Computer Networks xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks

Q1

Jasper Oosterman[a,*], Jie Yang[a], Alessandro Bozzon[a], Lora Aroyo[b], Geert-Jan Houben[a]

[a] Delft University of Technology, P.O. Box 5, 2600 AA, Delft, The Netherlands
[b] VU University, De Boelelaan 1081a, 1081 HV, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 31 October 2014
Revised 14 June 2015
Accepted 3 July 2015
Available online xxx

Keywords:
Crowdsourcing
Cultural heritage
Artwork annotation

ABSTRACT

Cultural heritage institutions more and more provide online access to their collections. Collections containing *visual* artworks need detailed and thorough annotations of the represented visual objects (e.g. plants or animals) to enable human access and retrieval. To make these suitable for access and retrieval, visual artworks need detailed and thorough annotations of the visual classes. Crowdsourcing has proven a viable tool to cater for the pitfalls of automatic annotation techniques. However, differently from traditional photographic image annotation, the *artwork* annotation task requires workers to possess the knowledge and skills needed to *identify* and *recognise* the occurrences of visual classes. The extent to which crowdsourcing can be effectively applied for artwork annotation is still an open research question. Based on a real-life case study from Rijksmuseum Amsterdam, this paper investigates the performance of a crowd of workers drawn from the CrowdFlower platform. Our contributions include a detailed analysis of crowd annotations based on two annotation configurations and a comparison of these crowd annotations with the ones from trusted annotators. In this study we apply a novel method for the automatic aggregation of local (i.e. bounding box) annotations, and we study how different knowledge extraction and aggregation configurations affect the identification and recognition aspects of artwork annotation. Our work sheds new light on the process of crowdsourcing artwork annotations, and shows how techniques that are effective for photographic image annotation cannot be straightforwardly applied to artwork annotation, thus paving the way for new research in the area.

1. Introduction

Visual *artwork*[1] annotation recently emerged as an important multidisciplinary discipline, fuelled by the growing needs of cultural heritage institutions. Galleries, Libraries, Archives, and Museums (GLAMs) have the mission of ensuring that the art produced by mankind is properly preserved, described, catalogued, and made accessible to the public. To unlock the value of their artwork collections, GLAMs must enable and facilitate browsing and retrieval for a broad yet unforeseen variety of users, having an unknown variety of needs. To this end, textual annotations are used to describe the instances of classes of *visual objects*, e.g. objects, plants, animals and human body parts, represented in the artworks. Image annotation is a notoriously hard problem for computers to solve but, thanks to recent progress in computer vision techniques [2,3], it is now possible to correctly identify

* Corresponding author. Tel.: +31152786346.
E-mail address: j.e.g.oosterman@tudelft.nl, jasper.oosterman@gmail.com (J. Oosterman).

[1] A visual artwork is an artistic expression represented on a flat surface (e.g., canvas or sheet of paper) in the form of a painting, printing or drawing [1].

http://dx.doi.org/10.1016/j.comnet.2015.07.008
1389-1286/© 2015 Published by Elsevier B.V.

the presence of several visual classes in *photographic* images. Alas, such techniques cannot be applied in cultural heritage collections due to the unique nature of *visual artworks* [1]. Therefore, GLAMs employ professionals, mostly art historians, to analyse artworks and create annotations about the occurrence of visual classes of interest; but the quality and extent of their annotation work is subject to temporal, monetary, and knowledge limitations.

Crowdsourcing has emerged as a viable solution to complement (or substitute) computer vision algorithms for many "difficult" visual analysis tasks, including the annotation of visual content [4–7]. Crowdsourced artwork annotation is a representative example of a *Crowdsourced Knowledge Creation* (CKC) task [8], i.e. a class of crowdsourcing tasks where workers are requested to stress their high-level cognitive abilities (e.g., knowledge synthesis, data interpretation), and draw from their experience or education, in order to solve problems for which a unique, factual solution might not exist. In the case of artwork annotation, a crowd worker must possess the *knowledge* and *skills* required to: (1) understand the abstract, symbolic, or allegorical interpretation of the reality depicted in the artwork to **identify** the occurrences of visual classes; and (2) **recognise** the type of such visual classes, describing them with an expressive text.

Crowdsourcing of artwork annotation is still an open research challenge.

Domain-specific experts are hard and expensive to recruit. Knowledgeable contributors (e.g. pro-amateurs and enthusiasts) might be present in anonymous human computation marketplaces, but must be located, engaged and motivated.

The **identification** and **recognition** of visual classes are aspects of artwork annotation that can be influenced by the CKC process design. The knowledge *extraction* step is of great importance, as it requires work interfaces that guide, but not constrain, their high-level cognitive and memory processes of contributors. This is in contrast to traditional "computational" crowdsourcing tasks, where a well-defined work interface guarantees execution efficiency and consistency. Also, the *aggregation* of knowledge from individual workers must account for the broad diversity of opinions and interpretations that a crowd knowledge elicitation task might imply. This work studies the crowdsourcing of artwork annotation, and addresses the following research questions:

RQ1: Can non-professional annotators from crowdsourcing platforms provide high quality artwork annotations?

RQ2: To what extent can the extraction and aggregation steps of a crowdsourced knowledge creation process influence the identification and recognition aspects of visual artwork annotation?

To answer these questions, we partnered with the Rijksmuseum Amsterdam[2], and set-up an extensive evaluation campaign aimed at testing the performance of workers from human computation platforms when asked to identify and recognise occurrences of visual object classes in artworks. We assembled a collection of 80 Rijksmuseum prints, and focused on the "flowers" class, to target an area of expertise that is likely present in a general population. Three trusted assessors created a reference annotation ground-truth to assess both the number, type, and location of flowers depicted in the dataset[3]. To test the effect of the CKC's *knowledge extraction* step, we evaluated two annotation configurations: an *Artwork–centric* configuration where textual annotations about visual objects are specified for the whole artwork; and a *Class–centric* configuration where occurrences of visual objects are identified using bounding boxes with distinct textual annotations.

The experiment was performed on the CrowdFlower human computation marketplace, and involved a crowd of 235 workers. For each *knowledge extraction* configuration, we tested the impact of different *aggregation* methods on the identification and recognition performance. We analyse the quality of annotations provided crowd workers to study the richness of their vocabulary, and its overlap with respect to annotations created by domain experts.

The main contribution of this paper is a study on how extraction and aggregation methods affect annotation quality of visual artworks in a Artwork–centric and Class–centric configuration. To enable our study we created a novel algorithm for aggregating annotations in the Class–centric configuration.

Results confirm the unique nature of the artwork annotation problem, showing how crowdsourcing techniques that are effective for photographic image annotation cannot be straightforwardly applied. The high percentage of workers who dropped out during recruitment testifies to the challenges related to the identification of visual objects, even when as simple as flowers. The experiments highlight the impact that the CKC process can have on the identification and recognition quality: a *Artwork–centric* configuration enhances recognition aspects, and comes with a richer annotation vocabulary; on the other hand, a *Class–centric* configuration guarantees better identification performance, but poorer recognition and vocabulary.

The remainder of the paper is structured as follows. In Section 2 we present the related work. Section 3 details and exemplifies the complexities of visual object identification and recognition in artwork annotation. Next, Section 4 describes the design and execution of our evaluation. Sections 5 and 6 present and discuss experimental results. Section 7 concludes and sets the scene for future work.

2. Related work

Recent literature [1,9] shows how in contrast to photographic images, which carefully represent the real world, artworks provide less and typically inconsistent visual information (texture, colour, depth, etc.); this, together with the lack of sufficiently large training sets and the presence of a sizeable number of visual classes to be recognised, are among the main causes for ineffective automatic artwork annotation algorithms. *Hybrid image annotation* methods [10,11] emerged as a promising solution to reduce costs and error rate by complementing automatic techniques with crowdsourced annotations. Inspired by these works, our paper focuses on

[2] http://rijksmuseum.nl. The Rijksmuseum Amsterdam is the largest and most prestigious museum in the Netherlands.

[3] The dataset and the results of this study are available for download from http://bit.ly/CN-SI-artworks.

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008

crowdsourcing visual object occurrence annotations, a form of annotation where strict quality control and effective aggregation techniques are needed to cater for the natural difficulties related to the drawing of geometric coordinates to bound image objects [12].

Previous works investigated how crowds can support the artwork annotation process [13–15]. For instance, "The Steve project" [13] studied crowd tagging of collections from more than 12 USA-based museums and compared crowd and professional taggers. Authors found crowd annotators, drawn from museum attendees, to use a different vocabulary than professional ones, but that such annotations were effective to improve the retrieval of the artworks. In [15], crowds without prior domain knowledge were engaged to annotate prints, while gaming mechanisms stimulated them to learn about the domain. Based on the gaming mechanism introduced by Luis von Ahns popular ESP-game [16], several games with a purpose [17,18] have been proposed to collect artwork metadata. Differently from previous works, which exploited domain-knowledgeable volunteers or museum attendees, our approach explicitly focuses on crowds drawn from human computation platforms, i.e. anonymous individuals for which no assumption can be made about their familiarity with artworks or their domain knowledge.

Recent studies [4,19,20] compared the performance of expert and human annotators from human computation platforms. All studies agreed on the potential and the scalability and reduced costs of crowds compared to experts, but also mentioned that additional actions, such as repetition and worker qualification, are needed to obtain high quality annotations. Standard aggregation techniques for crowd results include removing results failing qualification tests and subsequently using majority voting to combine the results [4].

Most studies however focused on tasks that required workers to have only basic skill and common knowledge. There is demand for more complex tasks, for example requiring creativity [21]. Only recently, several works advocated for specialised crowdsourcing techniques for knowledge creation tasks, such as *nichesourcing* [22] or *community sourcing* [23,24]. For instance, in the context of domain-specific ontology and taxonomy creation, Noy et al and Chilton et al. [25,26] found a crowd performance of around 80% correctness which, although being lower than that of domain expert, was very promising and above all scalable.

The results described in this paper are rooted in our previous work on crowdsourced knowledge creation. The early work [27] defined the problem of artwork annotation in the context of the Rijksmuseum Amsterdam in the Netherlands. The subsequent works [8,28] introduced an experimental methodology, and reports on preliminary results mainly focused on Artwork–centric knowledge extraction. To the best of our knowledge, our work is the first one that systematically studies the performance of Artwork–centric and Class–centric knowledge extraction techniques in human computation platforms, and assess their identification and location performance with respect to a high-quality ground truth.

