How Reliable are Annotations via Crowdsourcing?

A Study about Inter-annotator Agreement for Multi-label Image Annotation

Stefanie Nowak
Fraunhofer IDMT
Ehrenbergstr. 31
98693 Ilmenau, Germany
stefanie.nowak@idmt.fraunhofer.de

Stefan Rüger
Knowledge Media Institute
The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
s.rueger@open.ac.uk

ABSTRACT

The creation of golden standard datasets is a costly business. Optimally more than one judgment per document is obtained to ensure a high quality on annotations. In this context, we explore how much annotations from experts differ from each other, how different sets of annotations influence the ranking of systems and if these annotations can be obtained with a crowdsourcing approach. This study is applied to annotations of images with multiple concepts. A subset of the images employed in the latest ImageCLEF Photo Annotation competition was manually annotated by expert annotators and non-experts with Mechanical Turk. The inter-annotator agreement is computed at an image-based and concept-based level using majority vote, accuracy and kappa statistics. Further, the Kendall τ and Kolmogorov-Smirnov correlation test is used to compare the ranking of systems regarding different ground-truths and different evaluation measures in a benchmark scenario. Results show that while the agreement between experts and non-experts varies depending on the measure used, its influence on the ranked lists of the systems is rather small. To sum up, the majority vote applied to generate one annotation set out of several opinions, is able to filter noisy judgments of non-experts to some extent. The resulting annotation set is of comparable quality to the annotations of experts.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Human Factors, Measurement, Performance

Keywords

Inter-annotator Agreement, Crowdsourcing

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

1. INTRODUCTION

In information retrieval and machine learning, golden standard databases play a crucial role. They allow to compare the effectiveness and quality of systems. Depending on the application area, creating large, semantically annotated corpora from scratch is a time and cost consuming activity. Usually experts review the data and perform manual annotations. Often different annotators judge the same data and the inter-annotator agreement is computed among their judgments to ensure quality. Ambiguity of data and task have a direct effect on the agreement factor.

The goal of this work is twofold. First, we investigate how much several sets of expert annotations differ from each other in order to see whether repeated annotation is necessary and if it influences performance ranking in a benchmark scenario. Second, we explore if non-expert annotations are reliable enough to provide ground-truth annotations for a benchmarking campaign. Therefore, four experiments on inter-annotator agreement are conducted applied to the annotation of an image corpus with multiple labels. The dataset used is a subset of the MIR Flickr 25,000 image dataset [12]. 18,000 Flickr photos of this dataset annotated with 53 concepts were utilized in the latest ImageCLEF 2009 Photo Annotation Task [19] in which 19 research teams submitted 74 run configurations. Due to time and cost restrictions most images of this task were annotated by only one expert annotator. We conduct the experiments on a small subset of 99 images. For our experiments, 11 different experts annotated the complete set, so that each image was annotated 11 times. Further, the set was distributed over Amazon Mechanical Turk (MTurk) to non-expert annotators all over the world, who labelled it nine times. The inter-annotator agreement as well as the system ranking for the 74 submissions is calculated by considering each annotation set as single ground-truth.

The remainder of the paper is organized as follows. Sec. 2 describes the related work on obtaining inter-annotator agreements and crowdsourcing approaches for distributed data annotation. Sec. 3 explains the setup of the experiments by illustrating the dataset and the annotation acquisition process. Sec. 4 details the methodology of the experiments and introduces the relevant background. Finally, Sec. 5 presents and discusses the results of the four experiments and we conclude in Sec. 6.

2. RELATED WORK

Over the years a fair amount of work on how to prepare golden standard databases for information retrieval eval-

uation has been published. One important point in assessing ground-truth for databases is to consider the agreement among annotators. The inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators. Kilgarriff [15] proposes guidelines on how to produce a golden standard dataset for benchmarking campaigns for word-sense disambiguation. He concludes that the annotators and the vocabulary used during annotation assessment have to be chosen with care while the resources should be used effectively. Kilgarriff states that it requires more than one person to assign word senses, that one should calculate the inter-annotator agreement and determine whether it is high enough. He identifies three reasons that can lead to ambiguous annotations and suggests ways how to solve them. Basically the reasons lie in the ambiguity of data, poor definition of annotation scheme or mistakes of annotators due to lack of motivation or knowledge.

To assess the subjectivity in ground-truthing in multimedia information retrieval evaluation, several work has been performed on the analysis of inter-annotator agreements. Voorhees [28] analyses the influence of changes in relevance judgments on the evaluation of retrieval results utilizing the Kendall τ correlation coefficient. Volkmer et al. [27] present an approach that integrates multiple judgments in the classification system and compare them to the kappa statistics. Brants proposes in [2] a study about inter-annotator agreement for part-of-speech and structural information annotation in a corpus of German newspapers. He uses the accuracy and F-score between the annotated corpus of two annotators to assess their agreement. A few studies have been performed to study the inter-annotator agreement for word sense disambiguation [26, 5]. These studies often utilize kappa statistics for calculating agreement between judges.

Recently, different works were presented that outsource multimedia annotation tasks to crowdsourcing approaches. According to Howe [10],

> crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.

Often the work is distributed over web-based platforms. Utilizing crowdsourcing approaches for assessing ground-truth corpora is mainly motivated by the reduction of costs and time. The annotation task is divided into small parts and distributed to a large community. Sorokin et al. [25] were one of the first who outsourced image segmentation and labelling tasks to MTurk. The ImageNet database [7] was constructed by utilizing workers at MTurk that validated if images depict the concept of a certain WordNet node.

