Getting by with a Little Help from the Crowd: Practical Approaches to Social Image Labeling

Babak Loni[1], Jonathon Hare[2], Mihai Georgescu[3], Michael Riegler[1],
Xiaofei Zhu[3], Mohamed Morchid[4], Richard Dufour[4], Martha Larson[1]

[1]Delft University of Technology, Netherlands
[2]University of Southampton, United Kingdom
[3]L3S Research Center, University of Hanover, Germany
[4]LIA, University of Avignon, France

b.loni@tudelft.nl, jsh2@ecs.soton.ac.uk, georgescu@l3s.de, m.a.riegler@tudelft.nl
zhu@l3s.de, {mohamed.morchid, richard.dufour}@univ-avignon.fr, m.a.larson@tudelft.nl

ABSTRACT

Validating user tags helps to refine them, making them more useful for finding images. In the case of interpretation-sensitive tags, however, automatic (i.e., pixel-based) approaches cannot be expected to deliver optimal results. Instead, human input is the key. This paper studies how crowdsourcing-based approaches to image tag validation can achieve parsimony in their use of human input from the crowd, in the form of votes collected from workers on a crowdsourcing platform. Experiments in the domain of social fashion images are carried out using the dataset published by the Crowdsourcing Task of the Mediaeval 2013 Multimedia Benchmark. Experimental results reveal that when a larger number of crowd-contributed votes are available, it is difficult to beat a majority vote. However, additional information sources, i.e., crowdworker history and visual image features, allow us to maintain similar validation performance while making use of less crowd-contributed input. Further, investing in *expensive* experts who collaborate to create definitions of interpretation-sensitive concepts does not necessarily pay off. Instead, experts can cause interpretations of concepts to drift away from conventional wisdom. In short, validation of interpretation-sensitive user tags for social images is possible, with "just a little help from the crowd".

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Performance Evaluation; I.2 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Human Factors, Experimentation

1. INTRODUCTION

User tagging has played an important role in the rise of social image sharing on the Internet. Tags assigned by users make it possible to find images or to browse large image collections. However, users' tagging patterns vary widely, and are dependent on the motivations and incentives that drive tagging behavior within a particular tagging system [9]. Automatic approaches exploiting pixel processing have been proposed to address the social image labeling challenge, e.g., [20, 22]. These approaches have made an important contribution, but have not solved the problem of improving the reliability of human-contributed tags. Ultimately, humans themselves remain the best source of human-interpretable descriptions of images [12].

In this paper, we investigate how crowdsourcing can best contribute to refine labels that describe the depicted content of social images. We tackle two image labeling tasks, validating whether or not an image is related to the domain of clothing and fashion, and validating which type of clothing item or fashion accessory it depicts. These two tasks are chosen because they are *interpretation-sensitive*, meaning that there is no absolute definition of what constitutes a fashion image, or a fashion image depicting a particular item or accessory. Instead, humans must interpret these concepts in order to apply them to images. Image retrieval systems that exploit tags build on the assumption that user tagging behavior will match user search behavior, as pointed out by [14]. For this reason, interpretation-sensitive tags are potentially very useful to systems. The challenge is not in making user-contributed tags consistent with each other, but rather, making them more reliable.

Interpretation-sensitive image labeling tasks are understudied in the literature, because automatic image classification techniques require visual consistency in order to function well. The interaction between interpretation-sensitive labels and automatic content-based labeling techniques is complex, as described, for example, by [4]. Here, we take the view that visual consistency should not be a factor in validating tags, but instead, we turn to human input in the form of crowdsourcing.

The key difference between the original user-contributed tags and the crowd-based validation of those tags is incentivization. Users contribute tags in support of goals within the social image sharing setting. In contrast, crowdworkers contribute validations of tags because it is their primary goal. Here, we define crowdsourcing as the work that is carried out in microtask markets, e.g., online platforms such as Amazon Mechanical Turk, that offer crowdworkers a reward in exchange for carrying out a small amount of work. The incentivization mechanisms of crowdsourcing platforms allows them to deliver human input on demand. The disadvantage of crowdsourcing platforms is that crowd input is costly and is of uneven quality. In this paper,

CrowdMM'14, November 7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2660114.2660123.

we tackle the issue of how human input can, even in the face of uneven quality, be used parsimoniously in order to refine labels that describe the depicted content of social images.