3. Challenges in artwork annotation

This section elaborates on the major challenges related to the annotation of visual objects in artworks. We describe the process currently employed at the Rijksmuseum Amsterdam, highlighting typical annotation requirements, and exemplifying how the compliance to such requirements is hindered by the nature of visual artworks.

3.1. The need for professional annotation of artworks

The Rijksmuseum has a collection of over 1 million artworks, 700, 000 of which are prints, that the museum wants to make accessible for online consumption. The museum currently conducts the following digitization process. First, a high quality digital representation is created. Then, a team of 6 professional, in-house, annotators describe the artwork. During discussions with the museums curator of the online collection the museums' interest regarding artwork annotation was stated: descriptions of both the art-historical aspects (such as creator, material and date of creation) and the depicted *visual objects*, such as depicted persons, buildings, flora and fauna. People using or studying their collection often try to answer questions regarding a visual object class `X` such as: "How many prints depict `X`?"; "Which are the type of `X` most commonly represented in a given artistic period?"; or "How are `X` depicted in different genres and periods?".

To enable these retrieval scenarios, annotations must possess the following 2 properties: (1) *coverage*, i.e. *all* instances of the visual object classes represented in the artwork should be *identified*, (possibly) located, and annotated, and (2) *expressiveness*, i.e. visual objects should be *recognised* and annotated with texts that should serve a *broad* spectrum of knowledge levels; this is to allow both common and expert users to access the collection by using the most familiar language.

Professional annotators are given 25 min to retrieve an artwork from storage, analyse it, find relevant information and publications online and in their library, and describe by entering annotations and references in collection management software[4]. Twenty minutes are devoted to the description of art-historical aspects, while only 5 min are allocated to visual object classes. The latter involves two steps: the *identification* of instances of a given class, and subsequently, their *recognition* and association with a representative label.

The flower annotation case study. The size and importance of the Rijksmuseum make it an exemplary case of cultural heritage institution striving for high-quality annotation of digital (digitized) collections.

Let us consider the print in Fig. 1, drawn from the 700*K* prints collection. According to the requirements defined before, professional annotators from the Rijksmuseum must identify all the instances of depicted flowers, and describe them with their common and scientific names. The print depicts a woman holding flowers sitting in a flower decorated cart pulled by dogs; the print contains 71 flower instances, distributed in at least 8 areas (highlighted with black bounding boxes), not considering the flowers decorating the cart or the decorations resembling flowers on the cart's wheels.

While focusing on prints and flower annotations, we argue this use case to be representative of a broader class of

[4] The allocated time is constrained by budgetary considerations and cannot be significantly increased.

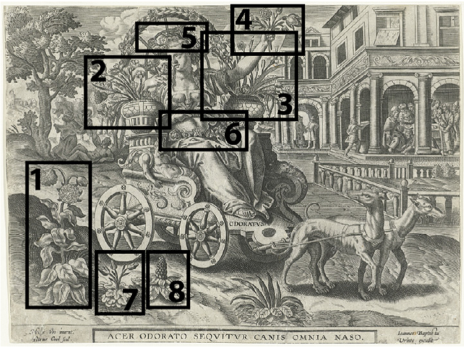


Fig. 1. "Scent" by Petrus Cool.

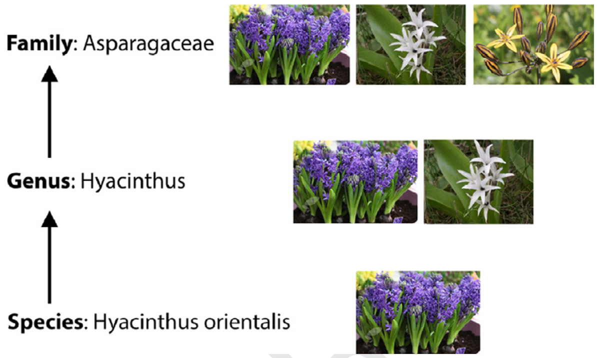


Fig. 2. An extract of the flower taxonomy for box 8 of Fig. 1.

artwork annotation problem: on the one hand, it reflects a typical annotation subject, i.e. a non-photographic representation of (a possibly symbolic) reality containing many, diverse visual objects. On the other hand, "flower" is an example of a common visual object class, which is not related to historical aspects, and for which frequency and specificity are a major bottleneck in the manual annotation process. The next sections describe two main challenges that affect the creation of qualitative annotations.

3.2. Challenges in visual object identification

Visual artworks often provide an abstract, symbolic, or allegorical interpretation of reality. In such context, the *identification* of *all* visual object occurrences is a very time-consuming and error-prone task, complicated by: (1) the lack of colours or details; (2) the abstract or stylised representation of the visual class occurrence; (3) the size, density, or composition of the depicted visual objects; and (4) subjective or personal interpretations. In such conditions, in order to identify all the occurrences of visual objects, an annotator must show both commitment to the annotation task (to account for the potentially high number of occurrences), and some degree of experience in the art domain, to be able to correctly infer the content of visual artworks.

3.3. Challenges in visual object recognition

To correctly *recognise* visual objects, and describe them with expressive text, domain-specific expertise is often required. Let us consider the domain of flowers: arguably, everyone is exposed, to some extent, to knowledge about flowers: in the mind of the writers, it is difficult to imagine someone not being able to recognise the red flower in Fig. 2 as a rose. However, going beyond such a shallow description requires domain-specific knowledge. Would the reader be able to tell a *Rosa canina* from a *Rosa multiflora*?[5]. And which terminology would the reader be able to use?

In our case study, the Rijksmuseum is interested in annotating each flower instance at different levels of specificity, according to the flowers (formally: *plants*) taxonomy. It ranges from the top element `Kingdom`, for example *Plantae* (all plants), to the most specific element `Species`, e.g. *Hyacinthus orientalis*, one plant. Fig. 2 depicts part of the flower description taxonomy. Above the`species` level is `genus`, which describes multiple flower `species`, e.g. the *Hyacinthus* genus. Above `genus` is `family`, which describes multiple flower `genus`, e.g., the *Asparagaceae* family. Annotations could use both the `common`, e.g. *Dutch hyacinth*, or the `botanical`, e.g. *Hyacinthus orientalis*, flower name.

4. Experimental design

Our goal is to study the annotation coverage and accuracy of non-professional annotators drawn from a crowd of anonymous workers, and to analyse their performance under different knowledge creation process configurations. To this end, we instrumented an extensive evaluation campaign, discussed in this section.

Section 4.1 describes the experimental dataset; Section 4.2 introduces the two annotation configurations subject of our study; Section 4.3 provides details about the setup and quality control mechanism of the experiments performed on the CrowdFlower platform; Section 4.4 describes the procedure employed for the normalisation of crowd annotations; Section 4.5 describes the gathering of annotations from domain experts; finally, Section 4.6 introduces the aggregation methods and evaluation metrics used for quantitative assessment.

4.1. Experimental dataset

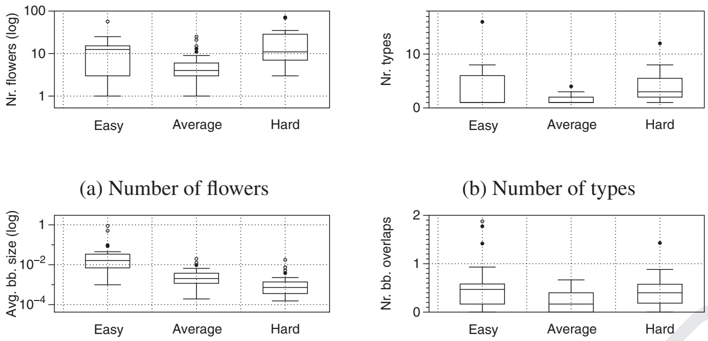
In collaboration with personnel of the Rijksmuseum Amsterdam, we selected 80 prints containing at least one flower instance. We then instrumented a ground-truth creation task, aimed at defining, for each print, the exact number and location of each contained flower instance.

We recruited 3 *trusted annotators*. They were selected from the SEALINCMedia project[6] staff; we considered individuals familiar with the targeted collections and with the technology used for image annotation, namely the

[5] Both *Rosa canina* and *Rosa multiflora* belong to the *rosa* genus and, through the eyes of a non-expert, they look pretty similar.

[6] https://sealincmedia.wordpress.com. SEALINCMedia is part of the Dutch national program COMMIT.

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008



(a) Number of flowers (b) Number of types

(c) Avg. bounding box size (d) Avg. overlapped bounding box ratio

Fig. 3. Distributions of the number of flowers and types and bounding box statistics per print difficulty class based on the ground-truth data created by the trusted annotators.

Annotorious library.[7] Annotorius allows users to draw (square) bounding boxes on images. Each trusted annotator was instructed to identify, in each print, all flowers conforming to the following definition:

Definition. A flower is considered to be the flowering part of a plant with petals and distinguishable from leaves. A branch can have multiple flowers (but each of those flowers has the same name). A flower bud counts as a flower.

For each print, they were asked to independently: (1) draw a bounding box of each flower instance; and, (2) count the number of different flower types depicted in the print. To guarantee maximum correctness and precision, annotators were given no time limits, and they were allowed to stop and resume at any time.

Upon completion of the independent annotation sessions we gathered the results and created a new version of each print featuring the bounding boxes from each annotator. We then organised three deliberation sessions (in total 8 h, spread over 3 days) where annotators were asked to discuss their work. For each print in the dataset, they needed to agree on a unique set of bounding boxes, and on a unique number of flower types. For each bounding box, the annotators were asked to also agree on its location and size; they were asked to redraw each bounding box as accurately as possible, so to guarantee that each flower occurrence was fully, but minimally, contained by it. Trusted annotators were also asked to provide comments and remarks about the issues they faced during the annotation process, including properties of the annotated prints such as the presence of little flowers, low contrast flower and leaves, flower orientation and overlap, that could have hindered the recognition activity.