Some studies have been conducted that explore the annotation qualities obtained with crowdsourcing approaches. Alonso and Mizarro [1] examine how well relevance judgments for the TREC topic about space program can be fulfilled by workers at MTurk. The relevance of a document had to be judged regarding this topic and the authors compared the results of the non-experts to the relevance assessment of TREC. They found that the annotations among non-expert and TREC assessors are of comparable quality. Hsueh et al. [11] compare the annotation quality of sentiment in political blog snippets from a crowdsourcing approach and expert annotators. They define three criteria, the noise level, the sentiment ambiguity, and the lexical uncertainty, that can be used to identify high quality annotations. Snow et al. [24] investigate the annotation quality for non-expert annotators in five natural language tasks. They found that a small number of non-expert annotations per item yields to equal performance to an expert annotator and propose to model the bias and reliability of individual workers for an automatic noise correction algorithm. Kazai and Milic-Frayling [13] examine measures to obtain the quality of collected relevance assessments. They point to several issues like topic and content familiarity, dwell time, agreement or comments of workers that can be used to derive a trust weight for judgments. Other work deals with how to verify crowdsourced annotations [4], how to deal with several noisy labellers [23, 8] and how to balance pricing for crowdsourcing [9].

Following the work of [25, 7], we obtained annotations for images utilizing MTurk. In our experiments, these annotations are acquired on an image-based level for a multi-label scenario and compared to expert annotations. Extending the work that was performed on inter-annotator agreement [1, 2], we do not just analyse the inter-rater agreement, but study the effect of multiple annotation sets on the ranking of systems in a benchmark scenario.

3. EXPERIMENTAL SETUP

In this section, we describe the setup of our experiments. First, the dataset used for the experiments on annotator agreements is briefly explained. Next, the process of obtaining expert annotations is illustrated by outlining the design of our annotation tool and the task the experts had to perform. Following, the acquisition process of obtaining ground-truth from MTurk is detailed. Finally, the workflow of posing tasks at Amazon MTurk, designing the annotation template, obtaining and filtering the results is highlighted.

3.1 Dataset

The experiments are conducted on a subset of 99 images from the MIR Flickr Image Dataset [12]. The MIR Flickr Image Dataset consists of 25,000 Flickr images. It was utilized for a multi-label image annotation task at the latest ImageCLEF 2009 [19] competition. Altogether, 18,000 of the images were annotated with 53 visual concepts by expert annotators of the Fraunhofer IDMT research staff. 5,000 images with annotations were provided as training set and the performance of the annotation systems was evaluated on 13,000 images. 19 research teams submitted a total of 74 run configurations. The 99 images utilized in our experiments on inter-annotator agreements and its influence on system ranking are part of the testset of the Photo Annotation Task. Consequently, the results of 74 system configurations in automated annotation of these images can serve as basis for investigating the influence on ranking.

3.2 Collecting Data of Expert Annotators

The set of 99 images was annotated by 11 expert annotators from the Fraunhofer IDMT research staff with 53 concepts. We provided the expert annotators a definition of each concept including example photos (see [18] for a detailed description of the concepts.). The 53 concepts to be annotated per image were ordered into several categories. In principle, there were two different kinds of concepts, optional concepts and mutual exclusive concepts. E.g. the category

Place contains three mutual exclusive concepts, namely *Indoor*, *Outdoor* and *No Visual Place*. In contrast several optional concepts belong to the category *Landscape Elements*. The task of the annotators was to choose exactly one concept for categories with mutual exclusive concepts and to select all applicable concepts for optional designed concepts. All photos were annotated at an image-based level. The annotator tagged the whole image with all applicable concepts and then continued with the next image.

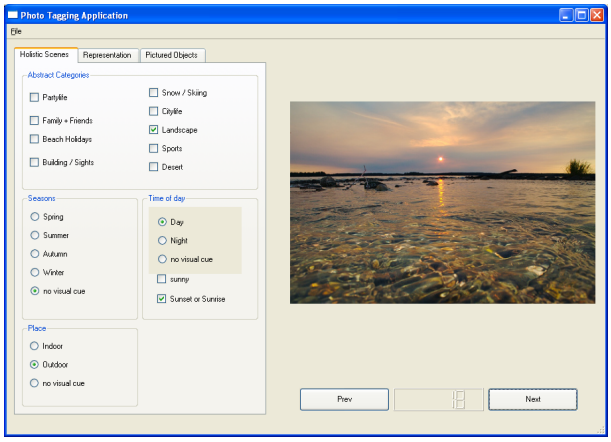

Figure 1: Annotation tool that was used for the acquisition of expert annotations.

Fig. 1 shows the annotation tool that was delivered to the annotators. The categories are ordered into the three tabs *Holistic Scenes*, *Representation* and *Pictured Objects*. All optional concepts are represented as check boxes and the mutual exclusive concepts are modelled as radio button groups. The tool verifies if for each category containing mutual exclusive concepts exactly one was selected before storing the annotations and presenting the next image.

3.3 Collecting Data of Non-expert Annotators

The same set of images that was used for the expert annotators, was distributed over the online marketplace Amazon Mechanical Turk (`www.mturk.com`) and annotated by non-experts in form of mini-jobs. At MTurk these mini-jobs are called HITs (Human Intelligence Tasks). They represent a small piece of work with an allocated price and completion time. The workers at MTurk, called turkers, can choose the HITs they would like to perform and submit the results to MTurk. The requester of the work collects all results from MTurk after they are completed. The workflow of a requester can be described as follows: 1) design a HIT template, 2) distribute the work and fetch results and 3) approve or reject work from turkers. For the design of the HITs, MTurk offers support by providing a web interface, command line tools and developer APIs. The requester can define how many assignments per HIT are needed, how much time is allotted to each HIT and how much to pay per HIT. MTurk offers several ways of assuring quality. Optionally the turkers can be asked to pass a qualification test before working on HITs, multiple workers can be assigned the same HIT and requesters can reject work in case the HITs were not finished correctly. The HIT approval rate each turker achieves by completing HITs can be used as a threshold for authorisation to work.