The paper arose out of observations that we made in the *Crowdsourcing Task of the MediaEval 2013 Multimedia Benchmark* [7]. In this task, a number of different research groups tackled the problem of how to combine multiple votes collected from a crowdsourcing platform into one high-quality validation of an image label. In the course of this research, we repeatedly made the observation that investing more resources into collecting human input did not always increase the quality of results. Specifically, we came to two insights, that form the major contributions of this paper: First, exploiting information from other sources makes it possible to "get by with just a little help from the crowd workers on the crowdsourcing platform". Second, although intuitively it seemed like a surefire solution to invest in a set of experts who could consult to create *absolute definitions* of labels, our *expensive experts* drifted away from conventional interpretations of the images. Including their votes caused performance to deteriorate. In sum, the added value of this paper is a systematic demonstration that more is not necessarily better when it comes to using crowdsourcing to improve the reliability of interpretation-sensitive tags for social image.

The rest of the paper is organized as follows, in the next section we discuss the background of this research including related work and the data set used by the benchmark task. Then, we discuss the label validation experiments that led to each of our insights, and end with a conclusion. Note that the novelty of this paper is not so much the specific nature of the approaches that we take to label aggregation and validation, but rather that it presents evidence that for a wide range of approaches a common insight applies, i.e., indiscriminately investing more resources is not an optimal approach to interpretation-sensitive image labeling.

2. BACKGROUND

We introduce our label aggregation approaches in the context of social image labeling. In this section we discuss some related work and describe the data set and the labeling task that is done to examine our proposed approaches.

2.1 Related Work

The basic problem of how to take judgements gathered from the crowd and aggregate them into a single high-quality judgement is referred to as *computing crowd consensus*. Many techniques have been developed for computing offline crowd consensus, and an excellent overview as well as reference implementations is offered by the SQUARE task framework [16]. In an earlier work, Sheng et al. [15] investigated how and to what extent labeling can be improved by repeated-labeling. They proposed different aggregation models that take into account the uncertainty of individual labels and model. They showed that *selective* acquisition of multiple labels can optimize the label quality/cost. The focus of their work is based on improving label quality by acquiring additional labels.

The approaches that we use in this paper model the overall work of crowdworkers (history-based approach) and also model the work of crowdworkers in conjunction with image difficulty (Nominal Label Extract [10]). Note that approaches exist that model even further dimensions of crowdworker performance [18]. However, for the purpose at hand, these approaches serve as good representations of crowd consensus approaches. Other related work involves hybrid automatic/human approaches to multimedia. An entry to this topic is provided by the framework presented by Bozzon et al. [1]. Also, work carried out in the area of assistive tagging, surveyed in [17], combines automatic and human image labeling. Here, we choose a simple yet elegant approach that is representative of approaches that combine visual image features with votes collected from the crowd.

Other work on image labeling has encountered interpretation-sensitivity or differences in crowd interpretations [2, 4], but in general does not actively embrace it. For any given image, human judges must consider both the visual content of the image, and also their understanding of the real world. For many areas, judges will share a common stable understanding of the world. Image labeling tasks in these areas are not considered interpretation sensitive. However, a concept such as fashion is open to different interpretations. It clearly includes some, but not all images of people wearing clothing. Different images give different impressions of whether that clothing is worn consciously, or incidentally. For this reason, whether or not an image is considered fashion is subject to interpretation. Such interpretations cannot be considered *personal* or *individual* perspectives, since a large amount of consensus does exist; however, the consensus falls short of being universal [5]. Our work on interpretation-sensitive image labeling tasks comes to a different conclusion that work focusing on image labeling tasks for which the common understanding of the world is less subject to interpretive variation. For example, [11] finds that the crowd is able to reproduce the labels generated by experts in the lab. In this paper, experts who consult with each other are shown to diverge from the conventional wisdom of the crowd, which, we argue, may ultimately be more useful in a social image search application.

2.2 Data set and task

The two label validation tasks are defined on a set of fashion-related images collected from the Flickr[1] photo-sharing platform. This data set, referred to as the *Fashion 10000 dataset*, is a publicly available data set containing nearly 32k social images collected from Flickr with Creative Commons licenses. The data set was used in the *Mediaeval 2013 Crowdsourcing Task* [7] and is described further in [6]. Example images are shown in Figure 1. The images are collected via a set of fashion-related keywords which queried over Flickr. Each image is labeled with a fashion category (e.g., *dress*, *trousers*, *tuxedo*). The name of the fashion category of the image is the keyword that was used to retrieve the image from Flickr at the time that the data set was collected. The retrieval was constrained such that the keyword was required to occur in the tag set of the image, and for this reason, corresponds to a user-assigned image tag [7].