These comments were uses to support an additional discussion session aimed at classifying each print in the dataset, according to the **difficulty** encountered in its annotation. We identified two difficulty dimensions, related to identification and recognition, while lead to three difficulty classes: `easy`, which identifies prints where flower instances are easy to identify and recognise; the class features 20 prints; `average` which identifies prints where identification is easy, but recognition is hard — 29 prints; and `hard`, which contains prints where both identification and recognition are hard to perform — 31 prints. Notably, no print in the dataset contained flower instance that were `easy` to recognise but `hard` to identify.

This carefully crafted annotation work led to the creation of 1047 bounding boxes. On average, each of the 80 prints contains 13 flower instances ($\sigma = 15$, $Max = 71$) and 3 flower types ($\sigma = 3$, $Max = 16$). Figs. 3a–d depict the distribution, for each print in the dataset, of: (1) the number of identified flower instances; (2) the number of unique flower types; (3) the average size of the annotation bounding boxes ; and (4) the average number of overlapping bounding boxes. Prints in the `average` difficulty class (easy to identify, hard to recognise) have, on average, the least number of flowers and overlapping bounding boxes. The annotation difficulty is strictly related to the average size of the bounding boxes. The number of flower types is similarly distributed across difficulty class.

4.2. Experimental configurations

We consider two experimental configurations, each testing a different knowledge extraction modality. In the *Artwork–centric* configuration, crowd workers define annotations on the artwork as a whole; in contrast, the *Class–centric* requires annotations to be defined for each visual object instance, thus including its exact location in the artwork. In both configurations, annotators are asked to *identify* and *recognise* the occurrences of objects in a given visual class. Then, for each configuration, we assess how different aggregation methods impact on the quality of the resulting annotations.

Artwork–centric knowledge extraction. In this configuration, we ask annotators to analyse the artwork as a whole. *Identification* is performed by counting the occurrences of the visual objects of interest, and the number of object class types therein represented. *Recognition* also occurs globally, by asking the worker to provide labels for each distinct type of object types identified in the print. The Artwork–centric

[7] http://annotorious.github.io/. Annotorius is a javascript image annotation library developed in the context of the Europeana project. Europeana is a large European content aggregator for Cultural Heritage and strives to make Cultural Heritage more accessible.

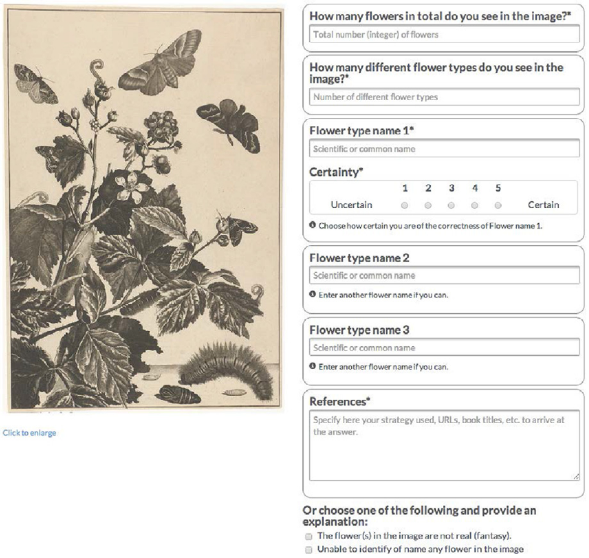


Fig. 4. Artwork–centric annotation user interface. Static image left and annotation fields right.

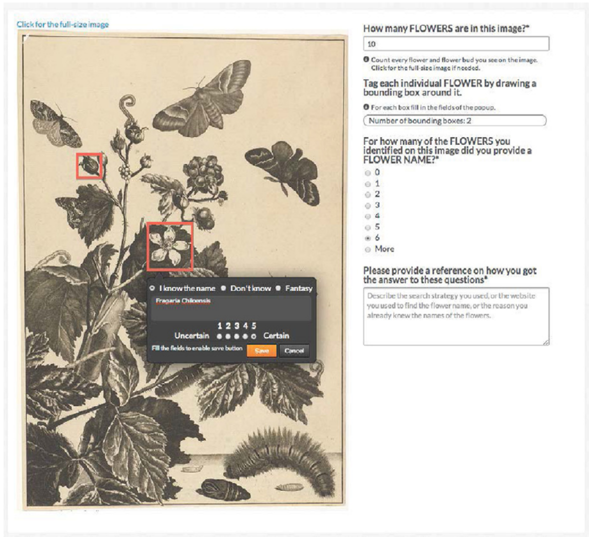


Fig. 5. Class–centric annotation user interface. Interactive image on which bounding boxes are drawn left. Annotation fields appear per drawn bounding box.

annotation configuration has as advantage the simplicity and speed by which an annotator can perform the assigned task. The main drawback, on the other hand, is that workers are not allowed to provide information about where and how (in terms of size, clarity, etc.) in the artwork a given object instance is represented. Likewise, labels are not directly associated with occurrences of the visual object class, thus providing no insights about their importance or role in the artwork. Fig. 4 depicts the user interface used in our evaluation for Artwork–centric annotation. The print to annotate is presented on the left-hand side. The worker can open the image at full screen by clicking on it. The right hand side lists the annotation fields. Workers are requested to indicate the number of flower instances, and the number of distinct flower types they identify in the print. Three text fields are available to provide up to three labels describing the flower types in the print. To avoid biases in the knowledge extraction process, no auto-suggestions functionality was provided: workers had to manually specify flower names, both in their `botanical` *or* `common` form.

To account for unidentified flower types, workers can report a print to contain only flower occurrences of imaginative nature (*fantasy*); or simply indicate their inability to specify any suitable label (*unable*). In both cases, an additional text box is prompted to collect comments about the reasons for their judgment.

Class–centric knowledge extraction. This configuration improves on the Artwork–centric one by providing workers with the ability to pinpoint the occurrences of the sought visual object in the artwork. Fig. 5 depicts the user interface used to collect Class–centric annotations: using the Annotorius library, workers were asked to draw *bounding boxes* directly on the image. Labels are specified for each identified occurrence, thus allowing for more fine-grained and localised annotations. A bounding box is supposed to fully contain the identified flower occurrence. Like in the Artwork–centric user interface, on the right hand side a set of annotation fields to allow workers to provide the number of occurrences of the objects of interest; and the number of distinct flower types for which a distinct label has been provided. An additional field is automatically filled with the number of bounding boxes drawn for the current print. Such fields are defined for quality control purposes, but are also used to further study the behaviour of crowd annotators.

4.3. Crowdsourced artwork annotation task design

The experiment was performed on *CrowdFlower*[8], a human computation platform that recruits workers worldwide.

We set up two annotation jobs, one for each annotation configuration. Each task required the annotation of five prints, and was rewarded with € 0.18. Compensation has been defined after several pilot experiments, used to test-drive the annotation interfaces and to receive feedback on the monetary reward (which resulted in a 3 out of 5 score). The moderate sum was set so to attract people intrinsically motivated while still giving some monetary appreciation. Workers were recruited in the *Level 1 Contributors* class of CrowdFlower[9] to exclude workers known by CrowdFlower to have a low quality.

Quality control mechanisms. Each annotation task was introduced by a description page, listing instructions about the requested input, an explanation of the user interface, examples of target annotations, and the same flower definition given to trusted annotators.

Each worker had to first perform a *qualification task*, as a necessary pre-condition to participate in the rest of the

[8] http://crowdflower.com

[9] Level 1 contributors are “high performance contributors who account for 60% of monthly judgments and maintain a high level of accuracy across a basket of CrowdFlower jobs.”

experiment. In the qualification task, workers were asked to annotate 5 prints, sampled from the ones belonging to the `easy` difficulty class, and having an unambiguously number of flowers and flower types, as determined by the trusted annotators. Each task paid € 0.18 upon completion. Workers were evaluated against the *number of flowers* and *number of types* defined in the ground truth; we allowed an *off-by-one* error to account for small counting errors. Workers that failed to correctly annotate more than one print in the qualification task were blocked from performing further tasks. Successful workers were allowed to continue with other annotation tasks; to avoid learning bias, prints used in the qualification tasks were not shown twice to the same worker. For the same reason, workers that performed tasks in the Artwork–centric job were not allowed to perform tasks in the Class–centric job, and vice-versa. As an additional quality control mechanism one hidden control question, a so-called honeypot, was added to each annotation task. Each control question contained a print with an unambiguously number of flowers and flower types were used, as determined by the trusted annotators. Workers were paid regardless of their performance with the control question. Prints were shown to workers in random order. Each worker could annotate every remaining print in the set, but only once. We designed the task such that each print could be annotated by at most 10 annotators. If the overall accuracy of a worker across all executed tasks on control questions dropped below 60%, then the annotations from such workers were discarded.

4.4. Annotation labels pre-processing

Labels produced underwent a pre-processing step aimed at providing a uniform label dictionary and, thus, support the subsequent analysis. Due to the employment of a worldwide workforce, labels were assumed to be provided in multiple languages, and to include spelling errors. To achieve maximum accuracy, each flower label has been manually mapped to a Wikipedia[10] page using the search functionality on that site. If no corresponding page could be found, we used the label plus the word *wiki* on a regular search engine to find wiki's in other languages. Subsequently we used the *Show this page in other languages* functionality of Wikipedia to revert back the English form. DBPedia is a semantic knowledge base, based on information from Wikipedia. As almost every page on the English Wikipedia has a similar page on DBPedia which has the same content. For example the DBPedia page about the Californian Rose is http://dbpedia.org/page/Rosa_californica. Thanks to the link with DBPedia, it has been possible to reconcile synonyms, and automatically associate each flower label as belonging to the `Genus`, `Species`, or `Family` level. Each label was also classified as a `common` name or as `botanical` name.

4.5. Label quality assessment

As the prior expertise of crowd workers is unknown, we assess the quality of labels they provide. For this purpose, we sought botanical experts through our social and work environments, but also looking for candidates in our geographical surroundings. Initially we only contacted members of a garden-historical society, but unfortunately those members did not have time available. We then expanded our inquiries to the Wageningen University of Agriculture in the Netherlands and to friends, family and acquaintances who were known to be working with, or are passionate about, flowers. Our efforts resulted in three plant-related researchers from Wageningen University, a former forest ranger and a practitioner in the flower domain, who volunteered their time to annotate our dataset collection with flower labels. We trusted their self-assessment of their capabilities after showing sample images from our print dataset. We let these domain experts annotate the images and test the crowd labels with respect to the ones produced by these annotators with known domain expertise.