3.3.1 Design of HIT Template

The design of the HITs at MTurk for the image annotation task is similar to the annotation tool that was provided to the expert annotators (see Sec. 3.2). Each HIT consists of the annotation of one image with all applicable 53 concepts. It is arranged as a question survey and structured into three sections. The section *Scene Description* and the section *Representation* each contain four questions, the section *Pictured Objects* consists of three questions. In front of each section the image to be annotated is presented. The repetition of the image ensures that the turker can see it while answering the questions without scrolling to the top of the document. Fig. 2 illustrates the questions for the section *Representation*.

Representation

1. How is the image illuminated? (Choose the most applicable)

   ○ Overexposed ○ Underexposed ○ Neutral Illumination

2. Is the image blurred? (Choose the most applicable)

   ○ Motion Blur ○ Out of focus ○ Partly Blurred / Depth of focus ○ No Blur

3. How is the content of the image represented? (choose all applicable)

   □ Portrait □ Macro Image □ Still Life □ Canvas

4. Is the image ...? (choose all applicable)

   □ of a high grade of overall quality? □ aesthetic? □ fancy?

Figure 2: Section *Representation* of the survey.

The turkers see a screen with instructions and the task to fulfil when they start working. As a consequence, the guidelines should be very short and easy to understand. In the annotation experiment the following annotation guidelines were posted to the turkers. These annotation guidelines are far shorter than the guidelines for the expert annotators and do not contain example images.

- Selected concepts should be representative for the content or representation of the whole image.
- Radio Button concepts exclude each other. Please annotate with exactly one radio button concept per question.
- Check Box concepts represent optional concepts. Please choose all applicable concepts for an image.
- Please make sure that the information is visually depicted in the images (no meta-knowledge)!

3.3.2 Characteristics of Results

The experiment at MTurk was conducted in two phases. In the first phase (also considered as test or validation phase), each of the 99 HITs was assigned five times, which resulted in 495 annotation sets (five annotation sets per image). One HIT was rewarded with 5 Cent. The second phase consists of the same 99 HITs that were annotated four times by the turkers. So altogether 891 annotation sets were obtained. The work of all turkers that did not follow the annotation rules was rejected. As the review of the annotation correctness is difficult and subjective, the rejection process was conducted on a syntactical level. Basically all images in which at least one radio button group was not annotated was rejected, as this clearly violates the annotation guidelines. Overall 94 annotation sets were rejected that belong to 54 different images (see Fig. 3(a)). In maximum for one image five HITs and for two others four HITs were rejected. Looking at the images in Fig. 3(a), no obvious reason can be found why so many turkers did not annotate all categories in these images. Fig. 3(b) illustrates the amount of images that were annotated per turker. The turkers are represented at the x-axis starting from the turker with most completed HITs to the turker with least completed HITs for both batches.

Statistics of first Batch.

The first batch, consisting of 495 HITs, was completed in about 6 hours and 24 minutes. In total 58 turkers worked on the annotations and spent in average 156.4 seconds per HIT. 58 annotation sets were rejected which corresponds to 11.72% of the batch. The time spent to annotate these images was in average 144.4 seconds, which does not substantially differ from the overall average working time. In all rejected images at least one category with mutual exclusive concepts was not annotated at all.

The first round also served as validation phase. Results were analysed to check whether there was a misconception in the task. In the survey, the category *Time of Day* consists as only category of mutual exclusive and optional concepts at the same time. The turker should choose one answer out of the radio buttons *Day*, *Night* and *No visual time* and optionally could select the concepts *Sunny* and *Sunset or Sunrise*. For this category, it seemed not clear to everybody that one radio button concept had to be selected. As a result the description for this category was rendered more precisely for the second round. The rest of the survey remained unchanged.

Statistics of second Batch.

The second batch was published four days after the first. Its 396 HITs were completed in 2 hours and 32 minutes by 38 workers. 9.09% of the HITs had to be rejected which is equal to 36 HITs. In average 137.5 seconds were needed for the annotation of one image. The rejected images were annotated in 117.8 seconds in average. Six workers worked on both batches. MTurk arranges that several assignments per HIT are not finished by the same turkers. However, as the second batch was published as a new task some days later, it was possible that the same turkers of the first round also worked on the second. All in all, there were 13 images that were annotated twice by the same person.

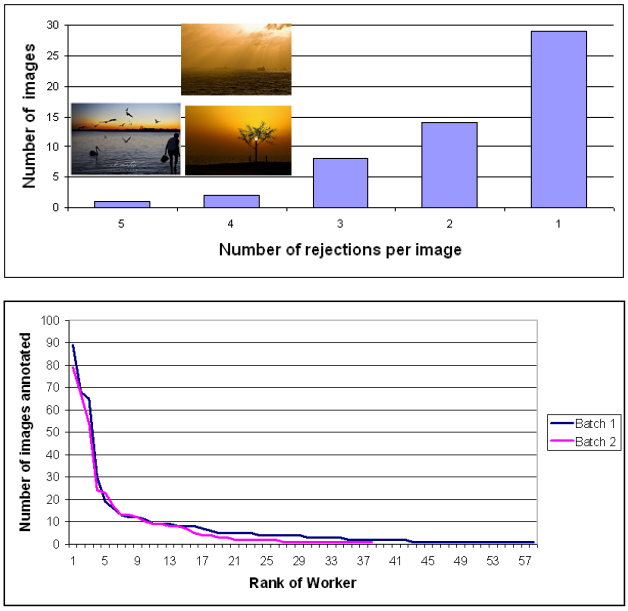

Figure 3: At the top, the number of rejections per image is plotted for all rejected images. At the bottom the amount of images annotated by each turker is illustrated.

Feedback.

Each HIT was provided with a comment field for feedback. The comments received can be classified into 1) comments about work, 2) comments about the content of the photo, 3) comments about the quality of the photo, 4) comments about feelings concerning a photo and 5) other comments. In Table 1 an excerpt of the comments is listed. Notably, no negative comment was posted and the turkers seemed to enjoy their task.