Although users have tagged images with fashion words, not all images are relevant to fashion or clothing. The first labeling task is to determine whether or not an image is truly related to fashion or clothing (`Fashion`). The second labeling task is to determine whether or not the fashion category of the image correctly characterizes its depicted content (`valid-cat.`). Three sources of information can be exploited to infer the correct label of an image: *a*) a set of *crowd votes* which are annotations collected from Amazon Mechanical Turk crowdsourcing (AMT) platform with a basic quality control mechanism (three votes are provided for each label on each image), *b*) the metadata of the images (such as title, description, comments, geo-tags, notes and context), and *c*) the visual content of the image. In this work we developed various algorithms that rely on these different sources of informations. The input from crowd workers for each of the two task can be either *yes*, *no* or *not sure*. However the ultimate output of the labeling tasks are to predict a binary label, i.e., the final estimated label should be either *yes* or *no* [7].

[1] www.flickr.com

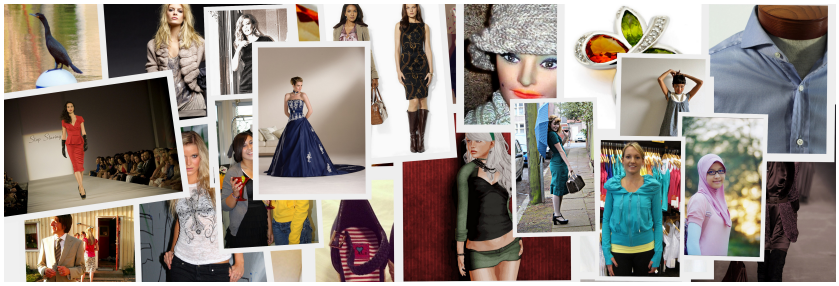

Figure 1: Sample images from the Fashion 10000 [6] dataset. Some images might be irreverent to fashion. Detecting whether an image is related to fashion or not is a typical interpretation-sensitive task which can benefits from the power of crowdsourcing.

A subset constituting 20% of the overall data set, containing 6262 images, was selected for testing purposes. This subset was annotated with addition crowd input collected from AMT using a state-of-the-art quality control mechanism [7]. For each image and each label, three high quality votes were collected. A majority vote is used to aggregate the three high quality votes to create high-fidelity label validation ground truth that are used for evaluating the experiments.

3. LESS LABELS LITTLE LOSS

In this section, we present two different approaches that combine crowd-contributed input and additional information sources to validate image labels. The first exploits worker history, and the second is a hybrid human/automatic approach that makes use of visual features.

3.1 Incorporating Worker's History

The *history* algorithm, introduced in [3], uses simultaneously assesses the worker reliability and the hidden labels. The aggregated crowd label of an instance i corresponds to L^i_{crowd} (i.e., *Yes* or *No*). L^i_{crowd} is computed by aggregating the individual worker labels $L^i_u \in \{Yes, No\}$. The worker confidence is a measure that indicates how well the worker is performing the task. We can either make a discrimination between the quality of the positive and negative answers or not. In the case of such a discrimination each worker is characterized by a positive confidence C^+_u and a negative confidence C_{u^-}, otherwise we use a single value for the worker confidence, C^*_u. In case we do not discriminate between positive and negative answer quality, the probability of an instance being labeled as positive is:

$$p^*_i = \frac{\sum_u C^*_u \cdot I(L^i_u = Yes)}{\sum_u C^*_u \cdot I(L^i_u = Yes) + \sum_u C^*_u \cdot I(L^i_u = No)} \quad (1)$$

In case we differentiate between the positive and negative answer quality this becomes:

$$p^+_i = \frac{\sum_u C^+_u \cdot I(L^i_u = Yes)}{\sum_u C^+_u \cdot I(L^i_u = Yes) + \sum_u C^-_u \cdot I(L^i_u = No)} \quad (2)$$

The probability of an instance being labeled as negative is obviously $p^-_i = 1 - p^+_i$. We will refer to the p^+_i and p^-_i as computed by using either method as **aggregated soft labels**. The final **aggregated hard label** assigned by the crowd is given by comparing the difference between the positive probability and the negative one $L^i_{crowd} = Yes$ if $p^+_i - p^-_i \geq 0$ and No, otherwise.

