

Image annotation: The effects of content, lexicon and annotation method

Zenonas Theodosiou · Nicolas Tsapatsoulis

Received: date / Accepted: date

Abstract Image annotation is the process of assigning metadata to images, allowing effective retrieval by text-based search techniques. Despite the lots of efforts in automatic multimedia analysis, automatic semantic annotation of multimedia is still inefficient due to the problems in modelling high level semantic terms. In this paper we examine the factors affecting the quality of annotations collected through crowdsourcing platforms. An image dataset was manually annotated utilizing: (i) a vocabulary consists of pre-selected set of keywords, (ii) an hierarchical vocabulary, and (iii) free keywords. The results show that the annotation quality is affected by the image content itself and the used lexicon. As we expected while annotation using the hierarchical vocabulary is more representative, the use of free keywords leads to increased invalid annotation. Finally it is shown that images requiring annotations that are not directly related to their content (i.e. annotation using abstract concepts), lead to accrue annotator inconsistency revealing in that way the difficulty in annotating such kind of images is not limited to automatic annotation, but it is generic problem of annotation.

Keywords Image Annotation · Crowdsourcing · Manual Annotation · Annotation Quality

Z. Theodosiou
Research Centre on Interactive Media, Smart systems and Emerging Technologies (RISE), Nicosia, Cyprus
E-mail: z.theodosiou@rise.org.cy

N. Tsapatsoulis
Dept. of Communication and Internet Studies, Cyprus University of Technology, Limassol, Cyprus E-mail: nicolas.tsapatsoulis@cut.ac.cy

1 Introduction

The enormous increase in the number of digital images generates the need to develop technologies for efficient archiving and access to visual content. Image retrieval can be performed either by content-based [66], [77], [48] or text-based [40], [15] methods. The content-based approach performs the retrieval by examining the collection of images and returning those with similar visual content to the image given to the user's query. On the other hand, text-based approach returns images which are accompanied by text similar to the user's text query. Text-based image retrieval (Fig. 1) remains the predominant choice, despite the successful development of several content-based multimedia retrieval platforms. For effective retrieval using text-based techniques, annotation process is an essential step.

Although image annotation can be accomplished through several methods [49], [74], [16], causes a significant difficulty in image retrieval for various reasons: It requires a huge effort, lots of images lack tagging because of their creators/ owners' unwillingness to describe them using textual information, manual tagging itself is not always accurate while the quality of annotation is questionable. These are some of the reasons why the content-based retrieval, despite its limitations, is still regarded as an alternative for accessing several image collections.

Many efforts have been made to achieve automatic image annotation [12], [42], [29] which take advantage of the benefits of both content-based and text-based retrieval methods and try to overcome their limitations. The ultimate goal is to provide keywords searching based on image content [75] without the need to explicitly tag all images in the collection. Automatic image annotation tries to emulate humans aiming to implicitly associate image visual content with semantic labels. Due to

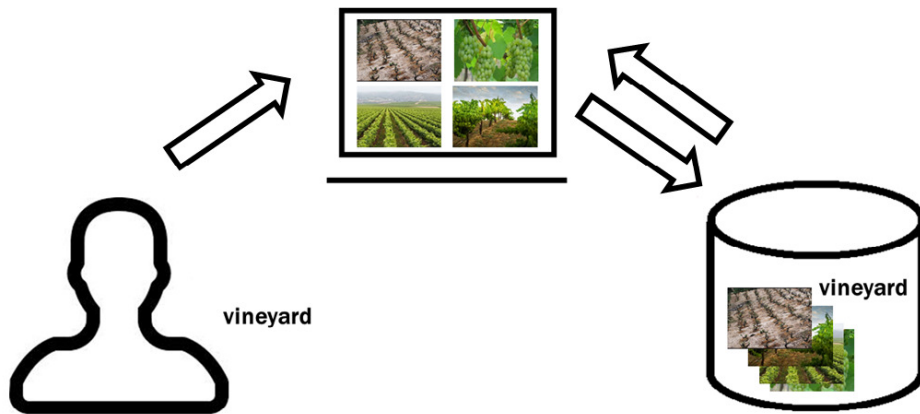


Fig. 1 Text-based image retrieval.

the broad range of its applications in addition to image retrieval, a large number of image techniques have been proposed that have been shown to achieve very promising results on various datasets. Automatic image annotation has many applications in the domain of image understanding and analysis and also used in other disciplines such as urban management, biomedical engineering, social media services, tourism industry, etc [12].

Manual image annotation is even more important for developing automatic image annotation methods using the training by example paradigm. Training examples that are used for automatic image annotation are pairs of images and annotations. Different approaches have been developed to model the correlation between visual features and text descriptions based on manually annotated datasets. The created models are then used to assign keywords to unseen data [76]. Generally speaking, clustering and classification based methods are among the most popular [36].

In clustering methods [71], [25] an algorithm quantizes the features from the training into clusters and replaces the continuous with the cluster centroids. Then, the probability of associating the concepts with clusters is calculated. The extracted features for a given image are compared with cluster centers and the closest ones are selected. The tag probabilities of the selected cluster centers are used to annotate the image. Several methods that analyzed the correlation between the clusters and keywords to discover hidden semantics had great success in automatic image annotation. Clustering based methods are fast and contain some level of generalization, but the priori unknown number of clusters and other clustering parameters remain the main problem of these methods. On the other hand in classification approaches [19], [20], keyword classifiers are developed based on the provided training exam-

ples, and are then used to classify an input image into one of various classes. According to the traditional machine learning flow, a feature extraction algorithm is applied to the input image and then the extracted features are used to create the classification model using a machine learning algorithm such as SVM [28], [39], Decision Trees [58], [59], Hidden Markov Models [54], Neural Networks [72], etc. The huge amount of data coming from visual sensing devices challenge the traditional machine learning approaches and calls the application of deep learning algorithms [38], [53], [44]. Convolutional Neural Networks (CNN) are now frequently used in recognition and detection tasks [11] achieving high performance which approach the humans on some tasks. The output of CNN layers can be interpreted as visual features and in this case the algorithm plays simultaneously the roles of feature extractor and classifier. In [68], a method based on recurrent neural networks (RNNs) is proposed to overcome the limitation of traditional methods to multilabel image classification [30] to explicitly use the label dependencies in an image.

The limited number of available training examples creates ineffective models of keywords without generalization ability and major problems in deep learning classification schemes which are typically trained on very large amount of labeled images [52]. In addition, the small number of available classes creates limitations to the retrieval results of text queries. Users, like the case of search engines, prefer to use free text rather than interfaces which include specific keyword sets.

Image annotation is a complex socio-cognitive process. It involves processing sensory input through classifying, abstracting, and mapping sensory data into concepts and entities often expressed through socially defined and culturally justified linguistic labels and identifiers [24]. The raw image data can not readily trans-

ferred to high-level semantics that usually appeared in users' queries [23]. This makes image annotation even more dependent on both humans and image content. There are many demographic factors that may affect the way that people interpret the image content [63]. The idea of collecting annotations through crowdsourcing [35] rather than an expert can significantly alleviate the subjectivity in image annotation, accelerate the whole process, reduce its cost and improve the ultimate efficiency of image retrieval systems.