4. EVALUATION DESIGN

In this section, the methodology of the inter-annotator agreement experiments is outlined and the utilized measures are introduced. Four experiments are conducted to assess the influence of expert annotators on the system ranking and whether the annotation quality of non-expert annotators is good enough to be utilized in benchmarking campaigns:

1. Analysis of the agreement among experts

   There are different possibilities to assess the inter-rater agreement among annotators in case each annotator annotated a whole set of images. One way is to calculate the accuracy between two sets of annotations. Another way is to compare the average annotation agreement on a basis of the majority vote for each concept or for each image.

   In the annotation of the images two principal types of concepts were used, optional and mutual exclusive ones. The question is how to assess when an annotator performs a decision:

   - Is a decision just performed by explicitly selecting a concept?

Content of photo	About work	Feelings about photo	Quality of photo	Other
Cupcakes Looks Like a dream land	Dolls aren't persons right really nice to work on this. this is very different and easy. Answer for Picture Objects 2 and 3 are not fitting correctly.	Cute And nice Just Beautiful Thats Creative I really like this one has nice composition.	Color effect can be better Interesting good representation for Logo or....	ha haa Sure For what purpose is this useful?

Table 1: The table depicts an excerpt of the comments posted by the turkers.

- Or does the annotator perform a judgment through the selection and deselection of concepts?

In case of the optional concepts, it is not assured that a deselected concept was chosen to be deselected or just forgotten during the annotation process. In case of the mutual exclusive concepts, the selection of one concept automatically leads to a deselection of the other ones in the same group. In this case, both the selection and deselection of a group of concepts is performed intentionally. The majority analysis of the agreement on images and concepts takes these two paradigms into consideration and compares its results.

2. Influence of different sets of expert annotations on ranking the performance of systems

   In the second experiment, the influence of annotator sets on the performance of systems is determined. The goal is to examine how much different ground-truths affect the ranks of systems in a benchmark scenario. Each set of expert annotations is regarded as ground-truth for the evaluation of annotation systems. The 74 run configurations of the ImageCLEF 2009 Photo Annotation task were trimmed to contain only the annotations for the 99 images. For each run, results against each ground-truth were computed with the evaluation measures Ontology Score (OS), Equal Error Rate (EER) and Area Under Curve (AUC), that were utilized in the official ImageCLEF campaign [19] (see Sec. 4.2). In a second step, the resulting ranked lists per annotator ground-truth are compared to each other with the Kendall τ correlation coefficient [14] and the Kolmogorov-Smirnov statistics [16].

3. Analysis of the agreement between experts and non-experts

   The third experiment analyses the agreement between expert and non-expert annotators. Its goal is to assess if there is a comparable agreement for non-experts to the inter-rater agreement of experts. In general, the annotations obtained from MTurk are organized as HITS. A HIT should cover a small piece of work that is paid with a small reward. As a consequence, the major differences between the expert annotation sets and the ones from MTurk are that at MTurk each set of 99 images is annotated by several persons. This allows to compare the agreement on the labels at a concept- and image-based level, but not to compare the correlation among annotators over the whole set. The analysis of non-expert agreements considers only approved annotation sets from MTurk and uses the annotation sets from both rounds combined.

   In this experiment, the annotation agreement for each image is calculated in terms of example-based accuracy and compared between both groups of annotators. Further, the expert annotation set determined with the majority vote is compared to the combined annotation set of the non-experts. Like in the first agreement experiment, the accuracy serves as the evaluation measure. To evaluate the inter-annotator agreement on a concept basis, the kappa statistics are utilized (see Sec. 4.4). They take all votes from annotators into account and derive an agreement value which excludes the agreement by chance. This experiment is performed for both groups of annotators and the results for the mutual exclusive categories and the optional concepts are compared.

4. Influence of averaged expert annotations compared to averaged non-expert annotations on system ranking

   Finally, the influence of the non-expert annotations on the system ranking is investigated. The combined annotations determined by the majority vote of the non-experts are used as basis for evaluation. The correlation in ranking between this ground-truth and the combined ground-truth of the expert annotators is computed for all three evaluation measures OS, EER and AUC. The Kendall τ correlation coefficient and the Kolmogorov-Smirnov statistics are calculated between both rankings.

In the following, the background needed to understand the experiments is briefly explained. First, the inter-annotator agreement computation based on accuracy is described. Second, the evaluation measures utilized in the ranking experiment are introduced. Next, the calculation of rank correlation is outlined. Finally, the kappa statistics are explained.

4.1 Accuracy for Agreement Assessment

Following [2] the accuracy between two sets of annotated images U and V is defined as

$$\text{accuracy}(U,V) = \frac{\#\text{ identically tagged labels}}{\#\text{ labels in the corpus}}, \quad (1)$$

where *labels* refer to the annotated instance of a concept over the whole set.

The accuracy between two sets of annotations per image can be calculated according to Eq. 2. This way of calculating the accuracy does not presume the existence of two persons that annotated the whole set of images, but evaluates the accuracy of annotations on an image-based level.

$$\text{accuracy}_{ex}(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{\#\text{ identically tagged labels in } X_i}{\#\text{ labels in } X_i}. \quad (2)$$

	A 2	A 3	A 4	A 5	A 6	A 7	A 8	A 9	A 10	A 11	Merged
A 1	.900	.877	.905	.892	.914	.912	.890	.894	.900	.916	.929
A 2		.885	.913	.905	.916	.915	.903	.911	.909	.925	.939
A 3			.900	.873	.902	.884	.878	.886	.892	.904	.918
A 4				.897	.926	.918	.899	.914	.915	.932	.947
A 5					.900	.917	.902	.901	.900	.911	.928
A 6						.925	.902	.918	.925	.932	.952
A 7							.900	.916	.918	.929	.945
A 8								.887	.892	.909	.920
A 9									.918	.918	.941
A 10										.919	.941
A 11											.958

Table 2: The confusion matrix depicts the accuracy among annotators averaged over a set of 99 images with 53 annotations per image. The column *Merged* contains the majority votes of all annotators.

