1). The *training* collection contains the same 2M creative commons (3) images as in Flickr-AIA, before any preprocessing is taken out, using user tags as ground truth (1). The role of a training set in tag recommendation differs from image annotation, in that images can be categorised with a wide range of tags, whereas images in an AIA training set are categorised for a small number of visual classes. Therefore, for Flickr-PTR, we chose *not* to remove the noisy tags from the collection. Therefore, PTR models can be evaluated for a real-life scenario. Our main contribution, however, lies in our test collection, where tags are clustered into coherent aspects. In order to overcome the discussed problems with synonyms, we cluster tags which describe the same aspect of 1000 random images using a crowdsourced experiment. By doing so, we are able to build a test collection where the ground truth describes *aspects* for each image (2), rather than tags, as required for diversification evaluation.

Crowd-sourcing (i.e. outsourcing a task to a network of online workers) experiments have grown in popularity in recent years [7,4] and have been adopted to carry out tasks which are often difficult for computers but easy for humans e.g. image classification [17,2]. Recently, Nowak *et al.* showed that by using a majority voting scheme for an image annotation task, the quality of Turker judgements were in-line with those made by experts [17]. The ImageNet collection was also built using a crowd-sourced experiment where internet images were mapped to WordNet nodes [2]. In our work, we instead use the crowd to cluster related tags which are already assigned to images.

Task Description: We conduct this experiment on the popular Amazon Mechanical Turk[5] platform. On this platform, human intelligence tasks (HITs) are taken out by workers called 'Turkers'. In our experiment, only those Turkers with the *Master Qualification*[6] are able to accept our HIT. On acceptance of our HIT, users are presented with the following task description:

- **What is required of you:** You will be presented with an image with the tags describing its contents. You must group the *synonyms* or the tags which refer to the *same aspect* of the image.
- **Details:** You will be presented with 20 images. You may skip up to 3 images. You have a maximum of 45 minutes to finish the experiment. To group tags, simply click and drag them into the displayed boxes, then click submit. All of the tags must belong to one group, and every group must contain at least one tag. This experiment is supported for Firefox and Chrome (Res 1024+).
- **Finally:** You must judge at least 17 images and be a *fluent English* speaker. You can do the experiment multiple times, although you must make sure to login if you are coming back.

[5] https://www.mturk.com/

[6] 'Workers who have demonstrated excellence in a type of HIT, for instance categorization, are awarded the Master Qualification'

The interface allows for users to easily click and drag each tag into a number of clusters. The user is able to define one of more clusters (up to the number of tags) for each image. A *video* tutorial and two *example images* accompany the task description, allowing the worker to fully understand what is expected of them before accepting the HIT. Turkers are paid if they agreed to and carried out the conditions of the experiment. On acceptance of these terms, the worker is presented with a registration questionnaire asking for the following details: TurkID, age, sex, occupation, education level and proficiency in English.

Ensuring Quality: One of the major problems with crowdsourcing, however, is that workers often spam or try to complete tasks with as little effort as possible in order to maximize their profits [7,4]. This can lead to poor quality submissions. Many existing works have resolved this problem by introducing a number of *'honeypots'* [7,4] i.e. tasks where the correct 'answers' are already known. In our experiments we therefore introduce a number of honeypot images, which aimed to identify spamming users. Specifically, for every 20 images, we present the user with *three* images where the tags had been pre-grouped by an expert. Care is taken in creating these clusters to ensure that there is no ambiguity in the groupings. Creating these honeypots allows us to indicate the users whom completed the HIT without *reasonable effort*. Any user which grouped the tags of these *honeypot* images differently than the expert is blocked and their work is discarded. Further, the work of any Turker whom describes their English level as less than fluent is also removed. Finally, each image has its tags clustered by three *different* workers, allowing clusters to be computed using an aggregation scheme (as described in the following Section), thus minimizing spam.