In this paper we propose a new framework for studying the quality of manual annotations with the aim to: (a) investigate how the use of structured and unstructured vocabularies affects the quality of annotation and at what cost (lost of useful and valid annotations), (b) explore the influence of using an explicitly designed multimedia annotation tool for image annotation with respect to annotation quality as well as to the richness of the created annotations, (c) specify to which extend and under what conditions free annotation can result in valid and useful image annotation, and (d) identify how the content of the image can affect the image annotation. A dataset consisted of 500 images which was manually annotated utilizing different methods was used for our experiments. The collected annotations of each method were analyzed independently, resulting in very useful results and future implications in the domain of manual image annotation.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed framework while Section 4 describes the evaluation process. Section 5 presents and discusses the results of the experiments. Finally, conclusions and further work hints are given in Section 6.

2 Related Work

Image annotation has attracted great interest the last years and a broad range of remarkable approaches has been presented [45]. The assignment of textual descriptions can be achieved using several methods including free keywords, vocabularies and ontologies [22]. In spite of the fact that the automatic assignment of keywords is characterized as a very difficult task, the tremendous need for efficient image retrieval forced the research community to focus on the development of automatic image annotation methods. The proposed efforts tried to face the following challenges: (i) the extraction of rich semantic information using visual features, and (ii) the absence of correlation between the textual information and image regions in the training data. Although some methods have been developed for the automatic assignment of keywords based on web-page's surrounding in-

formation [21], they have not yet reached the level of success of the manual annotation. On the other hand, the annotation of large-scale datasets is time consuming and extremely difficult and manual annotations are recognized as imprecise, ambiguous, inconsistent and have some limitations [46]. The assignment of several people into annotation tasks will result in the collection of various annotation per image improving the overall quality and overcoming the above issues.

Joachims et al. [31] showed that there are minor differences between implicit and explicit relevance judgments. This outcome led to the use of implicit relevance judgments as training data in machine learning frameworks for several information retrieval applications [43], [65]. This approach can easily address the problems in a cheap and quick way and helps task creators to exploit different opinions [34]. Implicit crowdsourced annotations can be easily gathered without burdening the involved users [64]. Since the collection is performed without supervision, it may contain several errors due to erroneous participant's feedback which is independent of whether or not they will receive a reward for their participation [10]. Consistency is a big issue so the inter-annotator and intra-annotator agreements are very crucial measurements for the quality of the collected annotations [7]. The inter-annotator agreement shows the level of consensus and homogeneity among annotators while the intra-annotator agreement shows how consistent is an individual annotator. All resources involved in an annotation assessment, such as annotators and vocabularies, should be carefully selected and effectively used [33].

Crowdsourcing allows the harvest of crowd's wisdom and contributions [27], [18] and has helped create datasets by outsourcing the work to a large crowd of workers [69]. The tasks which were traditionally undertaken by employees or contractors can now be assigned to an undefined crowd [26], [6]. Several works studying the crowdsourcing annotations under different perspectives highlighted the importance and promising future of this research area [27]. Academia and industry recognized crowdsourcing as a simple solution to easily collect annotations at little cost. From the early beginning, the Amazon Mechanical Turk (MTurk) extended the capabilities of the crowdsourcing tasks by introducing more comprehensive GUIs and reward tools [8]. There is the possibility to assign the tasks to a large number of workers and gather the results after a few hours. A primary goal is to reduce measurement error [32], or to enhance the quality [4] of the collected data. The MTurk workers [47] and the quality of collected data have been studied for a broad range of tasks [41].

Annotation quality obtained through crowdsourcing varies. Annotators provide erroneous or poor-quality labels either in the hope that they will be unnoticed and will not have payment penalties or because they have not properly understood the task in hand [69]. Different approaches have been proposed that investigate the quality of crowdsourcing annotations. Snow et al. [62] discovered that annotators in a crowdsourcing framework cannot reach the experts' effectiveness but the combination of their opinion may yield to annotations of good quality. This outcome indicates the importance of aggregating the annotations, showing that a large number of non-experts can achieve better annotation quality than a small number of experts on the same task.

Raykar et al. [57] tried to solve the difficult case which there are no ground truth but only poor quality annotations are available to be used in a supervised machine learning framework. This work emphasized the importance of effective annotators and tried to estimate sensitivity and specificity of each annotator. It also recommended the use of multiple annotations for each item combining different weights for each annotator related to his/her agreement on the ground truth dataset. Smith et al. [61] analyzed the difficulties of evaluating tasks which consist of several annotators without the presence of ground truth data. They tried to create a gold standard by combining the opinions of several experts and showed that the annotator's consensus can be used to evaluate the annotation quality. Annotation quality is also studied in [60] showing that repeated and careful labeling can increase the labeling quality. The consistency of annotators was also utilized in the context of paired games and CAPTCHAs [3] for the creation of ground truth dataset. The difficulty of non-expert annotation and the abilities of annotators are examined in [70], while a dedicated system which guides the users to give more informative and cost effective labels is presented in [67]. An interesting technique for the collection of ground truth data is presented in [5] which utilizes disagreement-aware metrics to evaluate the ambiguity inherent in crowdsourced annotations.

The assignment of several annotators into the same task, creates the difficulty to decide if an annotation is a positive or a negative in case there is a disagreement between the answers. The inter-agreement can be successfully used to evaluate the quality and, as the agreement rate increases, the annotation quality gets more accurate and reliable [55]. Kilgarriff [33] proposed an interesting approach to create gold standard datasets and underlined the advantage of using more people while keeping the inter-annotator agreement in a high level. The author also indicated the causes of am-

biguous annotations focusing mainly on the task definition and worker's profile. The annotation quality created through the MTurk platform as well as the quality of annotations given by non-experts are evaluated in [62]. The authors calculated the inter-annotator agreement between experts and non-experts and concluded that the average performance of many non-experts is converged to the performance of an expert. Callison-Burch [9] studied the quality of annotations collected through MTurk and showed that the aggregation of non-expert judgments can lead to a high agreement based on a gold-standard dataset. Nowak et al. [50] studied the differences between various sets of annotations given by experts and the reliability of non-expert annotations to provide ground truth data. MIR Flickr images which were previously annotated by experts were then annotated by non-experts using MTurk and their inter-agreement was evaluated under different experiments. The majority vote, accuracy and k statistics were utilized to calculate the inter-annotator agreement at image-based and concept-based level. More recently, a study was presented which examines the reliability of labeling through examining inter-annotator agreement when two or more analysts label the relations in a sample of compound noun [73]. The results indicate that the agreement was fairly high and disagreements were consistent.