The confidence in a worker is defined as:

$$C^*_u = \frac{tp_u + tn_u}{tp_u + tn_u + fp_u + fn_u} \quad (3)$$

$$C^+_u = \frac{tp_u}{tp_u + fp_u} \quad (4)$$

$$C^-_u = \frac{tn_u}{tn_u + fn_u} \quad (5)$$

The above worker confidence values are calculated based on the final aggregated hard labels using the following equations:

$$tp_u = \sum_i I(L^i_u = Yes) \cdot I(L^i_{crowd} = Yes) \quad (6)$$

$$tn_u = \sum_i I(L^i_u = No) \cdot I(L^i_{crowd} = No) \quad (7)$$

$$fp_u = \sum_i I(L^i_u = Yes) \cdot I(L^i_{crowd} = No) \quad (8)$$

$$fn_u = \sum_i I(L^i_u = No) \cdot I(L^i_{crowd} = Yes) \quad (9)$$

The algorithm uses the E step to compute the aggregated crowd labels, and the M step to update the worker confidences.

To evaluate the history-based approach based on a limited amount of crowd input (i.e., one worker annotation), we introduce a variation of this method which relies only on one crowd vote. This method evaluates worker confidence using a training set and chooses the vote of the worker associated with the highest confidence. This scenario corresponds to the situation where we have three votes available in the training data, and in order to validate the image label we can choose among three workers, one worker from whom we can request input. We select the worker who was the best on the training data and use that worker's vote. The worker's confidence is calculated in the same way that described in the previous section using $Eq.$ 5.

We evaluated the performance of the history algorithms in two different scenarios. In the fist scenario we use all the three available votes while in the second scenario we use the one-worker-only model. The performance of the two scenarios on the test set based on the F1 score are presented in Table 1. We use F1 score to evaluate our experiments because our labeling tasks can be considered as two binary classification tasks (i.e., `Fashion` and `valid-cat.`). F1 score is the harmonic mean of precision and recall and is one of the most widely used metrics for binary classification tasks. The F1 scores for each of the two tasks are calculated separately.

Both scenarios are tested based on whether or not there should be discrimination between the positive and negative confidences. The

Model	P/N Discrimination	Fashion	Valid-cat
3-workers-history	N	**0.9103**	**0.8393**
3-workers-history	Y	0.9056	0.8355
Majority Vote baseline		0.9053	0.8358
1-worker-history	N	**0.8970**	0.8121
1-worker-history	Y	0.8890	**0.8155**
One-worker-only baseline		0.8638	0.7795

Table 1: Performance of the 3-worker and 1-worker history-based approach, in terms of F1 score, compared to two baselines. The history-based approach loses performance when only one worker is used, but the drop is minimal compared to the drop between the baselines.

three workers cases are compared with a majority vote baseline and the one-worker-only cases are compared with a baseline which is calculated only using one worker. To calculate this baseline we assume that the input is contributed by only by one worker: the vote of the first worker who carried out the task is taken to be the final label.

As the results in Table 1 show, the performance of the three-workers case for the no-discrimination scenario is slightly better than the majority vote baseline. On the other hand, the performance of the one-worker-only scenarios are considerably better than one-worker-only baseline while they are very close to the three-workers cases. Indeed with loss of almost 2% of the F1 for both labels, the task can be done by only one worker resulting in less costs for the crowdsourcing task.

3.2 Visual-Based Methods

The general idea of the here presented visual-based algorithms is to use the visual features of the images. The classification itself is based on a search-based approach developed with the LIRE Framework [8]. A similar technique was used in [21]. The classifier is trained with a training set which has a high fidelity ground truth generated from expert annotations.

Using LIRE we extracted the global features CEDD, FCTH, JCD, PHOG, Edge Histogram, Color Layout, Gabor, Tamura, Luminance Layout, Opponent Histogram, JPEGCoefficient Histogram and Scalable Color (which are described and referenced in [8]). On these features feature selection is applied to select the features with the highest information gain for each label. The selected features then are used for feature combination. This combination is carried out with a late fusion method using random forest classifier similar to the approach in [13]. At first, each feature is used to classify a query image. In the next step these decisions are combined to a overall classification. The classification is performed as a search where the to-classifying-image is the query for the search. For each image query a ranked list of similar images is returned. Based on the classes of these pictures the algorithm decides which class should be assigned to the query image.