4.2 Image Annotation Evaluation Measures

The performance of systems in the ImageCLEF Photo Annotation task was assessed with the three evaluation measures OS, EER and AUC. The OS was proposed in [20]. It assesses the annotation quality on an image basis. The OS considers partial matches between system output and ground-truth and calculates misclassification costs for each missing or wrongly annotated concept per image. The score is based on structure information (distance between concepts in the hierarchy), relationships from the ontology and the agreement between annotators for a concept. The calculation of misclassification costs favours systems that annotate an image with concepts close to the correct ones more than systems that annotate concepts that are far away in the hierarchy from the correct concepts. In contrast, the measures EER and AUC assess the annotation performance of the system per concept. They are calculated out of Receiver Operator Curves (ROC). The EER is defined as the point where the false acceptance rate of a system is equal to the false rejection rate. The AUC value is calculated by summing up the area under the ROC curve. As the EER and the AUC are concept-based measures, they calculate a score per concept which is later averaged. In case a concept was not annotated at least once in the ground-truth of the annotator, it is not considered in the final evaluation score for EER or AUC.

4.3 Rank Correlation

The correlation between ranked lists can be assessed by Kendall's τ [14] or Kolmogorov-Smirnov's D [16]. Both use a non-parametric statistic to measure the degree of correspondence between two rankings and to assess the significance of this correspondence. The Kendall test computes the distance between two rankings as the minimum number of pairwise adjacent swaps to turn one ranking into the other. The distance is normalised by the number of items being ranked such that two identical rankings produce a correlation of +1, the correlation between a ranking and its perfect inverse is -1 and the expected correlation of two rankings chosen randomly is 0. The Kendall statistics assume as null hypothesis that the rankings are discordant and reject the null hypothesis when τ is greater than the $1 - \alpha$ quantile, with α as significance level. In contrast, the Kolmogorov-Smirnov's D [16] states as null hypothesis that the two rankings are concordant. It is sensitive to the extent of disorder in the rankings. Both tests are utilized in our experiment to see how much the different ground-truths affect the ranking of systems. The correlation is regarded as a kind of agreement between the annotators on the annotations, as a high correlation denotes an equal ranking in both lists, which points to a close annotation behaviour.

4.4 Kappa Statistics

Kappa statistics can be utilized to analyse the reliability of the agreement among annotators. It is a statistical measure that was originally proposed by Cohen [6] to compare the agreement between two annotators when they classify assignments into mutual exclusive categories. It calculates the degree of agreement while excluding the probability of consistency that is expected by chance. The coefficient ranges between 0 when the agreement is not better than chance and 1 when there is perfect agreement. In case of systematic disagreement it can also become negative. As a rule of thumb, a kappa value above 0.6 represents an adequate annotator agreement while a value above 0.8 is considered as almost perfect [17]. The kappa statistic used in the following analysis is called free-marginal kappa (see [3, 21]) and can be utilized when the annotators are not forced to assign a certain number of documents to each concept. It is suitable for any number of annotators.

5. RESULTS AND DISCUSSION

This section details the results of the four experiments. The first experiment analyses the agreement between different sets of expert annotations. The second one investigates the influence of the annotation sets on the performance ranking. The third one compares inter-rater agreement between experts and non-experts and the last one considers the influence of non-expert annotations on ranking.

5.1 Agreement Analysis among Experts

For each annotator the accuracy in comparison to the annotations of all other annotators is calculated. Additionally, the majority vote of all annotators is utilized as 12^{th} ground-truth, further denoted as *merged annotations*. Table 2 presents the results in a confusion matrix. In general, the accuracy between the annotations is very high. The overall accuracy is 0.912 with a minimum of 0.873 between annotator 3 and 5 and a maximum of 0.958 between annotator 11 and the merged annotations.

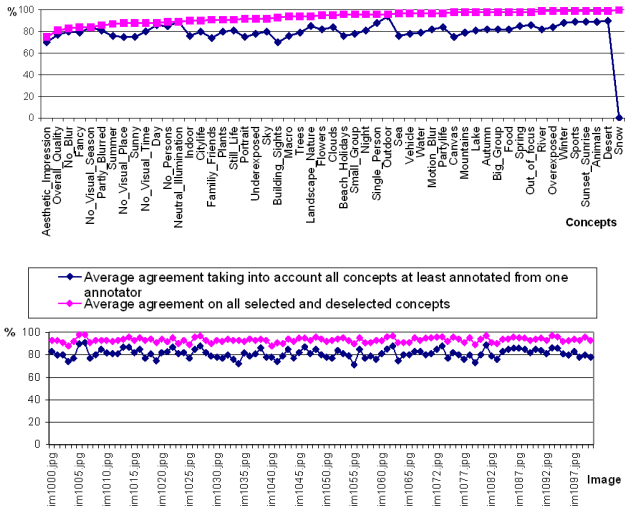

Figure 4: The upper figure depicts the agreement among annotators for each concept determined over the majority vote. The lower diagram shows the inter-annotator agreement for each image.

Fig. 4 presents at the top the agreement for each concept among the eleven annotators. The majority vote determines if a concept has to be annotated for a specific image. The percentage of the annotators that chose this concept is depicted in the figure averaged over all images. Note that by calculating agreements based on the majority vote, the agreement on a concept cannot be worse than 50%. The upper line represents the agreements on a concept averaged over the set of images in case the selection and deselection of concepts is regarded as intentional. This means that if no annotator chose concept C to be annotated in an image X, the agreement is regarded as 100%. The lower line represents the case, when only selected concepts are taken into account. All images in which a concept C was not annotated by at least one annotator are not considered in the averaging process. In case only a small number of annotators select one concept the majority vote determines the agreement and it is considered in the averaging process. This means if e.g. nine out of 11 annotators decided not to select concept C in an image X, the agreement on this concept would be about 82%. For the concept *Snow* the lower line represents an agreement of 0%. There was no annotator that annotated that concept in one of the 99 images. At the bottom of Fig. 4 the agreement among annotators is illustrated averaged for each image. Again, the average per image was calculated based on selected concepts and based on all concepts. The upper line represents the average agreement among annotators for each image when taking into account the selected and deselected concepts. The lower line illustrates the agreement per image when just considering the selected concepts. All in all, the agreement among annotators in case of averaging based on the majority vote shows a mean agreement of 79,6% and 93,3% per concept and 81,2% and 93,3% per image for selected and all concepts, respectively.