Crowdsourcing gave new insights into human computation. Often non-experts provide annotations of poor quality which are noisy and might require additional validation. In addition, the annotation method sometimes plays a decisive role in annotations quality. Nevertheless, the non-experts would be the majority of tomorrow's users in search machines and their judgments are very important for datasets' creation. The majority of current studies aim to justify the usefulness of crowdsourcing, identify the drawbacks and compare the non-expert annotation quality with expert ones. In the current work we extend the survey on crowdsourcing annotation and try to find out how the selected annotation method, image content itself and used lexicon affect the quality of manual image annotation.

3 Proposed Framework

In this section we present the key elements of the multi-step approach (Fig. 2) we followed in order to collect and analyze the crowdsourced data.

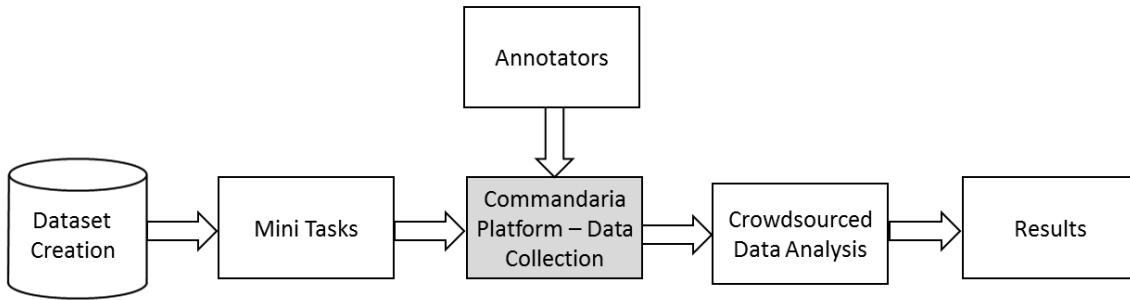


Fig. 2 The block diagram of the proposed method.

3.1 Commandaria Platform

The Commandaria platform [1] was developed under the Commandaria Project ¹ and provides users with the opportunity of profiling, uploading, annotating and searching information related to the Commandaria Cypriot Wine. During the project, people around the globe were encouraged to: (i) Upload multimedia content which relates Commandaria with the history and culture of Cyprus, (ii) Annotate the available content. The Commandaria dataset counts 7500 files in total where the 3500 files are digital manuscripts and scanned papers from books, journals and legislation documents and the remaining 4000 files are images and videos. The efficient retrieval of this information for several user categories requires proper annotation. Commandaria data are of priceless value for Cypriot heritage, thus, their collection, proper preservation and access is a task of tremendous importance [51]. Efficient indexing and retrieval depend upon accurate and rich data annotation which may be affected by various factors.

3.2 Annotation Methods

Images can be associated with several types of metadata. In this work, we focus on textual information which refers directly or indirectly to image’s content and is classified in the following categories [22]: (a) Content-descriptive metadata, which directly describes the semantic content of the image such as objects, scenes, meanings, emotions, etc., (b) Content-independent metadata, which indirectly describes the content of the image such as date, location, etc. The textual information for both categories can be assigned using keywords from vocabularies, taxonomies, hierarchical vocabularies or using free keywords. The first method restricts the annotator to choose keywords from a vocabulary created

based on Commandaria Taxonomy. The second method allows annotator to choose keywords from an hierarchical vocabulary which is based on the pre-selected keywords of the first method. There are many ways to classify the visual content of an image according to demographics factors [22]. Thus, a third method is offered to annotators, as an effort to overcome the limitations of the previous methods, which provides the opportunity of using free keywords. Although this method has no restrictions, it is subjected to several challenges due to spelling errors and grammatical mistakes. The use of a dedicated spell checker and/or an ontology are necessary to address these challenges. The three methods almost cover the whole range of manual annotation approaches. Due to the fact that each annotation method has its advantages and disadvantages, their combination provides a more complete annotation approach.

3.3 Dataset

A dataset of 500 images, randomly selected from the Commandaria collection was used for our experiments. The vocabulary and hierarchical vocabulary were consisted of 28 keywords of the annotation taxonomy compiled by Commandaria team. The dataset was uploaded on Commandaria platform in form of mini-jobs and annotated by non-experts. The annotators were students at the Department of Communication and Internet Studies of Cyprus University of Technology, who enrolled in Digitalisation of Cultural Heritage course. Each participant in the experiments, received partial credit toward completion of the course.

3.4 Task Design

The annotation process was consisted of mini-tasks, where each mini-task was concerned with the annotation of one image using one or more of the three proposed methods. The annotator was able to choose which image to annotate by choosing the relevant mini-task

¹ “The History of Commandaria: Digital Journeys Back to Time”, project funded by the Cyprus Research Promotion Foundation (CRPF) under the contract ANTHRO/0308(BE)/04.

from the list of thumbnails. A list of instructions was presented to the users who were also asked to classify the image based on their content into categories “abstract” and “specific” before starting the annotation. The instructions helped the annotators to complete the annotation process, including how each method works and the minimum time needed to complete a mini-task (the minimum time was equal to 60 seconds based on the expert’s calculation). An example of a mini-task is shown in Fig. 3. According to the first method, the users were able to assign annotations to an image by selecting the most appropriate pre-selected keywords from the list. In addition, the users were able to select keywords from the hierarchical vocabulary provided by the second method. The hierarchical vocabulary was classified in three main categories. Each category was further classified to a number of subcategories and each subcategory to a number of nodes and so on, providing an hierarchical annotation tree. The users were also provided the opportunity to add free keywords in English or Greek language using the third annotation method.

4 Evaluation Process

4.1 Mathematical Background

In this subsection we set the mathematical background of the evaluation process. We denote by A_i the i -th annotator ($i=1, \dots, N_A$). I_j indicates the j -th image ($j=1, \dots, N_I$) in the image dataset, while N_I denotes the total number of images in this dataset. t^{ij} indicates the set of keywords suggested by annotator A_i for image I_j . The total number of keywords suggested by the i -th annotator, T_{A_i} , and total number of keywords submitted for the j -th image, T_{I_j} , are computed by equations (1) and (2), respectively:

$$T_{A_i} = \bigcup_{j=1}^{N_I} t^{ij} \quad (1)$$

$$T_{I_j} = \bigcup_{i=1}^{N_A} t^{ij} \quad (2)$$

The representative keywords denote the valid keywords for each image. According to the evaluation metric, a representative keyword for an image is every keyword that was being suggested either by an expert or by the majority of the annotators or by more than one annotator (exclusion of keywords suggested by mistake). The set of representative keywords for the j -th image is denoted by $K^j = \{K_1^j, \dots, K_n^j\}$. The relative complement of K^j in T_{A_i} , denoted as $\mathcal{C}_{T_{A_i}}(K^j)$, indicates the keywords suggested by the i -th annotator for the j -th image that are invalid. Similarly, the relative complement of K^j in T_{I_j} , $\mathcal{C}_{T_{I_j}}(K^j)$, indicates the keywords

suggested for j -th image by all annotators that are invalid. The Venn diagrams explaining the two relative complements are shown in Fig. 4. Finally, the intersection between the representative keywords K^j for image I_j and the keywords that the annotator A_i suggested for the same image, denoted as v^{ij} , indicates the set of valid keywords that suggested by the i -th annotator for the j -th image. Therefore $v^{ij} \subseteq t^{ij}$ and $v^{ij} \subseteq K^j$.