Two domain experts annotated all (80) prints. The other three annotated 24, 13 and 3 prints, respectively. In total, 186 flower labels were provided. The majority of all the labels provided by experts (80%) were defined at the `genus` level. The `botanical` form was used in 37% of the labels. In 57 annotation tasks (28%) at least one expert reported *unable* to name any of the flowers in the print. Notably, for 12 prints (11 `hard` and 1 `average`) no expert was able to provide any label. The remaining 68 prints received on average 2.7 flower labels ($\sigma = 1.45$, $Max = 7$) each. The most frequently used labels have been *Rose* (20% of all labels), *Lily* (17%), *Tulip* (6%), *Carnation* (5%) and *Iris* (5%).

On average, experts respectively provided 2.40 ($\sigma = 1.60$), 2.40 ($\sigma = 0.73$), and 1.64 ($\sigma = 1.47$) unique labels for `easy`, `average`, and `hard` prints. The perceived experts vocabulary featured 59 distinct flowers names. By reconciling `botanical` and `common` name of flowers, such as *Helianthus* and. *Sunflower*, the number of unique labels decreases to 52. Fifty-six percent of experts' vocabulary is expressed in `botanical` form. Looking at the distribution of unique labels, most were expressed at the `genus` specificity level (66% versus 34% at the `species` level).

4.6. Evaluation and aggregation

In this subsection we define the metrics used to assess the quality of individual annotations and define our aggregation methods for multiple annotations. Section 4.6.3 describes our novel bounding box aggregation algorithm.

4.6.1. Evaluation of individual annotations

We assess crowd workers' contribution quality by comparing their annotation results with the ground-truth created by trusted annotators. In the Artwork–centric configuration, we compare the number of flowers and the number of flower types. In the Class–centric configuration, for which we have the bounding box information, we define 3 additional metrics related to the precision of the bounding boxes.

Matching: given a ground truth bounding box *gb*, a bounding box *b* created by a worker is defined as *matching* if it overlaps with *gb*. If *b* matches multiple *gb*'s, we take the closest one (measured by the L_1 distance between the bounding box vectors) as the matched ground truth bounding box.

[10] Although Wikipedia has information on many topics, the flower domain is very well represented having most of the flower taxonomy present.

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008

Centroid distance: If a worker bounding box b matches a ground truth bounding box gb, we define its centroid distance as

$$cDist = \sqrt{\left(\frac{gb_{cX} - b_{cX}}{gb_w}\right)^2 + \left(\frac{gb_{cY} - b_{cY}}{gb_h}\right)^2}$$

where gb_w, gb_h denotes the width and height of the ground truth bounding box, b_w, b_h denotes the width and height of the matched bounding box. Intuitively, *cDist* measures the distance between the centres of the two bounding boxes; to account for differences in bounding boxes size, *cDist* is calculated by normalising the coordinate distances by the ground truth bounding box width and height.

Area overlap: If a worker bounding box b matches a ground truth bounding box gb, we define its area overlap as

$$aOverlap = \frac{gb_a \cap b_a}{gb_a \cup b_a}$$

where gb_a, b_a denotes the area of ground truth bounding box and worker bounding box. Note that for *aOverlap* we simplify the intersection/union of the *geometric* areas of two bounding box using set operators. The value of the *aOverlap* varies in the range (0, 1], in which 1 means that gb and b perfectly cover each other.

Notice that *cDist* and *aOverlap* measure different properties: a worker bounding box close to ground truth bounding boxes (i.e., with a small value of *cDist*) might not necessarily cover a large portion of area of the same ground truth bounding boxes (i.e., a large value of *aOverlap*), and vice versa.

4.6.2. Artwork–centric annotation aggregation

As common in crowdsourcing, the works of several contributors are aggregated in order to produce a single "crowd truth".

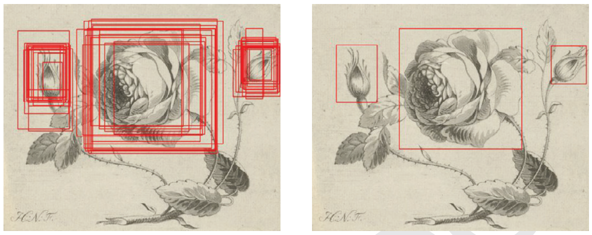
Identification. The *number of flowers* and *number of flower types* identified in a print are numerical variables. We therefore adopt as aggregation operators traditional numeric functions: (1) the `median` of the figures provided by each worker for a given print, which guarantees robustness to outliers; and (2) the `maximum` value provided by one of the annotators which, on the other hand, can be seen as a conservative measure. We decided to experiment the `maximum` function to account for the under-specification tendency of workers described in the previous section.

The ratio of erroneous number of flowers (ξ_{flower}) and number of types (ξ_{types}) reported for a print are defined as follows:

$$\xi_{flower} = \frac{|\#FL_{at} - \#FL_{gt}|}{\#FL_{gt}};\ \xi_{types} = \frac{|\#TY_{at} - \#TY_{gt}|}{\#TY_{gt}} \tag{1}$$

where $\#FL_{at}$ and $\#TY_{at}$ are, respectively the average number of flower and the average number of flower types provided by workers, and $\#FL_{gt}$ and $\#TY_{gt}$ are the number of lowers and number of flower types in ground truth.

Recognition. We aggregate the labels provided by the crowd for each artwork using three different methods: (1) the union of *all* crowd labels; (2) labels specified by *at least 2* crowd workers; and (3) labels specified by the *majority* of workers annotating a print.



(a)Individual (b)Aggregated

Fig. 6. An example of aggregating multiple individual bounding boxes.

4.6.3. Class–centric annotation aggregation

Identification. For this configuration we assess, as in the Artwork–centric, both the number of flowers and types using the `median` and `maximum`. Given the local nature of Class–centric annotations, we address the problem of bounding boxes aggregation. This task is not trivial, as it is hindered by several factors, e.g.: (1) the usage of different pointing devices, e.g. mouse or touch screens, which affects the drawing action; (2) workers might not be able to correctly identify an entity instance; and (3) the presence of malicious or poorly motivated workers, that might incorrectly (e.g. by drawing random or very big bounding boxes), or partially perform the task at hand. To cater for such problems, we propose a novel method to aggregate the Class–centric annotations produced by multiple workers, so that bounding boxes can be aggregated together to obtain a correct, high-quality identification. We propose a novel method to aggregate Class–centric annotations produced by multiple workers, each providing multiple bounding boxes to the same image. We note that a similar method, [29], has been used for aggregating bounding boxes from multiple sources. However, that method assumes each image has only one object, i.e. each worker only provides one bounding box to an image.

See Fig. 6 for an example of the application of a bounding box aggregation method.

Fig. 7 summaries the steps that compose our method for bounding box merging. The bounding boxes defined for a print are first pre-processed to remove low quality ones, i.e. bounding boxes having an area bigger than 3σ (3 standard deviations) of the mean value of all the bounding boxes specified for the same image. The threshold of 3σ is chosen to retain as many bounding boxes that could contribute to aggregation. Such an heuristic is justified by the empirical observation that some users draw *very large* bounding boxes that contains multiple flowers, in contrast with other, more committed and precise users.

Then, the method requires the identification of all groups of bounding boxes in a print. Such a step can be modelled as a clustering task, where the goal is to find clusters of bounding boxes such that all bounding boxes within the same cluster target the same visual object occurrence, and that bounding boxes of different clusters target different occurrences. We test the performance of three clustering techniques: (1) *simple geometric clustering*, where bounding boxes are defined as belonging to the same cluster if they overlap (also partially); (2) *k-means*; and (3) *Gaussian Mixture Model* (GMM) [30]. For the two unsupervised clustering methods, bounding box B_i is represented by the coordinates of its upper-left point p and its bottom-right point q, i.e., $B = (p_x, p_y, q_x, q_y)^T$.

Fig. 7. Steps of our Class–centric annotation aggregation method.

Note that for unsupervised clustering methods, we need to explicitly set the number of clusters *K*. Due to the wide range of flower occurrences in all images, *K* is set differently for each print according to annotations from workers. Different estimation strategies could be instantiated according to this principle. We compare: (1) the `maximum` number of bounding boxes drawn by one of the print annotators; (2) the `median` number among the workers; and (3) a value automatically obtained according to model selection such as Bayesian information criterion (BIC) [31].

This clustering step is then followed by the derivation, for each cluster or bounding boxes, of a single, representative bounding box. Similarly as for the setting of the *K* number of clusters, we compare the performance of several aggregation methods, which, for each of the four bounding box vertexes, select (1) the `maximum`; or (2) the `median` value among the ones of the bounding boxes in the cluster.

Recognition. To enable comparison, bounding box labels are managed as global artwork labels, and assessed using the same metrics as in the Artwork–centric configuration.

5. Results

This section discusses the outcomes of the experiments described in Section 4. For each configuration, we assess multiple annotation aggregation methods, and we analyse the resulting identification and recognition performance of crowd workers.

5.1. Artwork–centric knowledge extraction evaluation

A total of 151 workers started a qualification task, out of which 67 (44%) failed. Out of the remaining 84 workers, 40 (26%) decided to stop at the qualification task. The remaining 44 workers performed a total of 475 annotation tasks. Each worker annotated an average of 10.8 prints ($\sigma = 7.6$). Despite the high variance each print was sufficiently annotated by an average 5.9 workers ($\sigma = 1.4$).

5.1.1. Identification

For each annotation task, we analyse the number of flowers and flower types indicated by crowd workers. Table 1 describes the under-/over-specification of these values with respect to the ground-truth, broken down according to the difficulty class of the corresponding prints. Regardless of the difficulty class, workers tend to report less flowers than the ones actually contained in the print; `hard` prints, which on average contain more flowers, are also the ones for which this under-specification effect is more evident. A similar result is observable for `hard` prints for the reported number of flower types. Crowd workers also under-specified for `easy` and `average` prints but the effect is less evident, especially for the number of types. Note that this trend is highly correlated with the *number* of flower instances in the print, but less correlated with the *size* of such instances (which is smaller in `average` than `easy` prints).