5.2 System Ranking with Expert Annotations

In the ranking experiment, the effect of the annotator selection on classification accuracy is investigated. Results are presented in Table 3 which displays the Kendall τ correlation coefficients and the decisions of the Kolmogorov-Smirnov test. The upper triangle depicts the correlations between the ranked result lists by taking into account the different ground-truths of the annotators for the evaluation with the OS measure. In average, there is a correlation of 0.916 between all result lists. The list computed with the merged ground-truth has an overall correlation of 0.927 with the other lists. Despite three cases, the Kolmogorov-Smirnov test supported the decision of concordance in the rankings. Overall, annotator 11 has the highest correlation with all other annotators with 0.939 and annotator 10 has the lowest average correlation of 0.860.

The lower triangle of Table 3 contains the correlation of ranks in the result lists for the evaluation measure EER. All pairs of ranked lists have in average a correlation of 0.883. The ground-truth of the annotators correlates on average with the merged ground-truth in 0.906. For the rankings with EER, the Kolmogorov-Smirnov statistics assigned discordance in rankings in six times. The annotator with the lowest average correlation in its annotations is annotator 1 with 0.85 and the annotations with the highest correlation in average are the ones from annotator 6 with 0.901 (when not considering the merged annotations as having the highest correlation overall).

The results for the evaluation measure AUC are similar to the ones of the OS. The correlation between all runs is on average 0.936. The correlation between the ground-truth of each annotator and the merged ground-truth is on average 0.947. The lowest average correlation with all other annotations are from annotator 10 with 0.916. The highest average correlation could be achieved from annotator 11 with 0.947. In all cases the Kolmogorov-Smirnov statistics supported the Kendall's τ test results for concordance.

Summarizing, the results of two tests showed a high correlation of the ranked lists calculated against the ground-truths of the different expert annotators. Just in a few case the test results were contradicting. Depending on the evaluation measure with which the ranked lists were computed, the average correlation varies from 0.916 (OS), 0.883 (EER) to 0.936 (AUC). One can conclude from these results that the annotators have a high level of agreement and that it does not affect the ranking of the teams substantially which annotator to choose. For the measures OS and AUC the same two annotators, annotator 11 and annotator 10, show the highest and the lowest average correlation with the other annotations respectively.

5.3 Agreement Analysis between Experts and Non-experts

In the following, the results of the agreement analysis of non-expert annotations are presented and compared to the inter-annotator agreement of the experts.

5.3.1 Accuracy

The accuracy was computed for each image X among all annotators of that image. The averaged accuracy for each image annotated by the expert annotators is 0.81. The average accuracy among all turkers for each image is 0.79. A merged ground-truth file was composed from the HITs by computing the majority vote for each concept in each image. The accuracy between the ground-truth file of MTurk and the merged ground-truth file of the expert annotators is 0.92.

	A 1	A 2	A 3	A 4	A 5	A 6	A 7	A 8	A 9	A 10	A 11	Merged
A 1		.938	.938	.914	.938	.884	.959	.952	.890	.816	.927	.890
A 2	.842		.964	.960	.914	.939	.951	.898	.934	.869	.960	.947
A 3	.804	.890		.969	.901	.942	.945	.897	.937	.872	.976	.948
A 4	.874	.898	.872		.892	.967	.939	.878	.959	.892	.976	.973
A 5	.872	.892	.864	.892		.859	.944	.950	.869	.792	.898	.866
A 6	.845	.927	.904	.908	.905		.910	.846	.955	.918	.951	.978
A 7	.863	.880	.872	.893	.895	.906		.928	.910	.841	.948	.917
A 8	.881	.884	.859	.896	.888	.889	.876		.850	.777	.888	.851
A 9	.819	.868	.882	.888	.875	.880	.849	.861		.892	.946	.958
A 10	.847	.867	.861	.919	.874	.881	.869	.851	.893		.872	.914
A 11	.838	.915	.894	.890	.883	.919	.898	.861	.882	.891		.954
Merged	.865	.925	.889	.919	.914	.946	.898	.905	.903	.890	.912	

Table 3: This table presents the Kendall τ correlation for the evaluation with OS measure and varying ground-truth in the upper triangle. The lower triangle shows the correlation coefficient for the evaluation with EER score. The cells coloured in gray represent the combinations for which the Kolmogorov-Smirnov decided on discordance for the rankings.

In both cases, the accuracy between experts and non-experts is very high. Remembering Table 2, the results between the expert annotators and the merged expert annotator results are on average 0.94. In terms of the overall accuracy of the ground-truth files, the annotations from MTurk nearly show as good results as the expert annotators.

5.3.2 Kappa Statistics

The kappa statistics are calculated in three different configurations using [22]. The first configuration (denoted as *non-experts*) takes all votes of the turkers into consideration. The second, called *experts*, uses all annotations from the experts and the third (*combined*) computes the kappa statistics between the averaged expert annotation set and the averaged non-expert annotation set. In the following, the results are presented for the categories with mutual exclusive concepts and the optional concepts.

Kappa Statistics on Mutual Exclusive Concepts.