The visual classifier predicts the class label for image i as:

$$L^i_{visual} = \underset{c\in\{yes,no\}}{\arg\max} \{S_i(c)\} \tag{10}$$

where $S_i(c)$ indicates the *class score* of class c for when image i in the test is used as query and is defined as:

$$S_i(c) = |c|. \sum_{j\in\{j|Class(j)=c\}} R_i(j)^{-1} \tag{11}$$

where $R_i(j)$ indicate the *rank score* of image j in the set of retrieved images.

For the second label, the classifier works slightly different. This is necessary because of the different quantity of representative images in different categories. Therefore, a weight value based on the number of available images per each category is applied. For the calculation of the weight the classifier uses the total number of images that belong to a category. The calculated weight is then applied to produce the result of the search-based classifier. Categories with fewer available images get a higher weight and vice versa. The proposed weighting approach makes sure that the classifier is not influenced by the large amount of images which are retrieved from the high cardinal categories.

In addition to our visual-only approach, we are also interested to understand the extent to which visual features can contribute to image-labeling performance when they are combined with human input in the form of a single vote from the crowd. We carried out in experiments in two more scenarios that combine visual features with crowd votes. In the first scenario, the final label is predicted using the same, previously used late fusion method. This method combines one vote from the crowd (effectively, the one-worker-only baseline in Table 1) with the visual features (Visual + One-worker-only). In the second scenario, all the three available crowd votes are aggregated using the worker's history method in Section 3.1 and then combined with visual features (Visual + 3-workers-history) using the same late fusion method.

Table 2 reports the F1 score of visual-only approach compared with the two crowd-incorporated approaches. These results demonstrate that if three workers are available, visual features do not provide added value. Specifically, the Visual-only method, as shown in Table 2, does not outperformn the Majority Vote basline. Unsurprisingly, no gain is achieved over the Majority Vote baseline when visual features are added to the 3-worker-history method. However, in situations in which only the input of one worker is available, visual features do provide an advantage. In Table 2, we see that "Visual + One-worker-only" method outperforms the "One-worker-only" baseline both for the `fashion` label and for `valid-cat` label. In fact, in the case of `fashion` label, the performance of the "Visual + One-worker-only" approach is indistinguishable from that achieved by the Majority Vote. Effectively, here, visual features have replaced the input of two crowdworkers. The ability of visual features to replace crowdworkers is less dramatic for `valid-cat`, but the performance of "Visual + One-worker-only" above and beyond Majority Vote still serves to demonstrate the ability of visual features to make it possible to get by with less input from the crowd.

For completeness we note that in the case of `fashion` label, the Visual-only method on its own does not outperform the Dominant Class baseline. The Dominant Class baseline is a classifier that assigns every item to the class which is represented by the majority of the items in the training set. In the case of `fashion` label the majority class is 'yes'. The F1 of the Dominant Class baseline for `fashion` label is 0.7832 and the Visual-only method achieves an F1 score of 0.7421. The comparison with respect to F1 obscures the potential of the Visual-only method to contribute to the performance of the classifier. The recall of the Dominant Class baseline is 1.0, but the precision is 0.6437. The Visual-only method is able to contribute because its performance is balanced differently between precision and recall.

In sum, we see that in practical situations, information about worker history, or visual-based methods, allows us to achieve competitive performance while using less information from workers (i.e., fewer crowd votes). In the next section, we turn to investigate the question of whether we can also get by with only a little help from the crowd if we have access to workers with specific expertise in the domain in which we are working.

Model	Fashion	Valid-cat
Visual-only method	0.7421	0.7290
Visual + One-worker-only	0.8957	0.8031
Visual + Visual + 3-workers-history	0.9078	0.8217
Dominant class baseline	0.7832	0.7403
Majority Vote baseline	0.9053	0.8358
One-worker-only baseline	0.8638	0.7795

Table 2: Performance of the visual-only approach compared with two hybrid visual-crowd methods in terms of F1 score.

4. WARPED WISDOM OF EXPENSIVE EXPERTS

In this section, we address the question of how to best use additional resources available to collect human judgements. Intuitively, it would appear that the problem of tag validation can be solved if we simply contract a group of expensive experts to consult with each other and arrive at agreement on how to validate the user tags.