Cluster Aggregation: As three different Turkers cluster the tags of each image, the votes from each are aggregated using a majority voting scheme, as adopted by [17,9]. Two tags are grouped if they are grouped in the *same cluster* by the *majority* of the three users i.e. two or more. The clusters are iteratively built, where clusters are merged if they contain a *common co-occurring* tag.

Workers: In total 197 different Turkers accepted the hit, with 20 users failing to pass the honeypot test. Therefore, work is accepted from 177 Turkers. Each HIT (20 images) is completed in 23 minutes and 40 seconds, on average. Turkers are paid between \$1 and \$3 for their work, which equates to \$5.98/hour on average. From the entry questionnaire, around 70% of users say English is their *first language* and around 30% describe their English proficiency as *fluent*. Further, 49% of Turkers are female and 51% male, with an average age of 34. Finally, 80% of workers describe their education level as 'college' or higher.

Summary: After aggregation, each image in our test collection contains around 9.87 clusters (with each containing around 2.18 tags), on average. Considering that images in our test set are annotated with 21.5 tags on average, this indicates that more than half of the tags in our test collection are deemed *redundant* (if we assume that each tag in a cluster describes a single aspect of an image).

5 Conclusion and Future Work

This paper highlighted a number of problems which exist in using *three* popular image annotation and *two* popular photo tag recommendation evaluation col-

lections. Most importantly, synonyms exist in annotation ground truths for all collections, which may result in misleading performance measures. Aside from this, we highlight six additional problems with annotation collections and two additional problems for tag recommendation collections. As a result, we introduce two new collections, namely Flickr-AIA and Flickr-PTR, which aim to overcome these issues and are created with fair evaluation in mind. For each collection, we also include extensive meta data relating on an image's photographer, location and time taken. Future work aims to include state-of-the-art image features and to increase the size of each collection.

References

1. K. Athanasakos, V. Stathopoulos, and J. M. Jose. A Framework for Evaluating Automatic Image annotation Algorithms. *ECIR '10.*
2. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. *IEEE CVPR '09*
3. P. Duygulu et al. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *ECCV 02.*
4. C. Eickhoff and A. P. Vries. Increasing Cheat Robustness of Crowdsourcing Tasks. *Inf. Retr., '13.*
5. N. Garg and I. Weber. Personalized, Tnteractive Tag Recommendation for Flickr. *ACM RecSys '08.*
6. M. Grubinger, P. Clough, H. Mller, and T. Deselaers. The IAPR TC-12 Benchmark - A New Evaluation Resource for Visual Iinformation Systems, 2006.
7. M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cheat-detection Mechanisms for Crowdsourcing. Technical Report, University of Würzburg, 8 2010.
8. M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. *MIR '08.*
9. A. P. D. V. J. Vuurens and C. Eickhoff. How much Spam can you take? An Analysis of Crowdsourcing Results to Increase Accuracy. *ACM SIGIR '11.*
10. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-media Relevance Models. *ACM SIGIR '03.*
11. D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. *WWW '09.*
12. A. Makadia, V. Pavlovic, & S. Kumar. Baselines for image annotation. *IJCV '10.*
13. B. T. Mark J. Huiskes and M. S. Lew. New trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. *MIR '10.*
14. P. J. McParlane, Y. Moshfeghi, and J. M. Jose. On Contextual Photo Tag Recommendation. *SIGIR '13.*
15. G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM, 1995.*
16. H. Müller, S. Marchand-Maillet, and T. Pun. The Truth about Corel - Evaluation in Image Retrieval. *CIVR '02.*
17. S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. *MIR '10.*
18. B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation based on Collective Knowledge. *WWW '08.*
19. L. von Ahn and L. Dabbish. Labeling images with a computer game. *CHI '04.*
20. T. Westerveld and A. P. de Vries. Experimental Evaluation of a generative probabilistic image retrieval model on 'easy' data. *ACM SIGIR '03.*