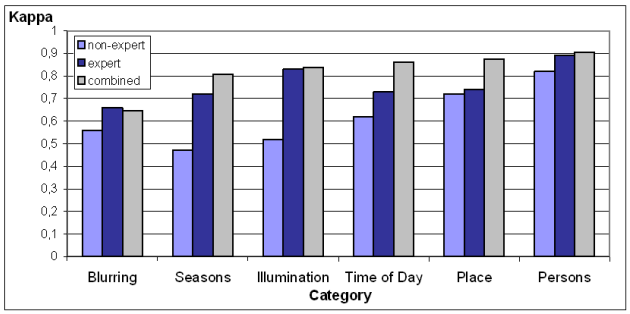

Figure 5: The diagram depicts the kappa values for the mutually exclusive categories.

The images were annotated with six categories that contain mutual exclusive concepts. Fig. 5 presents the results for the kappa analysis for the categories *Blurring*, *Season*, *Time of Day*, *Place*, *Illumination* and *Persons*. *Time of Day* and *Place* contain three concepts, *Blurring*, *Illumination* and *Persons* four concepts and the category *Season* is assigned with five concepts. The results show that the kappa value is higher for the expert annotators for each category. In all categories the expert annotators could achieve a kappa value higher than 0.6. For the categories *Illumination* and *Persons* an agreement higher than 0.8 could be obtained. Considering the non-expert annotations, only for half of the categories the kappa value is above the threshold. The kappa value for the categories *Season* and *Illumination* is indeed quite low. A possible reason for the category *Season* lies in the fact, that the images should only be annotated with concepts that are visible in the image. In most cases the season is not directly visible in an image and the expert annotators were trained to assign the concept *No visual season* in this case. However, the turkers may have guessed, which leads to a lower agreement. The kappa statistics for the combined configuration shows that the majority of the experts has a good agreement to the majority of non-experts. Despite the category *Blurring* all agreements are higher than 0.8.

Kappa Statistics on Optional Concepts.

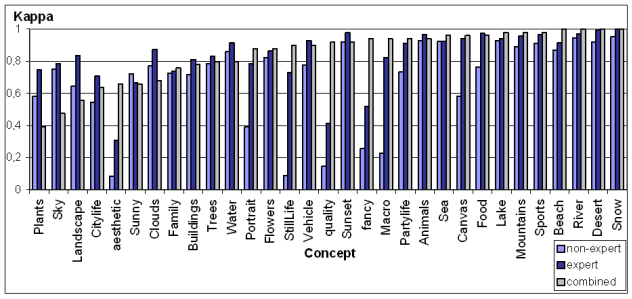

Figure 6: The diagram depicts the kappa values for the optional concepts.

In the dataset, 31 optional concepts were annotated. For each optional concept the kappa statistics are exploited separately in a binary scenario for the described three configurations. Fig. 6 presents the kappa statistics for the optional concepts. On average, the non-expert annotators agree with a value of 0.68, the experts with a value of 0.83 and the combined kappa value is 0.84. For a few concepts (*aesthetic*, *Still Life*, *Quality*, *Macro*...) the non-expert agreement is very

low. However, the combined agreement for these concepts is quite high, also slightly better on average than the agreement among experts. The results indicate that the majority vote is able to filter the noise from the non-expert annotations of most concepts and raise the averaged annotations to the level of the expert annotations. For a few concepts like *Plants*, *Sky* and *Landscape* the agreement among the combined annotation sets is low. These concepts are depicted quite frequently in the images, so apparently there is no major agreement about how to annotate these concepts. For other concepts the agreement in the combined configuration is very high. Taking into account the contents of the images, the reasons for this are twofold. On the one hand, there are some concepts that simply are not often depicted in the 99 images (e.g. the concept *Snow* is not depicted at all). So it is correct, that all annotators agree that the concept is not visible. But the results for these images can not be considered as a general annotation agreement for this concept, as this may change on images that depict these concepts. On the other hand, there is the problem of which annotation paradigm to apply as illustrated in Sec. 4. If just the annotations for images are considered in which at least one annotator selected the concept, the agreement would decrease.

In contrast to the agreement results in terms of accuracy (Sec. 5.3.1), the inter-annotator agreement evaluated with kappa statistics shows major differences between experts and turkers. While the experts could achieve a satisfiable agreement on concepts for most categories and optional concepts, the results of the turkers are not comparable well. They only cross the 0.6 threshold for half of the categories. In case of the optional concepts 22 of 31 concepts have a non-expert agreement higher than 0.6. For other concepts, there exist major differences in agreement.

5.4 Ranking with Non-expert Annotations

This experiment explores how the different ground-truths of expert annotators and turkers affect the ranking of systems in a benchmark scenario. For this experiment the combined annotation sets obtained over the majority vote are utilized as ground-truth for the system evaluation. The Kendall τ test assigns a high correlation in ranking between the combined ground-truth of the turkers and the combined ground-truth of the experts. Evaluated with the OS, the correlation is 0.81, evaluated with EER the correlation is 0.92 and utilizing the AUC measure, the correlation coefficient is 0.93. This corresponds approximately to the correlation coefficient the single expert annotators had in comparison to the combined expert list as illustrated in Sec. 5.2. In that experiment the correlation coefficient for the OS is 0.93, for the EER 0.91 and for the AUC 0.95 on average. Consequently, the ranking of the systems is affected most in case of the OS. For the EER the non-experts annotations show even a higher correlation with the merged list as the single experts annotations. These results are supported by the Kolmogorov-Smirnov test. This test decides for concordance in case of EER and AUC, but for discordance in case of OS.

Following, these results confirm that the majority vote seems to filter some noise out of the annotations of the non-experts. However, as shown in Sec. 5.3.2, for some concepts the agreement is still low. It is surprising that the concept-based measures EER and AUC show such a high average correlation in the ranking of systems, even if there are a few concepts for which the annotations differ substantially among expert and non-expert annotators. These results pose the question whether the evaluation measures used for the evaluation of image annotation are sensitive enough.