4.2 Evaluation Metrics

After identifying and manually correcting the keywords given as free text, we utilized the below measurements for answering the questions set in Section 1:

1. Annotators consistency

This measurement calculates the annotators consistency by comparing the representative keywords proposed for each image, with keywords proposed by each annotator for the same image. A keyword is considered representative if the number of times being suggested is equal to or above a threshold $Th=2$, indicating that it was submitted at least by 2 annotators.

The overall consistency C^i , of the annotator A_i is given by summing its consistency across all images he/she annotated:

$$C^i = \sum_{i, t^{ij} \neq \emptyset} \frac{|(v^{ij})|}{|(K^j)|} \quad (3)$$

2. Total number of suggestions for each free keyword

The second measurement focuses on the use of free keywords by calculating the total number of suggestions for each one.

3. Percentage of the suggested free keywords for each image

The third measurement aims to examine if there is a significant difference in the use of the free keywords for “abstract” and “specific” images.

4. Percentage of invalid keywords suggested by each annotator

The fourth measurement aims to determine the percentage of keywords submitted by each annotator and are invalid. A keyword is considered as invalid, first, if it was not suggested by the expert, and second, if it was not suggested by more than the half of the annotators. The invalid keywords submitted by i -th annotator for the j -th image is the relative complement the first measurement. The percentage P_{A_i} of invalid keywords suggested by the i -th annotator is calculated using the following formula:

$$P_{A_i} = \frac{|\bigcup_{j=1}^{N_I} (\mathcal{C}_{t^{ij}}(K^j))|}{|T_{A_i}|} \quad (4)$$

File Annotation

Annotate the file: **Winery**

1. Select the most representative vocabulary keywords

<input type="checkbox"/> Grape Cultivation	<input type="checkbox"/> Producers	<input type="checkbox"/> Culture Events
<input type="checkbox"/> Grape Collection	<input type="checkbox"/> Wine Judges	<input type="checkbox"/> Campaigns
<input type="checkbox"/> Mellowing Draining	<input type="checkbox"/> Historical People	<input type="checkbox"/> Public Advertisements
<input type="checkbox"/> Wine Production	<input type="checkbox"/> Consumers	<input type="checkbox"/> Private Advertisements
<input type="checkbox"/> Consumption	<input type="checkbox"/> Writers	<input type="checkbox"/> Location Places
<input type="checkbox"/> Legislation	<input type="checkbox"/> Merchant Dealer Trader	<input type="checkbox"/> Stories Legends
<input type="checkbox"/> Books	<input type="checkbox"/> Ancient Times	<input type="checkbox"/> Cyprus
<input type="checkbox"/> Research	<input type="checkbox"/> Middle Times	<input type="checkbox"/> Commandaria Region
<input type="checkbox"/> Wine Review Results	<input type="checkbox"/> Modern Times	<input type="checkbox"/> Other Regions
		<input type="checkbox"/> World

2. Select the most representative hierarchical vocabulary keywords

<input type="checkbox"/> Community	<input type="checkbox"/> Time	<input type="checkbox"/> Documents
<input type="checkbox"/> People	<input type="checkbox"/> Production Cycle	<input type="checkbox"/> Publications
<input type="checkbox"/> Producers	<input type="checkbox"/> Grape Cultivation	<input type="checkbox"/> Legislation
<input type="checkbox"/> Wine Judges	<input type="checkbox"/> Grape Collection	<input type="checkbox"/> Books
<input type="checkbox"/> Historical People	<input type="checkbox"/> Mellowing Draining	<input type="checkbox"/> Research
<input type="checkbox"/> Consumers	<input type="checkbox"/> Wine Production	<input type="checkbox"/> Wine Review
<input type="checkbox"/> Writers	<input type="checkbox"/> Consumption	<input type="checkbox"/> Dissemination
<input type="checkbox"/> Merchant Dealer Traders	<input type="checkbox"/> Period	<input type="checkbox"/> Cultural Events
<input type="checkbox"/> Area	<input type="checkbox"/> Ancient Times	<input type="checkbox"/> Campaigns
<input type="checkbox"/> Cyprus	<input type="checkbox"/> Middle Times	<input type="checkbox"/> Public Advertisements
<input type="checkbox"/> Commandaria Region	<input type="checkbox"/> Modern Times	<input type="checkbox"/> Private Advertisements
<input type="checkbox"/> Other Region		<input type="checkbox"/> Location Places
<input type="checkbox"/> World		<input type="checkbox"/> Stories Legends

3. Add free Keywords

Select




Fig. 3 A mini-task of image annotation process.

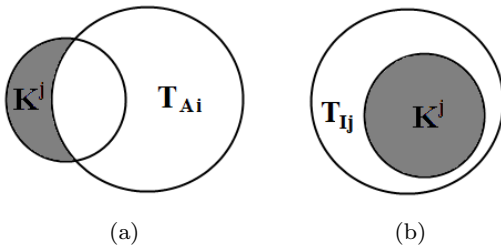


Fig. 4 Venn Diagrams for: (a) the relative complement of K^j in T_{A_i} , (b) the relative complement of K^j in T_{I_j} .

5. Percentage of invalid suggested keywords for each image

We extend our study on invalid keywords with the fifth measurement which identifies the percentage of invalid keywords submitted per image. The percentage of invalid keywords for the j -th image is calculated as follows:

$$P_{I_j} = \frac{|\bigcup_{i=1}^{N_A} (G_{t^{ij}}(K^j))|}{|T_{I_j}|} \quad (5)$$

6. Agreement analysis between expert and non-experts

The sixth measurement allows the estimation of the accuracy agreement between the expert and the non-experts on image basis. Based on the formula proposed by Brants [7], the accuracy between non-experts' and expert's annotations for the I dataset can be defined as follows:

$$Accuracy(I) = \frac{1}{N_I} \sum_{j=1}^{N_I} \frac{|(K_j)|}{T_{I_j}} \quad (6)$$

Where, the T_{I_j} denotes the total number of keywords given by the non-experts for the j -th image and K_j denotes the representative keywords given by the expert for the same image.