5.1.2. Recognition

Workers provided a total of 461 flower name labels. Almost all labels were provided in English, except for some labels provided in Italian, Spanish and Dutch. All prints received at least one flower label from at least one worker. In 69 annotations tasks (14%) at least one crowd worker reported *unable* to name any flowers in the print. Figs. 8 and 9 depict the distribution of workers' labels in the 3 print difficulty classes. *Fantasy* labels were equally distributed. Prints where multiple workers were *unable* are distributed as follows: `easy` (1), `average` (7), and `hard` (11). As expected, `easy` prints received on average more unique labels than both `average` and `hard`. We account the higher number of unique labels for `hard` prints to the higher number of depicted flower types.

Fig. 10a depicts the distribution of label specificity for prints belonging to the three difficulty classes. `Genus` labels and *common* names are consistently the most used by crowd workers (respectively 85% and 77% of all annotations). Family names are rarer, but mainly specified in `easy` prints. Workers' vocabulary included 74 distinct flowers labels (58 after reconciliation of botanical and common version of the same flower name). On average, respectively 4.30 ($\sigma = 2.90$), 2.13 ($\sigma = 1.22$), and 4.03 ($\sigma = 2.06$) unique labels were provided for `easy`, `average`, and `hard` print. Twenty-eight percent of flower labels were defined at `species` level, 67% at `genus`, 5% at `family` level. Sixty-one percent of workers' vocabulary is expressed in *botanical* form. The labels most frequently used by crowd workers are *Rose* (33% of all labels), *Lily* (15%), *Tulip* (7%), *Sunflower* (6%) and *Carnation* (5%). Fig. 10b depicts the vocabulary size per worker. Four workers, which provided only *unable* and *fantasy* labels are not reported. On average each worker has a vocabulary size of 5.4 ($\sigma = 3.5$, $Max = 16$) distinct labels. The vocabulary size is strongly correlated ($c = 0.75$, $p \ll 0.005$) with the amount of labels a workers provided . To account for this we calculate a worker's label diversity using Shannon entropy which is also shown in 10 b (indicated by a line). Users who provide more labels, also use such labels more often.

5.1.3. Aggregation

Table 2 compares the results, in terms of error ratio, and broken down according to the difficulty class of the corresponding prints, obtained after applying the `median` and `maximum` aggregation functions. Performance is better (i.e. higher identification precision) for prints in the `easy` and `average` than in the `hard` difficulty class. On the other hand, workers perform worse on the identification of the number of flowers in the `easy` class than in the `average` class. Such a result is mostly due to 5 prints in the `easy` class, for which there is an error rate higher than 50%. Manually inspecting the annotations for these prints we observed that

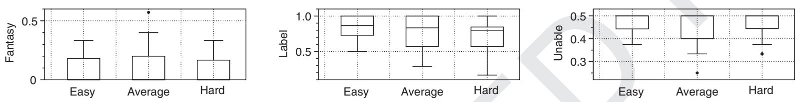
Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008

Table 1
Average difference and under-/over-specification of number of flowers and types by workers in Artwork–centric annotation.

	Number of flowers				Number of types			
	Under	Equal	Above	Avg. diff.	Under	Equal	Above	Avg. diff.
`Easy`	80%	20%	0%	−5.35 ± 7.93	40%	40%	20%	−0.94 ± 1.76
`Average`	72%	21%	7%	−0.90 ± 0.96	24%	52%	24%	−0.11 ± 1.09
`Hard`	97%	0%	3%	−9.66 ± 11.14	71%	23%	6%	−1.52 ± 1.88

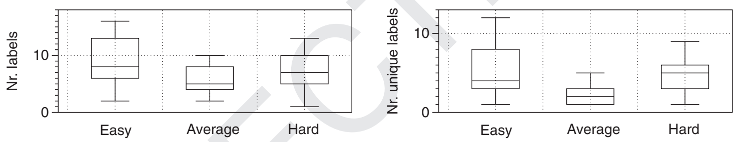
Table 2
Error rate for *number of flowers* and *types* per class for different aggregation methods in Artwork–centric annotations.

	# Flowers error ratio			# Types error ratio		
Method	`Easy`	`Average`	`Hard`	`Easy`	`Average`	`Hard`
`Median`	0.34 ± 0.30	0.14 ± 0.20	0.45 ± 0.25	0.11 ± 0.17	0.11 ± 0.23	0.30 ± 0.27
`Maximum`	0.11 ± 0.20	0.16 ± 0.25	0.41 ± 0.50	0.27 ± 0.54	1.27 ± 3.75	0.22 ± 0.25



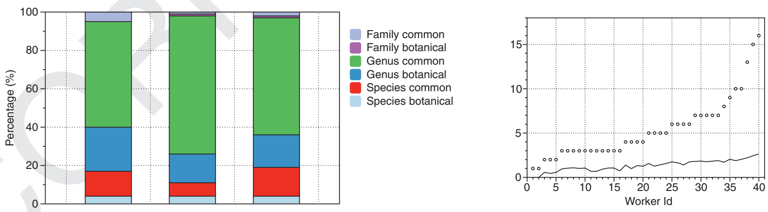
(a) Fantasy (b) Label (c) Unable

Fig. 8. Percentage of workers that specified at least 1 a) fantasy, b) label, c) unable across different print classes in Artwork–centric annotation.



(a) Number of labels (b) Number of unique labels

Fig. 9. Number of (unique) labels on prints across different classes in Artwork–centric annotation.



(a) Specificity of used labels by per print class. (b) Vocabulary size (points) and label diversity (line) per worker

Fig. 10. Analysis of labels in Artwork–centric annotations.

such errors are accountable to badly followed instructions (contrary to our flower definition, workers did not count flower buds) or to wrong interpretation of a flower composition (some workers counting each individual flower of the plant — a hyacinth — while others counted them as a single flower).

5.1.4. Interpretation of the results

Workers tended to under-specify the number of flowers and flowers types, especially for prints in the `hard` category. We account this result to the fatigue effect [32] that might occur when the number of visual classes instances increases; we interpret the propensity of workers to conservatively estimate the number as an indication of genuine effort.

Despite no requirements from our side, a surprisingly large proportion of the labels (23% in total and 61% of their vocabulary) used the botanical form for flower names. This result hints to two possible explanations: (1) the annotators *knew* the botanical name; or (2) annotators actively looked up the flower in (Web) knowledge bases to retrieve a suitable name. We interpret the result as a sign of knowledgeable and intrinsically motivated workers, which stood out from the crowd despite the task complexity and moderate reward.

No aggregation function consistently provided better performance. The `maximum` aggregation method performs better for the identification of number of flowers. This result can be intuitively explained as follows: even when workers tend to under-specify the number of object instances, a single good performer suffices for good-quality identification. On the other hand, `median` has better performance in the identification of number of flower types. Again, such a result can be motivated by the behaviour of crowd workers who,

Table 3
Average difference and under-/over-specification of number of flowers and types by workers in Class–centric annotations.

	Number of flowers				Number of types			
	Under	Equal	Above	Avg. diff.	Under	Equal	Above	Avg.diff
`Easy`	75%	15%	10%	-3.70 ± 4.45	45%	20%	35%	-1.29 ± 2.33
`Average`	72%	21%	7%	-0.75 ± 0.86	45%	31%	24%	-0.26 ± 0.69
`Hard`	87%	3%	10%	-11.53 ± 14.83	71%	6%	23%	-1.63 ± 2.14

Table 4
Quality and number of bounding boxes compared to ground truth annotations in Class–centric annotations.

	Quality of workers			Number of bounding boxes			
	Matching ratio	*cDist*	*aOverlap*	Under	Equal	Above	Avg. diff.
`Easy`	0.70 ± 0.30	0.15 ± 0.14	0.23 ± 0.19	80%	10%	10%	-3.92 ± 4.54
`Average`	0.76 ± 0.29	0.25 ± 0.30	0.49 ± 0.20	72%	28%	0%	-1.20 ± 1.31
`Hard`	0.43 ± 0.28	0.33 ± 0.45	0.59 ± 0.23	100%	0%	0%	-12.59 ± 15.13

on average, were more often correct. In such a condition, using the majority of worker annotations is a better approach, which can compensate to some extent the presence of outliers.

5.2. Class–centric knowledge extraction evaluation

Eighty-fourworkers started the qualification tasks, out of which 21 (25%) failed. Of the remaining 63 workers, 17 (27%) decided to stop at the qualification task. The remaining 46 workers created 552 annotations. All labels were provided in English. On average each worker annotated 12.0 prints ($\sigma = 17.1$). Despite the high variance each print was sufficiently annotated by, on average, 6.9 workers ($\sigma = 1.37$). A total of 3442 bounding boxes were drawn.

5.2.1. Identification

Table 3 reports the identification performance in terms of number of flowers and number of types for Class–centric annotations. Workers tend to define fewer flowers and fewer flower types than the ones actually contained in prints, especially for the `hard` ones. On the other hand, `average` prints, which contains the least number of flowers and types, are the ones where workers perform best. Table 4 shows the quality of created bounding boxes across different print classes. `Matching ratio` in this table denotes the ratio of matched bounding boxes out of the ones defined in the ground truth for the considered prints.

The better matching ratio is obtained for `average` prints, while `hard` prints have significantly worse workers' bounding box quality. Considering the reference ground-truth, this result suggests a correlation between the number of flowers in a print, as prints with less flower features better matching ratio. *cDist* and *aOverlap* are, however, measures which are more influenced by the sizes of the ground truth bounding boxes: intuitively, it is easier for workers to more accurately position boxes for large flowers, although their overlap is not necessarily better (i.e. it is more difficult for workers to draw accurate bounding boxes).

In the right part of Table 4 we report the performance, in terms of under-/over-specification of number flowers, which can be achieved by counting the number of bounding boxes specified by workers. When comparing these result with the ones in Table 3, we observe how workers often (22% of the executed tasks) draw less bounding boxes than the ones they specify in the dedicated text field of the user interface for the same annotated print. Only in a handful of tasks (0.7%) they drew more bounding boxes.