6. CONCLUSION AND FUTURE WORK

Summarizing, this paper illustrates different experiments on inter-annotator agreement in assessing ground-truth of multi-labelled images. The annotations of 11 expert annotators were evaluated on a concept- and an image-based level and utilized as ground-truths in a ranking correlation experiment. All expert annotators show a high consistency in annotation with more than 90% agreement in most cases depending on the measure utilized. The kappa statistics show an agreement of 0.76 on average for the exclusive categories and a kappa of 0.83 for the optional concepts. A further experiment analyses the influence of judgments of different annotators on the ranking of annotation systems. This indirect agreement measure exploits the fact that system are ranked equally for the same ground-truth, so a low ranking correlation points to a low agreement in annotation. The results show that a high correlation in ranking is assigned among the expert annotators. All in all, the ranking of systems in a benchmark scenario is in most cases not major influenced by evaluating against different expert annotations. Depending on the evaluation measure used to perform the ranking in a few combinations a discordance between rankings could be detected. This leads to the conclusion that repeated expert annotation of the whole dataset is not necessary, as long as the annotation rules are clearly defined. However, we suggest that the inter-rater agreement is validated on a small set to ensure quality as depending on the concepts also the expert annotation agreement varies.

The same experiment was conducted with non-expert annotators at Amazon Mechanical Turk. Altogether nine annotation sets were gathered from turkers for each image. The inter-annotator agreement was not judged consistently by the different approaches. The accuracy shows a high agreement of 0.92 which is very close to the agreement among expert annotators. However, the kappa statistics report an average of 0.62 for the exclusive categories and an average of 0.68 for the optional concepts for the non-experts. The value of the exclusive concepts is close to the lower threshold for what is regarded as acceptable for annotator agreements. When comparing the averaged ground-truth file of the non-experts with the one from the experts, the inter-annotator agreement in terms of kappa rises to 0.84 on average. The majority vote used for generating this ground-truth file seems to filter some noise out of the annotations of the non-experts. Finally, the ranking experiment shows a high correlation with combined ground-truth of the non-expert annotators in comparison to the one of the expert annotators. These results indicate that the differences in annotations found by the kappa statistics, even with the combined ground-truth, do not major influence the ranking of different systems in a benchmark scenario. This poses new questions concerning the sensitivity of image annotation evaluation measures, especially in the case of concept-based evaluation measures.

Concluding, data annotation utilizing a crowdsourcing approach is very fast and cheap and therefore offers a prospective opportunity for large-scale data annotation. The results obtained in these experiments are quite promising when re-

peated labelling is performed and support results of other studies on distributed data annotation [25, 1, 24]. However, as the experiments were conducted on a small database, future work has to explore if the results remain stable on a larger set. This work does not answer the question how many annotation sets of non-experts are necessary to obtain comparable results to expert annotators. Additionally, further analysis needs to be performed to answer the question why the concept-based evaluation measures for ranking systems in a benchmark scenario do not reflect the differences in annotation quality to a great extent, as the kappa statistics or the OS do. A deeper analysis about which evaluation measures are more sensitive to varying annotation quality will be part of future work.

7. ACKNOWLEDGMENTS

This work has been partly supported by grant 01MQ07017 of the German research program THESEUS which is funded by the Ministry of Economics. The work was partly performed at Knowledge Media Institute at Open University thanks to the German Academic Exchange Service (DAAD) scholarship. Thanks for all annotation effort conducted at Fraunhofer IDMT and for the work of the turkers at Mechanical Turk.

8. REFERENCES

[1] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *SIGIR 2009 Workshop on the Future of IR Evaluation*.

[2] T. Brants. Inter-annotator agreement for a German newspaper corpus. In *Proc. of the 2nd Intern. Conf. on Language Resources and Evaluation*, 2000.

[3] R. Brennan and D. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687, 1981.

[4] K. Chen, C. Wu, Y. Chang, and C. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia*, 2009.

[5] T. Chklovski and R. Mihalcea. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of RANLP 2003*, 2003.

[6] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. CVPR*, pages 710–719, 2009.

[8] P. Donmez, J. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD intern. conference on Knowledge discovery and data mining*, pages 259–268. New York, USA, 2009.

[9] D. Feng and S. Zajac. Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. *ACL-IJCNLP 2009*.

[10] J. Howe. Crowdsourcing - A Definition. `http://crowdsourcing.typepad.com/cs/2006/06/`, last accessed 24.11.2009, 2006.

[11] P. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 2009.

[12] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *Proc. of ACM Intern. Conf. on Multimedia Information Retrieval*, 2008.

[13] G. Kazai and N. Milic-Frayling. On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation*.

[14] M. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30:81–89, 1938.

[15] A. Kilgarriff. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(4):453, 1998.

[16] A. Kolmogoroff. Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari*, 4(1):83–91, 1933.

[17] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[18] S. Nowak and P. Dunker. A Consumer Photo Tagging Ontology: Concepts and Annotations. In *THESEUS-ImageCLEF Pre-Workshop 2009, Corfu, Greece*, 2009.

[19] S. Nowak and P. Dunker. Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. *CLEF working notes 2009, Corfu, Greece*, 2009.

[20] S. Nowak and H. Lukashevich. Multilabel Classification Evaluation using Ontology Information. In *Proc. of IRMLeS Workshop, ESWC, Greece*, 2009.

[21] J. Randolph. Free-Marginal Multirater Kappa (multirater κ free): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In *Joensuu Learning and Instruction Symposium*, 2005.

[22] J. Randolph. Online Kappa Calculator. `http://justus.randolph.name/kappa`, 2008.

[23] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[24] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[25] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008*, 2008.

[26] J. Véronis. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, 1998.

[27] T. Volkmer, J. Thom, and S. Tahaghoghi. Modeling human judgment of digital imagery for multimedia retrieval. *IEEE Trans. on Multimedia*, 9(5):967–974, 2007.

[28] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.