7. Reliability of agreement among the annotators

The reliability of the agreement among the annotators on keyword basis is computed in the seventh and final measurement using the kappa statistics. Cohen [13] first proposed this statistical mea-

sure for estimating the inter-rater agreement which takes values between 0 and 1, where 0 denotes the agreement resulting from random assignment and 1 the full agreement. Kappa values less than 0 denote agreement below the anticipated from random assignment. A sufficient agreement is denoted by a kappa value greater than 0.6 while the perfect agreement is denoted by a kappa value greater than 0.8 [37]. The free-marginal kappa statistic [56] which can be used for any number of annotators who do not assign a specific number of images to each keyword, was utilized in a binary scenario for each one of the 28 pre-selected keywords.

5 Results and Discussion

Each one of the 500 mini-tasks was assigned 50 times resulted in 25000 annotations. All annotations completed before the minimum required time were rejected. Overall 36 annotation sets were rejected that belong to 20 different images. The remaining annotations were used to evaluate the annotators' consistency and compare the efficiency of annotation methods. Furthermore, we have explored the significance of image content in the quality of crowdsourced annotation.

From the 50 annotators participated in our experimental setup, the 31 utilized the vocabulary keywords, while the remaining 19 utilized the hierarchical vocabulary keywords to describe the set of 500 images. Additionally, 45 annotators improved the image annotation by adding free keywords.

The consistency of annotators is presented in Fig. 5. In the case of vocabulary keywords, the majority of annotators show medium to low average annotation score indicating their weakness to understand the meaning of the pre-selected keywords. Free keywords improve the annotation score but a significant improvement occurs when the annotators use the hierarchical vocabulary keywords. The results indicate that the hierarchical structure of the vocabulary keeps the level of the annotation score high even though some annotators did not assign the same keywords to the same image.

Concerning the average annotation consistency, it is evident from Table 1 that annotations using the hierarchical vocabulary or free keywords are more consistent than using vocabulary keywords. Furthermore, the t-test was used to evaluate if there is an adequate evidence to claim a difference between the three annotation methods. Each annotation method was separately compared with the other two. The specific statistical test compares the mean values of the two distributions to verify whether the hypothesis that there is no difference between the two methods is rejected. Using the

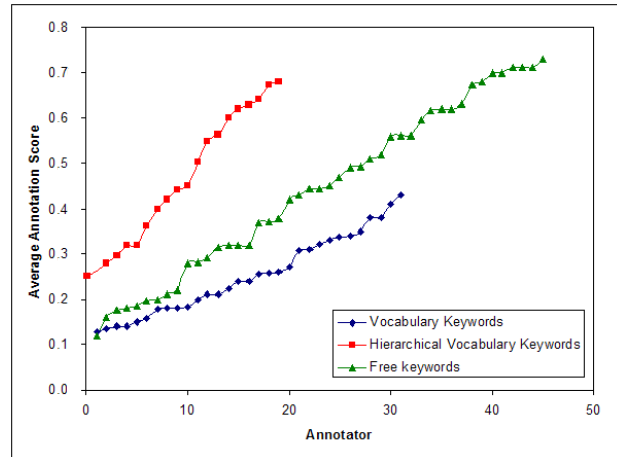


Fig. 5 Annotators consistency using vocabulary, hierarchical vocabulary and free keywords.

Table 1 Average Annotation Consistency.

<i>Annotation Method</i>	<i>Average Consistency</i>
Vocabulary Keywords	0.25
Hierarchical Vocabulary Keywords	0.47
Free Keywords	0.44

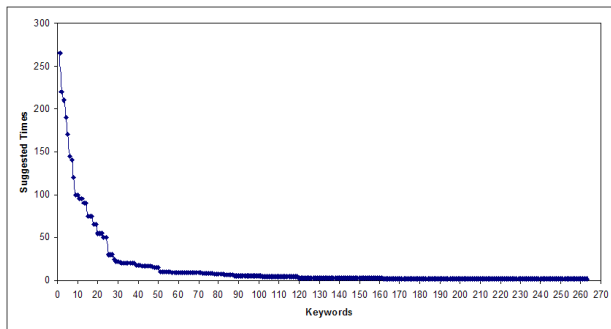
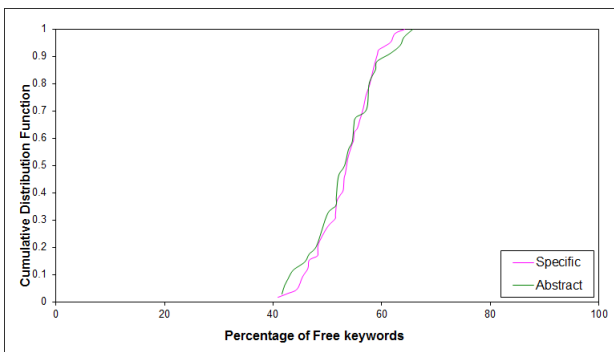
annotation consistency of each method, the t-test at the 5% confidence level [14] was used to test the significance of the differences. The results in Table 2 show that there are significant differences between the use of: (i) vocabulary keywords and hierarchical vocabulary keywords, and (ii) vocabulary keywords and free keywords, while there is no significant difference between hierarchical vocabulary keywords and free keywords.

The majority of the annotators chose to improve the image description using free keywords. The example image depicted in Fig. 6 was described by a number of annotators as “producers”, “wine judges”, or “historical people” while others chose to describe it based on its semantic representation and used free keywords such as “old man” or “grandfather”. A total number of 818 different free keywords were suggested by 45 annotators. Fig. 7 presents the total number of suggestions for the 263 keywords that suggested twice or more. Some keywords received a high number of suggestions indicating their importance for annotating the set of images used for the experimental setup.

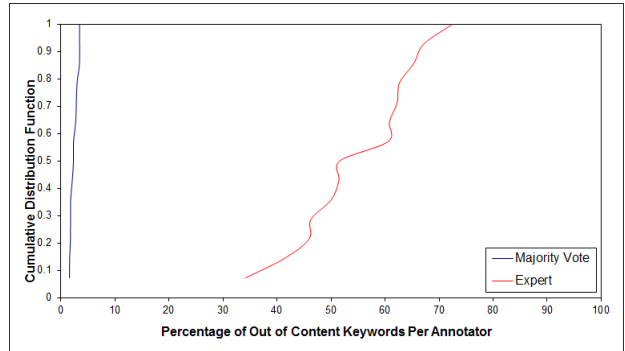
As shown in Fig. 8, the content of the image does not affect the use of free keywords. The annotators have an average percentage of free keywords equal to 53.44 % for all images, 53.24% for abstract and 53.51% for specific images. The frequent use of free keywords during annotation indicates first, the weakness of the predefined keywords to cover all the important content-descriptive metadata, and second, the weakness of annotators to understand the meaning of that keywords.

Table 2 Consistency differences in the use of keywords.

	<i>Vocabulary Keywords</i>	<i>Hierarchical Vocabulary Keywords</i>	<i>Free Keywords</i>
<i>Vocabulary Keywords</i>	-	s. (significant)	s.
<i>Hierarchical Vocabulary Keywords</i>	s.	-	n.s. (not significant)
<i>Free Keywords</i>	s.	n.s.	-

**Fig. 6** Example used in our experiments.**Fig. 7** The total number of suggestions for the 263 free keywords that were suggested twice or more.**Fig. 8** Cumulative Distribution of Percentage of Free Keywords.