5.2.2. Recognition

Workers drew 3,442 bounding boxes, of which 1,583 (46%) contained a label that could be mapped to a DBPedia resource. 616 (18%) were annotated as fantasy, while for 1149 bounding boxes (33%) workers indicated they were *unable* to name the flower. The ratio of *unable* annotations is similarly distributed across print difficulty class — respectively 34% in `easy` prints , 31% in `average` prints, and 37% in `hard` prints. The size of bounding boxes has no effect on the ability of workers to provide a label for a flower. The same applies to the number of flowers or flower types in the print.

Fig. 11 a shows the specificity of the provided labels per print difficulty category. Both `family` names (the most generic) and `botanical` forms of `species` are rarely used. The large majority of labels, for all three difficulty classes uses the `genus` specificity in `common` name form. For prints in the `average` difficulty class less `species` labels are used than in the other classes.

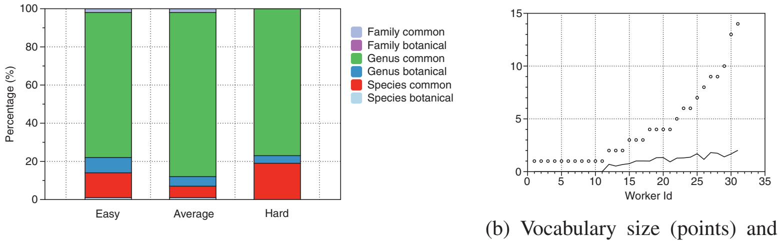
In total 1566 labels corresponding to flower names were provided. Workers featured a vocabulary of 45 distinct flowers labels (43 after reconciliation of `botanical` and `common` version of the same flower name). On average, respectively 4.05 ($\sigma = 3.78$), 2.03 ($\sigma = 1.12$), and 2.97 ($\sigma = 2.18$) unique labels were provided for `easy`, `average`, and `hard` prints.

Crowd workers specified 24% of the labels at `species` level, 69% at the `genus` level, and 7% at the `family` level of the flower taxonomy. Thirty-six percent of their vocabulary is expressed in `botanical` form. The label most frequently used by workers are *Rose* (43%), *Tulip* (14%), *Lily* (11%), *Daisy* (6%) and *Sunflower* (4%). The *botanical* form of labels is used in 6% of the labels.

5.2.3. Aggregation

Similarly to the analysis performed for the Artwork–centric configuration, in Table 5 we first report the

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008



(a) Specificity of used labels by per print class. (b) Vocabulary size (points) and label diversity (line) per worker

Fig. 11. Analysis of labels in Class–centric annotations.

Table 5
Error rate for *number of flowers* and *types* per class for different aggregation methods in Class–centric annotation.

	Flowers error ratio			Types error ratio		
Method	`Easy`	`Average`	`Hard`	`Easy`	`Average`	`Hard`
`Median`	0.25 ± 0.29	0.12 ± 0.17	0.48 ± 0.27	0.17 ± 0.28	0.13 ± 0.23	0.40 ± 0.31
`Maximum`	0.12 ± 0.22	0.18 ± 0.24	0.46 ± 0.70	0.57 ± 0.87	1.11 ± 4.04	0.72 ± 0.75

Table 6
Error rates of different methods for setting the number of clusters *K*.

	Error rate			Error sign		
Method	`Easy`	`Average`	`Hard`	Under	Equal	Above
`Max`	0.15 ± 0.28	0.11 ± 0.19	0.34 ± 0.26	32%	39%	29%
`Median`	0.23 ± 0.28	0.12 ± 0.16	0.49 ± 0.26	68%	30%	2%
`BIC`	0.96 ± 0.82	1.26 ± 0.97	0.59 ± 0.42	45%	5%	50%

performance of different aggregations methods on the number of flowers and number of flower types explicitly specified by workers. `Median` always outperforms `maximum` for the aggregation of the number of types. For the number of flowers however, `maximum` outperforms `median` in both `easy` and `hard` prints, while having worse performance in `average` prints.

Next, we compare the performance of the three *K* estimation techniques per print difficulty class for the clustering step of our novel algorithm presented in Section 4.6.3. To this end, we define the error rate

$$\xi = \frac{|\#BBs - \#FL_{gt}|}{\#FL_{gt}}$$

which measures the normalised difference between *K* determined by a configuration and the ground truth. Table 6 reports the resulting performance. Simpler methods substantially outperform `BIC`. This results suggest that, for the purpose of estimating the number of clusters, the number of bounding boxes given by the workers is a less noisy signal than the overall coordinates of all input bounding boxes. `Maximum` outperforms `median` in both the `easy` and `hard` difficulty classes, while being only slightly better for `average` prints. The result can be justified by the fact that the `average` category have the lowest number of flowers. By comparing the figures in Table 6 and Table 5, we observe how using the `maximum` value is best both for aggregating the reported number of flowers and bounding boxes; however, for `average` and `hard` prints, estimating the number of flowers in a print using bounding boxes lead to more precise results.

Table 7
Performance of different algorithm configurations.

Configuration	Matching ratio	*cDist*	*aOverlap*
`Geometric_median`	0.46 ± 0.12	0.31 ± 0.20	0.45 ± 0.16
`k-means_median`	0.82 ± 0.22	0.28 ± 0.20	0.48 ± 0.20
`GMM_median`	0.81 ± 0.22	0.28 ± 0.20	0.49 ± 0.16
`Geometric_max`	0.48 ± 0.29	0.58 ± 0.51	0.28 ± 0.18
`k-means_max`	0.83 ± 0.21	0.49 ± 0.92	0.38 ± 0.18
`GMM_max`	0.83 ± 0.21	0.48 ± 0.92	0.39 ± 0.18

Being the best aggregation method, we select the value returned by `maximum` as an input for the next steps. Table 7 compares the performance in terms of matching ratio, *cDist* and *aOverlap* of the different algorithm configurations that can be obtained by varying clustering techniques and strategies for picking the representative.

In all clustering techniques, `median` performs significantly better than `maximum` for both *cDist* and *aOverlap*, but with a slightly worse `Matching ratio`. Such a result can be explained by the fact that larger representative bounding boxes are more likely to match with ground truth bounding boxes, but will also feature bigger and less accurate areas. Among the tested clustering techniques, the ones based on unsupervised learning clearly outperform the *simple geometric clustering*, while there is no significant difference between *k*-means and GMM.

To conclude, we investigate how the performance of GMM and `k-means` vary according to the print difficulty class, and report the results Fig. 12 (red for `k-means` and blue for GMM). It can be observed how GMM is less effective for `easy` prints,

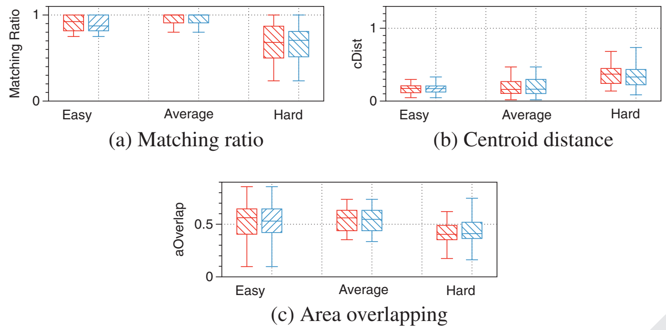


Fig. 12. *k*means_median (red) and GMM_median (blue) performance on images of different difficulty class.

while more effective for `hard` prints. Technically, the result could be explained by the nature of GMM, a more flexible version of *k*-`means` that models the variance and covariance of bounding box coordinate vector. The coordinates of ending point in drawing a bounding box for small flower is more influenced by the starting point than for large flower, which could make GMM perform better than the simpler *k*-`means` clustering on `hard` prints; on the other hand, by being less robust than *k*-`means` on this aspect, GMM performs worse in `easy` prints, where the effect of little pointing errors are more evident. We note that for their performance, however, none of the differences in the three print classes are statistically significant.

5.2.4. Interpretation of the results

In this configuration, workers were asked to both count the number of flowers in the print, and to draw bounding boxes around them (see Section 4); a counter in the user interface reported the number of currently drawn bounding boxes. Despite the presence of a self-defined anchor, workers often drew a number of bounding boxes less than the number of identified flowers. The result can be interpreted in two ways. On the one hand, as there was a direct relation between the number and size of flowers in prints, workers could have simply ignored very small instances; on the other hand, due to fatigue, workers simply stop annotating even if they knew that more flowers existed in the print. An analysis performed over several annotation tasks suggests that both explanations are correct to some extent. The result hints to the importance of intrinsically motivated workers for Class–centric artwork annotation, as a high number of instances to annotate might be discouraging.

Annotation specificity hints at a similar conclusion: 36% of the annotations, but only 6% of their vocabulary, are expressed in botanical form. This suggests less knowledgeable workers, but also workers less available for online research (e.g. in knowledge bases), possibly due to the additional effort needed to draw bounding boxes, which diminishes their motivation to extend their domain knowledge.

Our method for bounding boxes aggregation shows promising results. The significantly improved aggregation results (up to 45% improvement in matching ratio, 20% in bounding box distance, and 30% in area overlap) achieved by the GMM and *k*-`means` algorithms demonstrates the need for sophisticated aggregation techniques. Due to workers generally under-specifying the number of flowers and flowers types in a print, the `maximum` aggregation function is better suited for the estimation of the number of flower instances which, in turns, provide better identification (matching ratio) performance. In contrast, the location of bounding boxes (centre and area) can be better achieved using a consensus-based function such as `median`.

These results demonstrate how, in this context, and contrary to traditional image annotation techniques, better outcomes can be achieved by selectively using methods not based on crowd consensus. This is supported by recent work on crowd disagreement [33].

6. Discussion

This section elaborates on the results reported in Section 5 with respect to the research questions defined in the introduction.