The experiments in this section use a model called *Nominal Label Extract* (NLE). Like the history-based model, this model incorporates information about the user past performance in terms of a per-worker confidence, but also additionally also incorporates information on per-image difficulty. Exploratory experiments demonstrated that this model does not work well when only one worker annotation is available per image (intuitively, it is rather hard to estimate the difficulty of an image from a single vote). Rather, the strength of NLE lies in cases where more annotations are available, such as discussed here.

The NLE model, illustrated by the plate diagram in Figure 2, originated from work by Mineiro [10] and extends the model of Whitehill et al. [19] by incorporating a hierarchical Gaussian prior on the elements of the confusion matrix (the γ hyper-parameter in the figure). The model assumes an unobserved ground truth label z combines with a per-worker model parametrized by vector α and scalar item difficulty β to generate an observed worker label l for an image. The hyper-parameter γ moderates the worker reliability as a function of the label class. The model parameters are learnt using a 'Bayesian' Expectation-Maximisation algorithm. For our experiments with this model, we used the NLE implementation published by Paul Mineiro[2] with uniform class priors. Note that the software was applied to data from each of the two labels separately. This model is unsupervised, so it can be applied directly to the test data, however, it can also be used in an extended semi-supervised fashion by fitting the model using the entire dataset of training and test data together. Applying it to the entire dataset implies higher cost (more labels are used, and thus had to be collected). Intuitively, one might expect that using both the testing and training data should improve the fit of the model (more data is available for each worker), however, the additional data makes virtually no difference in our experiments.

In order explore the behavior of the model under the availability of more votes, extra votes were collected in two ways: Firstly, we randomly selected 1000 images from the test set and had them annotated by two reliable experts. The two experts first annotated the data independently, arriving at agreement in 671 cases (across both questions). For the images they did not agree on for either question, they collaboratively came to a decision about the 'true' vote for both questions. The relatively low-level of initial agreement between the experts is an indication of the interpretation-sensitive nature of the labeling task being

[2]http://code.google.com/p/nincompoop/downloads/

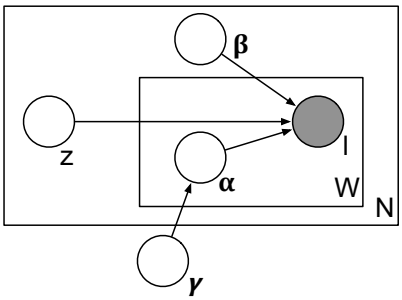

Figure 2: The *nominal label extract* generative model, incorporating per-item difficulty and per-worker reliability

Model	Fashion	Valid-cat
NLE	0.9055	0.8397
NLE + additional experts	0.8719	**0.8445**
NLE + additional non-experts	**0.9061**	0.8372
Majority Vote baseline	0.9053	0.8358

Table 3: Performance of the Nominal label extract (NLE) method, in terms of F1 score, compared with the situation where additional expert and non-expert annotations are collected. Only the test data is used for these experiments.

performed (especially with respect to Fashion label). Secondly, for the images in the entire data set (both test and training) that were assigned at least two 'not sure' votes by crowdworkers, we gathered more responses through additional crowdsourcing using the CrowdFlower[3] platform. In total we gathered additional of 824 responses over 421 images from this extra crowdsourcing task. Table 3 shows the effect of these additional votes on the F1 scores, with aggregation using the NLE algorithm. In the case where expert votes were used, the additional votes were used to *clamp* the model at the respective images in order to obtain a better fit.

The results in Table 3 indicate that although the performance of the prediction can be improved slightly when additional non-expert votes are added, but interestingly additional expert votes hurt the performance for Fashion label and improve the valid-cat. label slightly.