The frequent use of free keywords increases the number of invalid keywords which can be examined under two perspectives: First, in terms of the limited perception of annotators regarding the semantic content of each image and second, in terms of the difficult and confusing meaning of the selected images especially for

**Fig. 9** Cumulative Distribution of invalid Keywords Per Annotator.

people who are not experts of the domain. Concerning the perception of annotators, the percentage of invalid keywords was calculated for each annotator based on majority vote and expert annotations as presented in Fig. 9. Based on majority vote, the annotators gave an average percentage of invalid keywords equal to 2.44% for all images while the average percentage of invalid based on expert annotations is equal 55.14%. The significant difference (t-test) between the two percentages indicates the limited perception of annotators according to the expert one. Furthermore, the low percentage of invalid keywords shown when based on majority vote implies the agreement between the non-experts in annotating the image dataset.

The difficult and confusing meaning of the selected images for the non-experts is examined through the accuracy of agreement on image basis. During the annotation process the 500 images were classified by annotators into “abstract” and “specific” categories. Based on their visual content, the 330 images were classified as “abstract” while the remaining 170 were classified as “specific”. Examples of the images classified in two categories are presented in Fig. 10. As shown in Fig. 11, more than 80% of the images classified as “specific” and the 38% of the images classified as “abstract” were annotated with accuracy agreement more than 0.7. The average accuracy for “abstract” images is 68.87% while the average accuracy for “specific” images is 92.42%. Furthermore, the averaged accuracy between the expert and the majority of annotators is 0.84. In terms of “specific” and overall accuracy the annotations from the non-experts show good results as the expert. Com-



Fig. 10 Examples used in our experiments which were classified as: (a) “abstract”, (b) “specific”.

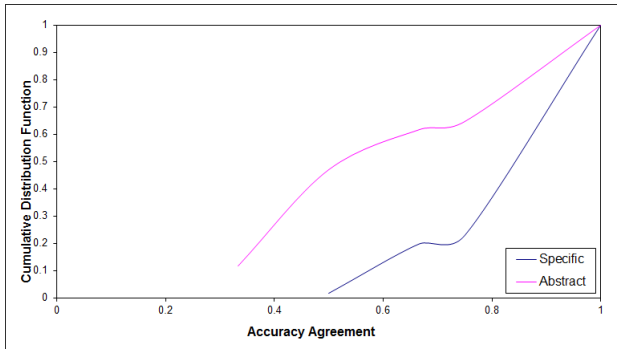


Fig. 11 Cumulative Distribution of Accuracy agreement.

paring the accuracy agreement between “specific” and “abstract” images, the results are in full agreement with the idea drawn by Fujisawa [17] which indicates that the terminology and description of cultural heritage are often too technical and difficult for nonprofessional users of the domain.

Finally, the inter-annotator agreement is calculated using an online kappa calculator [2]. Fig. 12 presents the kappa statistics for 28 pre-selected keywords. On average the 31 annotators used the vocabulary keywords to annotate the image dataset agree with a kappa value of 0.83 and the 19 annotators used the hierarchical vocabulary keywords with the value of 0.86. Although the annotation score based on the first measurement is medium to low, the reliability of their agreement on vocabulary keyword basis is almost perfect. The agreement for the majority of keywords is greater than 0.8. More abstract keywords like “Cyprus”, “Location Places” and “Commandaria Region” present lower agreement among the annotators. Some of the pre-selected keywords were not selected to annotate the image dataset like the “Legislation” which presents one of the highest value indicating that all annotators (except one) agree with it. Nevertheless, the resulted annotations cannot be considered as a general annotation agreement for this keyword.

The results show that the pre-selected keywords based on Commandaria taxonomy create an extra difficulty to the non-expert annotators and play an important

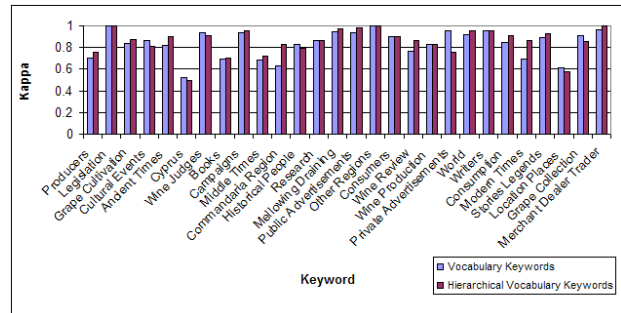


Fig. 12 The kappa values for the lexicon keywords.

role in the annotation results. The structure of the hierarchical vocabulary helped the annotators to use the keywords whether or not they understood their meaning. The fact that the selection of a keyword assigns automatically all the keywords that belong to same hierarchy enhances the annotators’ consistency. The difficulty in understanding the meaning of pre-selected keywords made significant the difference between the use of vocabulary and hierarchical vocabulary, increased the number of suggested free keywords and therefore the number of invalid keywords. The high inter-annotator agreement indicates that non-experts used the pre-selected keywords almost in the same way. The accuracy agreement between non-experts is higher for the images classified as “specific”, indicating that visual content affects the quality of the ever, the content does not play important role in the use of free keywords since annotators suggested almost the same number of free keywords for both “specific” and “abstract” images. Although the annotation for culture heritage images is difficult to achieved by non-experts, the combination of the annotation methods as well as the annotations based on majority vote can achieve very desirable results which are very close to those given by an expert.

Although the research has focused on cultural heritage images, the results are more general in nature and can contribute to a better understanding of the factors that affect the quality of image annotation. The fact that there is a difference between expert and non-expert annotations, difficulty in understanding the pre-selected keywords which is tempered by the use of hierarchical vocabulary and free keywords, high accuracy agreement between non-expert annotators, high accuracy between expert’s annotations and annotations determined by the majority vote of non-experts, and difficulty in understanding and annotating “abstract” images, may be related to the annotation of any image dataset. However, each dataset and its related knowledge create different difficulties in the way non-experts annotate images. In our case, the pre-selected keywords and the content of the images created difficulties to

the non-expert annotators who preferred to annotate the image dataset in colloquial language using free keywords. In any case, different image dataset, vocabulary and annotation framework may lead to different findings.