6.1. Research question 1

In this section we provide answers to **RQ1**: *Can non-professional annotators from crowdsourcing platforms provide high quality artwork annotations?*

6.1.1. Identification

Section 5.1 and Section 5.2 provide quantitative evidence of the effectiveness of crowd workers in *identifying* and *locating* flower instances. Despite the moderate reward and the demanding nature of the annotation tasks, our experiments attracted a considerable number of skilled and motivated workers; they often matched the *identification* performance of our trusted annotators, especially on prints of `easy` and `average` difficulty (matching rates of 89% and 84%, respectively); `hard` prints lead to worse, but still satisfactory figures (a matching rate of 59%). Even with the Class–centric configuration, which was more labour intensive, workers achieved good *location* accuracy (bounding box matching rates of 70%, 76% and 43% for easy, average and hard prints, respectively). These results are indicative of the *willingness* and *ability* of crowds to identify visual classes in artworks. Indeed, the identification of elements on images is a task that is familiar to crowd workers and requires little to no domain knowledge. On the other hand, artworks often present additional complexities with respect to photographic images (e.g. abstract, symbolic, or allegoric interpretations): the results of our experiments are in line with the outcomes

Table 8
Comparison of number of common and new labels, split by print difficulty class and aggregation methods, in Artwork–centric annotations.

	Matching label			New label		
Method	`Easy`	`Average`	`Hard`	`Easy`	`Average`	`Hard`
`All`	1.30 ± 0.80	0.79 ± 0.50	1.00 ± 0.79	3.00 ± 2.88	1.25 ± 2.35	3.45 ± 2.35
`Lab.freq.` ≥ `2`	0.95 ± 0.60	0.68 ± 0.55	0.65 ± 0.67	0.65 ± 1.09	0.29 ± 0.46	0.65 ± 0.81
`Majority`	0.70 ± 0.66	0.64 ± 0.56	0.40 ± 0.50	0.15 ± 0.37	0.18 ± 0.39	0.25 ± 0.55

Table 9
Label comparison per print difficulty class for different aggregation methods in visual class annotation in Class–centric annotation.

	Matching label			New label		
Method	`Easy`	`Average`	`Hard`	`Easy`	`Average`	`Hard`
`All`	1.21 ± 1.24	0.79 ± 0.57	1.05 ± 1.00	2.84 ± 2.98	1.18 ± 1.06	2.35 ± 2.39
`Label.freq.` ≥ `2`	1.11 ± 1.24	0.64 ± 0.62	0.80 ± 0.89	1.79 ± 2.04	0.89 ± 0.88	1.70 ± 1.42
`Majority`	0.95 ± 1.18	0.64 ± 0.62	0.80 ± 0.89	1.53 ± 1.58	0.89 ± 0.83	1.45 ± 1.19

of other recent studies [25,26], and provide additional evidences on the suitability of crowdsourcing as an accurate tool for cultural heritage content annotation.

6.1.2. Recognition

In terms of recognition performance, workers consistently provide a high number of labels and show a rich vocabulary. In Table 8 and Table 9 we compare, for each experimental configuration, the number of *unique* labels which are provided by the domain experts and the crowd (Matching label) and the number of *unique* labels provided by the crowd but not by the experts (New label)

The vocabulary size of crowd workers is comparable (58 and 43 compared to 52 for the Artwork–centric and Class–centric, respectively) to the one of experts, and, for both configurations, we observed how workers often used labels with lower specificity (genus and family) in the flower classification taxonomy. This result suggests familiarity with the domain-specific vocabulary. On the other hand, it can also be interpreted as an indicator of the potential information need of crowd-workers, who prefer using laymen terminology to describe botanical entities: workers were allowed to look up flower names online, but they deliberately choose to specify `common` names at lower level of specificity.

The previous interpretation is supported by the following observation. In both configurations (Artwork–centric and Class–centric), and even using the most conservative label aggregation policy (i.e. using *all* crowd labels), experts and crowd workers respectively share a limited vocabulary: 34%, 28%, 34% (Artwork–centric) and 43%, 44%, 52% (Class–centric) respectively for `easy`, `average`, and `hard` prints. The size of the shared vocabulary only slightly decreases (on average) with stricter aggregation conditions. On the other hand, the decrease is more evident in terms of new labels (right part of Table 8 and Table 9): the number of new labels introduced by crowd workers is relatively consistent across aggregation methods and print difficulty.

Both crowd workers and experts showed a similar tendency to have *low agreement* on their labels. This is an interesting phenomena, also observed in other knowledge-intensive content annotation use cases (e.g. medical [33]). The result suggests the need for more articulated annotations campaigns, possibly organised in workflows [34] that interleave automatic and human operations. While this is subject of future work, we can envision the following annotation flow: crowd labels are used to instrument a Web retrieval step, where images of flowers associated with the provided label are collected. Such images can then be used in another crowdsourcing task as a comparison term, to visually verify the similarity of the labelled flower instance with the real-world examples.

6.2. Research question 2

In this section we provide answers to **RQ2**: *To what extent can the extraction and aggregation steps of a crowdsourced knowledge creation process influence the identification and recognition aspects of visual artwork annotation?*

6.2.1. Extraction

The experimental evaluation shows that the adoption of different *knowledge extraction* interfaces has a relevant effect on the identification and recognition performance. Such an effect is not uniformly distributed across print annotation difficulties, but it allows for interesting considerations.

By comparing the figures of Table 1 (Artwork–centric) and Table 3 (Class–centric) we observe how the presence of the bounding box functionality, which should push workers to a more precise identification of flower instances, does not result in a significant difference (less than 10%) in the distribution of under- and over-specification of the number of flowers. However, we observe that using drawn bounding-boxes as a way to count the number of flowers do lead to significantly better results (see Tables 2, 5 and 6). This is especially observable with more difficult prints; 21% and 24% decrease of the error rate for `average` and `hard` prints using the optimal aggregation method, respectively. On the other hand, the identification performance concerning the *number of flower types* shows a different trend, as workers in Artwork–centric annotation achieve a lower error than in Class–centric annotation. These results suggest that the identification aspect could benefit from an annotation interface that benefits from both global (Artwork–centric) and local

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008

(Class–centric) knowledge extraction interfaces, to be used according to the difficulty of the print to annotate.

Different observations can be made for the *recognition* aspect. The presence of a bounding box forces workers into providing a label for each identified visual class instance. Despite the availability of a “Don’t Know” option for labels, workers often provided the same annotation even for flower having clear visual differences. Moreover, by comparing Fig. 10 with Fig. 11, we observe that the labels provided in the Class–centric configuration are more frequent of a more generic nature (genus and family) compared to the Artwork–centric configuration. A higher proportion of labels is also specified using the *common* name. The vocabulary size of workers in the Class–centric configuration is also smaller. These results suggest that, while a Class–centric configuration can help obtaining higher-granularity annotations, the overhead required for drawing bounding boxes might penalise the recognition capabilities of workers. A possible explanation of this result can be attributed to the well-known fatigue effect [32], that often occurs with repetitive tasks. Due to fatigue, workers could be led to provide wrong labels, thus introducing noise. On the other hand, the cost of providing a “Don’t Know” annotation was of a single click, considerably lower than typing a work, or copy/paste it. We weren’t able to find a comprehensive explanation to justify the different *recognition* performance with the Class–centric configuration. We rely on future work to obtain a better understanding of the impact of the cognitive load of an annotation interface over the quality of the retrieved annotation.

6.2.2. Aggregation

The experiments show that the *aggregation* step can impact the quality the visual class *identification* results. The `maximum` and `median` functions find different optimal applications. The former allowed a better estimation of the *number of flowers*, whole the latter was more suited to estimate the *number of flower types* contained in the print.

Once more, we can explain these results as a consequence of a fatigue effect [32] that emerged with workers. This was most apparent for images with many small instances in the Class–centric configuration. Some workers drew fewer bounding boxes than the number of flowers they had counted and reported. Countermeasures to the fatigue effect are also studied in the field of crowdsourcing, for example in [35], where they studied the effect of inserting micro-breaks during tasks. Our results show that the adoption of a different aggregation function, e.g. `maximum`, can provide satisfactory results without the need for additional task execution time and, consequently, cost.

Altogether, the results discussed in this sub-section clearly indicate how artworks annotation demands from different *aggregation* techniques with respect to photographic image annotation: consensus-based knowledge aggregation techniques [4,16,17] need to be supported by other methods, to counter some of the additional visual complexity of artworks.

7. Conclusion

In this paper we report the results of an evaluation, conducted in collaboration with the Rijksmuseum Amsterdam, and aimed at studying how different knowledge extraction and aggregation configurations affect the *identification* and *recognition* aspects of artwork annotation. We instrumented two *knowledge extraction* configurations: an *Artwork–centric* design, where textual annotations about visual objects are specified for the whole artwork; and a *Class–centric* design, where occurrences of visual objects are identified using bounding boxes with distinct textual annotations. To support the *Class–centric* design, we proposed a novel bounding-box aggregation algorithm. Then, we experimented with different annotation *aggregation* methods, and tested their impact on identification and recognition performance.

We engaged with 235 workers from a crowdsourcing platform, and asked them to annotate 80 Rijksmuseum prints of varying annotation difficulty; *easy*, where both identification and recognition of flower instances is easy; *average*, where identification is easy and recognition is difficult; and *hard*, where both identification and recognition is difficult.

Both *knowledge extraction* configurations (Artwork–centric and Class–centric) resulted in satisfactory identification performance (Sections 5.1.1, 5.2.1 and 6.1.1). For tasks of easy and average difficulty crowd workers can achieve an identification performance comparable to trusted annotators. However, as print difficulty increases the performance of crowd workers lowers considerably.

In terms of recognition performance (Sections 5.1.2, 5.1.2, and 6.1.2), we observed that the crowd provided a rich vocabulary with little overlap with respect to domain experts, regardless of the knowledge extraction configuration. In the Class–centric configuration more workers provide a single label than in the Artwork–centric configuration. These results suggest that, while a Class–centric configuration can help obtaining higher-granularity annotations, the work overhead required for drawing bounding boxes might penalise the recognition capabilities of workers.

Crowd workers consistently provide labels at varying level of specificity

(species, genus, family) and show their familiarity with domain specific (i.e. botanical) names. The crowd provides more distinct labels per image than domain experts. This suggests an opportunity for the creation of annotation sets that are complementary to the ones of experts, and that can accommodate a broader variety of information needs. On the other hand, the general low agreement calls for further studies, in order to better assess the quality of each annotation.

Our experiments with different annotation aggregation functions (Sections 5.1.3, 5.2.3 and 6.2.2) show performance diversities. In terms of identification performance, the `maximum` function outperforms the `median`, to counter the workers’ tendency to under-specify the number of identified instances in complex artworks. On the other hand, aggregating the location of bounding boxes (centre and area) can be better achieved using a consensus-based function such as `median`.