In order to understand how expert votes could hurt the image label validation performance, we did a hand analysis of cases in which the experts and the general crowd did not agree with each other. Three sample cases are shown in Figure 3. We noticed that the experts had a narrower range of images that they considered to reflect fashion than non-experts; in particular, experts considered that clothing worn for a particular function should not be considered fashion. In the image on the left, a man wearing a diving suit is shown. In the image in the middle, a group of religious devotees is pictured. The clothing that the people are wearing is part of what they are doing in the picture. In other words, the people in the picture could be argued not to have *chosen* their clothing. The experts do not consider clothing in this situation to count as fashion, but the crowd is more liberal. The most general assumption about the world is that *anyone can wear anything anywhere* (for example, to a costume party), and if Web users searching for fashion inspiration make this assumption, then the crowdworkers and not the experts should be considered right. The image on the right shows fabric. Here, the experts considered the image to be related to fashion (since fabric is the first step in making a fashion item), but the crowdworkers did not. Again, if Web users

[3]http://crowdflower.com

Figure 3: Samples from the test set for which experts and crowdworkers disagree on being fashion images or not.

searching for fashion want to see something that can be worn, then the interpretation of the crowd should be considered right. Note that we do not claim that the general crowd necessarily applies the same interpretations as Web users searching for images. Rather we point out that our experiments reveal evidence that experts consulting with each other do not necessarily contribute valuable votes, since they might drive interpretations away from the ones that are ultimately most useful.

5. CONCLUSION

In this paper, we have presented approaches demonstrating how human input collected from a crowdsourcing platform can be used parsimoniously in order to validate, and thus improve user tags. The result is better descriptions of social images, which ultimately transfers into improved search and browsing of social image collections. We focus on the fact that crowdsourcing is expensive with respect to automatic approaches and ask the question how we can benefit from crowdsourcing, while calling as little as possible on the crowd. Our finding have revealed that additional information sources can often compensate for crowdworker input, and that good validation performance can be achieved with only "a little help from the crowd". Further, we have observed that experts collaborating arrive at understandings of concepts that differ considerably from the conventional wisdom of the crowd. If tags are to be used to support image search that serves general user demographic, the effort of experts invested in making tags more consistent could ultimately be detrimental to the reliability of image tag validation.

Acknowledgments

This work was partially funded by the European Commission's FP7 project under grant agreement no. 287704 (CUbRIK).

6. REFERENCES

[1] A. Bozzon, I. Catallo, E. Ciceri, P. Fraternali, D. Martinenghi, and M. Tagliasacchi. A framework for crowdsourced multimedia processing and querying. In R. A. Baeza-Yates, S. Ceri, P. Fraternali, and F. Giunchiglia, editors, *CrowdSearch*, CEUR Workshop Proceedings, 2012.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] M. Georgescu and X. Zhu. Aggregation of crowdsourced labels based on worker history. In *Proceedings of International Conference on Web Intelligence*, 2014.

[4] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *Proceedings of the International Conference on Multimedia Information Retrieval*, 2010.

[5] M. Larson, M. Melenhorst, M. Menendez, and P. Xu. Using crowdsourcing to capture complexity in human interpretations of multimedia content. In *Fusion in Computer Vision*, 2014.

[6] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, M. Larson, and L. Gottlieb. Fashion 10000: An enriched social image dataset for fashion and clothing. In *Proceedings of the 5th ACM Multimedia Systems Conference*, 2014.

[7] B. Loni, M. Larson, A. Bozzon, and L. Gottlieb. Crowdsourcing for social multimedia at mediaeval 2013: Challenges, data set, and evaluation. In *MediaEval 2013 Workshop*, 2013.

[8] M. Lux and O. Marques. Visual information retrieval using java and lire. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2013.

[9] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, 2006.

[10] P. Mineiro. Modeling Mechanical Turk Part II. `http://www.machinedlearnings.com/2011/01/modeling-mechanical-turk-part-ii.html`.

[11] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, 2010.

[12] T. Pavlidis. Why meaningful automatic tagging of images is very hard. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009.

[13] M. Riegler, M. Lux, and C. Kofler. Frame the crowd: global visual features labeling boosted with crowdsourcing information. *MediaEval 2013 Workshop*, 2013.

[14] A. Rorissa. A comparative study of Flickr tags and index terms in a general image collection. *Journal of the American Society for Information Science and Technology*, 2010.

[15] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08. ACM, 2008.

[16] A. Sheshadri and M. Lease. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation*, 2013.

[17] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*

[18] P. Welinder, S. Branson, S. Belongie, and P. Perona. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems 23*. 2010.

[19] J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, page 2035–2043, December 2009.

[20] R. C. Wong and C. H. Leung. Automatic semantic annotation of real-world web images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[21] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *Proceedings of the international conference on Multimedia*. ACM, 2010.

[22] D. Zhang, M. M. Islam, and G. Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 2012.