6 Conclusions and Future Work

With the tremendous use of visual devices and social media the number of available images is ever increasing. Therefore, the efficient search, retrieval and access of images through search engines is an urgent need. Despite of significant work in the domain of automatic image annotation, the manual image annotation remains the trustworthy way for assigning textual descriptions to images. In this paper we propose a framework in which factors such as the visual content, used lexicon and annotation method are investigated how they affect the annotations collected through crowdsourcing. The preliminary results show that annotations using hierarchical structured lexicon or free keywords are more consistent than annotations based on lexicon keywords. Nevertheless, the kappa statistics allow an agreement of 0.83 on average for the lexicon and 0.86 for XML lexicon keywords. The numerous uses of free keywords show the lack of understanding of the pre-selected keywords and their limited ability to describe the content the dataset. The large amount of out of content keywords indicates that the annotation of cultural heritage data is often difficult for non experts. Furthermore, the majority of out of content keywords was suggested for images have abstract content. However, the combination of the proposed annotation methods can achieve competitive results. The experimental results are very encouraging and indicate that image annotation tasks can be outsourced online or addressed by tags created online while keeping the annotation quality in high standards and achieving wide and diverse participation. The proper image annotation can lead to efficient data retrieval related to culture heritage and can also be used for further research in the domain of crowdsourcing annotation and data retrieval. In addition, the outcomes of this work can contribute to better understanding which circumstances affect the quality of crowdsourced image annotation.

Our future work includes the vocabulary keywords enrichment with the most frequent suggested free keywords. The annotation platform will be extended to collect more data related to annotation such as the timestamp when a keyword is proposed for a specific image, etc. This information will give new insights into annotators consistency since the keywords proposed first may have more possibilities to be relevant to the image

content, as opposed to the keywords proposed later. In addition, the future work will also include the experimentation on larger datasets and data collected through online social networks. Finally, the demographics of annotators will be explored in which extend influence the annotation quality.

Acknowledgements This work has been partly supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 739578 (RISE – Call: H2020-WIDE-SPREAD-01-2016-2017-TeamingPhase2) and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

References

1. <https://commandaria.cut.ac.cy//>
2. Randolph, J. J.(2008). Online Kappa Calculator. Retrieved April 5, 2019, from <http://justusrandolph.net/kappa/>
3. Ahn, L.V., Maurer, B., McMillen, C., Abraham, D., Blum, M.: recaptcha: Human-based character recognition via web security measures. *Science* **321**(5895), 1465–1468 (2008)
4. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S.: Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* **17**(2), 76–81 (2013)
5. Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* **36**(1), 15–24 (2015)
6. Brabham, D.: Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* **14**(1), 75–90 (2008)
7. Brants, T.: Inter-annotator agreement for a german newspaper corpus. In: Proc. of the 2nd International Conference on Language Resources and Evaluation, pp. 1–5. Athens, Greece (2000)
8. Brawley, A.M., Pury, C.L.S.: Work experiences on mturk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* **54**, 531 – 546 (2016)
9. Callison-Burch, C.: Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In: Proc. of Conference on Empirical Methods in Natural Language Processing, pp. 286–295. Singapore (2009)
10. Chen, K.T., Wu, C.C., Chang, Y.C., Lei, C.L.: A crowdsourcing evaluation framework for multimedia content. In: Proc. of the 17th ACM international conference on Multimedia, pp. 491–500. Beijing, China (2009)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 834–848 (2018)
12. Cheng, Q., Zhang, Q., Fu, P., Tu, C., Li, S.: A survey and analysis on automatic image annotation. *Pattern Recognition* **79**, 242 – 259 (2018)
13. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46 (1960)

14. Cowles, M., Davis, C.: On the origins of the .05 level of statistical significance. *American Psychologist* **37**(5), 553–558 (1982)
15. Dutta, A., Verma, Y., Jawahar, C.V.: Automatic image annotation: the quirks and what works. *Multimedia Tools and Applications* **77**(24), 31991–32011 (2018)
16. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video (2019)
17. Fujisawa, S.: Automatic creation and enhancement of metadata for cultural heritage. In: Bull. IEEE Tech. Committee on Digital Libraries (TCDL) (2007)
18. Ghezzi, A., Gabelloni, D., Martini, A., Natalicchio, A.: Crowdsourcing: A review and suggestions for future research. *International Journal of Management Reviews* **20**(2), 343–363 (2018)
19. Glowacz, A.: Acoustic-based fault diagnosis of commutator motor. *Electronics* **7**(11) (2018)
20. Glowacz, A.: Fault diagnosis of single-phase induction motor based on acoustic signals. *Mechanical Systems and Signal Processing* **117**, 65 – 80 (2019)
21. Gulati, P., Yadav, M.: A novel approach for extracting pertinent keywords for web image annotation using semantic distance and euclidean distance. In: M.N. Hoda, N. Chauhan, S.M.K. Quadri, P.R. Srivastava (eds.) *Software Engineering*, pp. 173–183. Springer Singapore, Singapore (2019)
22. Hanbury, A.: A survey of methods for image annotation. *Journal of Visual Languages & Computing* **19**(5), 617–627 (2008)
23. Hare, J.S., Lewis, P.H., Esner, P.G.B., J., S.C.: Mind the gap: Another look at the problem of the semantic gap in image retrieval. In: Proc. of Multimedia Content Analysis, Management and Retrieval 2006 SPIE. San Jose, California, USA (2006)
24. Heidorn, P.B.: Image retrieval as linguistic and nonlinguistic visual model matching. *Library Trends* **48**(2), 303–325 (1999)
25. Hong, S., Choi, J., Feyereisl, J., Han, B., Davis, L.S.: Joint image clustering and labeling by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(7), 1411–1424 (2016)
26. Howe, J.: The rise of crowdsourcing. *Wired Magazine* **14**(6), 176–183 (2006)
27. Howe, J.: *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business (2008)
28. Huang, Y., Yang, H., Qi, X., Malekian, R., Pfeiffer, O., Li, Z.: A novel selection method of seismic attributes based on gray relational degree and support vector machine. *PLOS ONE* **13**(2), 1–16 (2018)
29. Jin, C., Sun, Q.M., Jin, S.W.: A hybrid automatic image annotation approach. *Multimedia Tools and Applications* **78**(9), 11815–11834 (2019)
30. Jing, X., Wu, F., Li, Z., Hu, R., Zhang, D.: Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing* **25**(6), 2712–2725 (2016)
31. Joachims, T., Granka, L., Pang, B., Hembrooke, H., G., G.: Accurately interpreting clickthrough data as implicit feedback. In: Proc. of the 28th Annual International ACM SIGIR Conference, pp. 154–161. Salvador, Brazil (2005)
32. Jr., F.F.: *Survey Research Methods*, 5 edn. SAGE Publications Inc. (2014)
33. Kilgarriff, A.: Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language* **12**(3), 453–472 (1998)
34. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in wikipedia: quality through coordination. In: Proc. of the 2008 ACM conference on Computer supported cooperative work, pp. 37–46. San Diego, CA, USA (2008)
35. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K.: Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision* **10**(3), 177–243 (2016)
36. Kwasnicka, H., Paradowski, M.: Machine learning methods in automatic image annotation. In: *Advances in Machine Learning II, Studies in Computational Intelligence*, vol. 263, pp. 387–411 (2010)
37. Landis, J.R., Koch, G.K.: The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977)
38. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
39. dit Leksir, Y.L., Mansour, M., Moussaoui, A.: Localization of thermal anomalies in electrical equipment using infrared thermography and support vector machine. *Infrared Physics & Technology* **89**, 120 – 128 (2018)
40. Li, A., Sun, J., Ng, J.Y., Yu, R., Morariu, V.I., Davis, L.S.: Generating holistic 3d scene abstractions for text-based image retrieval. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1942–1950 (2017)
41. Lovett, M., Bajaba, S., Lovett, M., Simmering, M.J.: Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of amazon’s mechanical turk masters. *Applied Psychology* **67**(2), 339–366 (2017)
42. Ma, Y., Liu, Y., Xie, Q.: Cnn-feature based automatic image annotation method. *Multimedia Tools and Applications* **78**(3), 3767–3780 (2019)
43. Macdonald, C., Ounis, I.: Usefulness of quality click-through data for training. In: Proc. of the 2009 Workshop on Web Search Click Data, pp. 75–79. Barcelona, Spain (2009)
44. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55**(2), 645–657 (2017)
45. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Proc. of European Conference on Computer Vision, pp. 316–329. Marseille, France (2008)
46. Matusiak, K.K.: Towards user-centered indexing in digital image collections. *OCLC Systems & Services* **22**(4), 283–298 (2006)
47. McCredie, M.N., Morey, L.C.: Who are the turkers? a characterization of mturk workers using the personality assessment inventory. *Assessment* (2018)
48. Nazir, A., Ashraf, R., Hamdani, T., Ali, N.: Content based image retrieval system by using hsv color histogram, discrete wavelet transform and edge histogram descriptor. In: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1–6 (2018)
49. Nguyen, D.T., Hua, B., Yu, L., Yeung, S.: A robust 3d-2d interactive tool for scene segmentation and annotation. *IEEE Transactions on Visualization and Computer Graphics* **24**(12), 3005–3018 (2018)
50. Nowak, S., Ruger, S.: How reliable are annotations via crowdsourcing a study about inter-annotator agreement for multi-label image annotation. In: Proc. of the International Conference on Multimedia Information Retrieval, pp. 557–566. Philadelphia, PA, USA (2010)
51. Papadopoulos, K., Tsapatsoulis, N., Lanitis, A., Kounoudes, A.: The history of commandaria: Digital journeys back to time. In: Proc. of the 14th International Conference on Virtual Systems and Multimedia. Limassol, Cyprus (2008)

52. Penna, A., Mohammadi, S., Jojic, N., Murino, V.: Summarization and classification of wearable camera streams by learning the distributions over deep features of out-of-sample image sequences. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4336–4344 (2017)
53. Perina, A., Mohammadi, S., Jojic, N., Murino, V.: Summarization and classification of wearable camera streams by learning the distributions over deep features of out-of-sample image sequences. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
54. Quintero, R., Parra, I., Lorenzo, J., Fernández-Llorca, D., Sotelo, M.A.: Pedestrian intention recognition by means of a hidden markov model and body language. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–7 (2017)
55. R., A.: Inter-annotator agreement. In: N. Ide, P. J. (eds.) Handbook of Linguistic Annotation. Springer, Dordrecht (2017)
56. Randolph, J.J.: Free-marginal multirater kappa: An alternative to fleiss' fixed-marginal multirater kappa. In: In Joensuu University Learning and Instruction Symposium. Joensuu, Finland (2005)
57. Raykar, V., Zhao, S., Yu, L., Jerebko, A., Florin, C., Valadez, G., Bogoni, L., Moy, L.: Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In: Proc. 26th Annual International Conference on Machine Learning, pp. 889–896. Montreal, Canada (2009)
58. Ristin, M., Guillaumin, M., Gall, J., Gool, L.V.: Incremental learning of random forests for large-scale image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(3), 490–503 (2016)
59. S. Piramanayagam W. Schwartzkopf, F.W.K.E.S.: Classification of remote sensed images using random forests and deep learning framework (2016)
60. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 614–622. Las Vegas, NV, USA (2008)
61. Smyth, P., Fayyad U. amd Burl, M., Perona, P., Baldi, P.: Inferring ground truth from subjective labeling of venus images. *Advances in Neural Information Processing Systems* **7**, 1085–1092 (1995)
62. Snow, R., O Connor, B., Jurafsky, D., Ng, A.: Cheap and fast but is it good evaluating nonexpert annotations for natural language tasks. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Honolulu, HI, USA (2008)
63. Theodosiou, Z., Kasapi, C., Tsapatsoulis, N.: Semantic gap between people: An experimental investigation based on image annotation. In: Seventh International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 73–77. Luxembourg (2012)
64. Theodosiou, Z., Tsapatsoulis, N.: Crowdsourcing annotation: Modelling keywords using low level features. In: Proc. of the 5th International Conference on Internet Multimedia Systems Architecture and Application. Bangalore, India (2011)
65. Tsikrika, T., Diou, C., De Vries, A.P., Delopoulos, A.: Image annotation using clickthrough data. In: Proc. of the 8th International Conference on Image and Video Retrieval, pp. 1–8. Santorini, Greece (2009)
66. Tyagi, V.: Content-Based Image Retrieval Techniques: A Review, pp. 29–48. Springer Singapore, Singapore (2017)
67. Vijayanarasimhan, S., Grauman, K.: What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2262–2269. Miami, FL, USA (2009)
68. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
69. Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost effective labels. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 25–32. San Francisco, CA, USA (2010)
70. Whitehill, J., Ruvolo, P., Bergsma T. Wu, J., Movellan, J.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Proc. 23rd Annual Conference on Neural Information Processing Systems, pp. 2035–2043. Vancouver, Canada (2009)
71. Wigness, M., Draper, B.A., Beveridge, J.R.: Efficient label collection for image datasets via hierarchical clustering. *International Journal of Computer Vision* **126**(1), 59–85 (2018)
72. Xie, F., Fan, H., Li, Y., Jiang, Z., Meng, R., Bovik, A.: Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Transactions on Medical Imaging* **36**(3), 849–858 (2017)
73. Yadav, P., Jezek, E., Bouillon, P., Callahan, T., Bada, M., Hunter, L., Cohen, K.B.: Semantic relations in compound nouns: Perspectives from inter-annotator agreement. *Studies in health technology and informatics* **245**, 644–648 (2017)
74. Yang, C.M., Choo, Y., Park, S.: Semi-automatic image and video annotation system for generating ground truth information. In: 2018 International Conference on Information Networking (ICOIN), pp. 821–824 (2018)
75. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. *Pattern Recognition* **45**, 346–362 (2012)
76. Zhang, R., Zhang, Z., Li, M., Zhang, H.J.: A probabilistic semantic model for image annotation and multi-modal image retrieval. *Multimedia Systems* pp. 27–33 (2006)
77. Zhu, L., Shen, J., Xie, L., Cheng, Z.: Unsupervised visual hashing with semantic assistant for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering* **29**(2), 472–486 (2017)