These results shows how: (1) the adoption of different *knowledge extraction* configurations and *aggregation* methods influences both the identification and recognition performance; (2) artworks annotation demands for different *aggregation* techniques than the ones used for photographic image annotation (e.g. image-centric annotations and majority voting).

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008

While promising, these results were obtained by studying a single knowledge domain. Further investigations are needed in order to assess the impact of the crowdsourced knowledge creation process in area of knowledge that are less common in the general population (e.g. annotation of birds, castles, etc.). As part of the future work we also plan to test the impact of other steps of the Crowd Knowledge Creation process, e.g. the discovery of expert workers from the crowd to dynamically assign annotation tasks to the most suited performer, also proposed as challenge in "The Future of Crowd Work" [34] . Other future work includes the assessment of our novel algorithm for aggregation bounding boxes in other contexts and on other datasets.

Acknowledgements

This publication was supported by the Dutch national program COMMIT. We like to thank the members of the SEALINCMedia worktable and in particular the Rijksmuseum Amsterdam for their support.

References

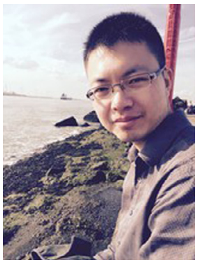
[1] G. Carneiro, Artistic image analysis using graph-based learning approaches, IEEE Trans. Image Process. 22 (8) (2013) 3168–3178. http://dx.doi.org/10.1109/TIP.2013.2260167.
[2] Q. Le, Building high-level features using large scale unsupervised learning, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 8595–8598. http://dx.doi.org/10.1109/ICASSP.2013.6639343.
[3] Q. Wang, K. Boyer, Feature learning by multidimensional scaling and its applications in object recognition, in: Conference on Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI, 2013, pp. 8–15. http://dx.doi.org/10.1109/SIBGRAPI.2013.11.
[4] S. Nowak, 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, in: MIR '10, ACM, New York, NY, USA, 2010, pp. 557–566. http://doi.acm.org/10.1145/1743384.1743478.
[5] T. Yan, V. Kumar, D. Ganesan, Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones, in: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, in: MobiSys '10, ACM, New York, NY, USA, 2010, pp. 77–90. http://doi.acm.org/10.1145/1814433.1814443.
[6] N. Sawant, J. Li, J.Z. Wang, Automatic image semantic interpretation using social action and tagging data, Multimedia Tools Appl. 51 (1) (2011) 213–246. http://dx.doi.org/10.1007/s11042-010-0650-8.
[7] S. Park, G. Mohammadi, R. Artstein, L.-P. Morency, Crowdsourcing micro-level multimedia annotations: the challenges of evaluation and interface, in: Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia, in: CrowdMM '12, ACM, New York, NY, USA, 2012, pp. 29–34. http://doi.acm.org/10.1145/2390803.2390816.
[8] J. Oosterman, et al., Crowd versus experts: Nichesourcing for knowledge intensive tasks in cultural heritage, in: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web, 2014, pp. 567–568. http://dx.doi.org/10.1145/2567948.2576960.
[9] E.L. Spratt, A. Elgammal, Computational beauty: aesthetic judgment at the intersection of art and science, CoRR abs/1410.2488 (2014). http://arxiv.org/abs/1410.2488.
[10] J. Whitehill, T.-f. Wu, J. Bergsma, J.R. Movellan, P.L. Ruvolo, Whose vote should count more: optimal integration of labels from labelers of unknown expertise, in: NIPS'09, 2009, pp. 2035–2043.
[11] P. Welinder, S. Branson, P. Perona, S.J. Belongie, The multidimensional wisdom of crowds, in: Advances in neural information processing systems, 2010, pp. 2424–2432.
[12] H. Su, J. Deng, L. Fei-Fei, Crowdsourcing annotations for visual object detection, in: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[13] J. Trant, B. Wyman, Steve, Investigating social tagging and folksonomy in art museums with steve.museum, in: Proceedings of the WWW'06 Collaborative Web Tagging Workshop, 2006.
[14] A. Ellis, D. Gluckman, A. Cooper, A. Greg, Your paintings: a nation's oil paintings go online, tagged by the public, Museum and the Web (2012).
[15] M.C. Traub, J. van Ossenbruggen, J. He, L. Hardman, Measuring the effectiveness of gamesourcing expert oil painting annotations, in: M. de Rijke, T. Kenter, A. de Vries, C. Zhai, F. de Jong, K. Radinsky, K. Hofmann (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, 8416, Springer International Publishing, 2014, pp. 112–123. http://dx.doi.org/10.1007/978-3-319-06028-6_10.
[16] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in: CHI '04, ACM, New York, NY, USA, 2004, pp. 319–326. http://doi.acm.org/10.1145/985692.985733.
[17] R. Gligorov, L.B. Baltussen, J. van Ossenbruggen, L. Aroyo, M. Brinkerink, J. Oomen, A. van Ees, Towards integration of end-user tags with professional annotations, in: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, 2010.
[18] C. Wieser, F. Bry, A. Bérard, R. Lagrange, Artigo: building an artwork search engine with games and higher-order latent semantic analysis, in: First AAAI Conference on Human Computation and Crowdsourcing, 2013.
[19] A. Kittur, E.H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in: CHI '08, ACM, New York, NY, USA, 2008, pp. 453–456. http://doi.acm.org/10.1145/1357054.1357127.
[20] R. Snow, B. O'Connor, D. Jurafsky, A.Y. Ng, Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, in: EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 254–263. http://dl.acm.org/citation.cfm?id=1613715.1613751.
[21] A. Kittur, Crowdsourcing, collaboration and creativity, ACM Crossroads 17 (2) (2010) 22–26.
[22] V. Boer, M. Hildebrand, L. Aroyo, P. Leenheer, C. Dijkshoorn, B. Tesfa, G. Schreiber, Nichesourcing: harnessing the power of crowds of experts, in: Knowledge Engineering and Knowledge Management, in: Lecture Notes in Computer Science, 7603, Springer Berlin Heidelberg, 2012, pp. 16–20. http://doi.acm.org/10.1007/978-3-642-33876-2_3.
[23] K. Heimerl, B. Gawalt, K. Chen, T. Parikh, B. Hartmann, Communitysourcing: engaging local crowds to perform expert work via physical kiosks, in: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, ACM, 2012, pp. 1539–1548.
[24] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, G. Vesci, Choosing the right crowd: Expert finding in social networks, in: Proceedings of the 16th International Conference on Extending Database Technology, in: EDBT '13, ACM, New York, NY, USA, 2013, pp. 637–648. http://doi.acm.org/10.1145/2452376.2452451.
[25] N.F. Noy, J. Mortensen, M.A. Musen, P.R. Alexander, Mechanical turk as an ontology engineer?: Using microtasks as a component of an ontology-engineering workflow, in: Proceedings of the 5th Annual ACM Web Science Conference, in: WebSci '13, ACM, New York, NY, USA, 2013, pp. 262–271. http://doi.acm.org/10.1145/2464464.2464482.
[26] L.B. Chilton, G. Little, D. Edge, D.S. Weld, J.A. Landay, Cascade: crowdsourcing taxonomy creation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in: CHI '13, ACM, New York, NY, USA, 2013, pp. 1999–2008. http://doi.acm.org/10.1145/2470654.2466265.
[27] C. Dijkshoorn, J. Oosterman, L. Aroyo, G. Houben, Personalization in crowd-driven annotation for cultural heritage collections, in: Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization, Montreal, Canada, July 16-20, 2012, 2012. http://ceur-ws.org/Vol-872/patch2012_paper_3.pdf.
[28] J. Oosterman, A. Nottamkandath, C. Dijkshoorn, A. Bozzon, G. Houben, L. Aroyo, Crowdsourcing knowledge-intensive tasks in cultural heritage, in: ACM Web Science Conference, 2014, pp. 267–268. http://doi.acm.org/10.1145/2615569.2615644.
[29] P. Welinder, P. Perona, Online crowdsourcing: rating annotators and obtaining cost-effective labels, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 25–32. http://doi.acm.org/10.1109/CVPRW.2010.5543189.
[30] C.M. Bishop, et al., Pattern recognition and machine learning, 1, springer New York, 2006.
[31] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464. http://dx.doi.org/10.1214/aos/1176344136.
[32] R.A. Hennfng, S. Sauter, G. Salvendy, E.F.J. Krieg, Microbreak length, performance, and stress in a data entry task, Ergonomics 32 (7) (1989) 855–864 pMID: 2806221. http://dx.doi.org/10.1080/00140138908966848.

[33] O. Inel, K. Khamkham, T. Cristea, A. Dumitrache, A. Rutjes, J. van der Ploeg, L. Romaszko, L. Aroyo, R. Sips, Crowdtruth: machine-human computation framework for harnessing disagreement in gathering annotated data, in: The Semantic Web - ISWC 2014-13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II, 2014, pp. 486–504. http://dx.doi.org/10.1007/978-3-319-11915-1_31.

[34] A. Kittur, J.V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, J. Horton, The future of crowd work, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, in: CSCW '13, ACM, New York, NY, USA, 2013, pp. 1301–1318. http://dx.doi.org/10.1145/2441776.2441923.

[35] J.M. Rzeszotarski, E.H. Chi, P. Paritosh, P. Dai, Inserting micro-breaks into crowdsourcing workflows, in: Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts, An Adjunct to the Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, November 7-9, 2013, Palm Springs, CA, USA, 2013. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7544.

Born in Amsterdam, the Netherlands, **Jasper Oosterman** studied Computer Science at the Delft University of Technology. His MSc thesis, titled 'Finding Experts in a Corporate Environment', concluded his MSc in 2011. Following his master, Jasper started a PhD student position supervised by Alessandro Bozzon at the Web Information Systems research group led by Geert-Jan Houben. The topic of his thesis is the maximization of Crowd Generated Knowledge and he plans to finish in 2015.



Jie Yang is a PhD student at the Web Information Systems Group of the Faculty of Engineering, Mathematics and Computer Science (EEMCS/EWI), Delft University of Technology (TU Delft). Jie's research is focused on crowdsourcing for knowledge intensive tasks.

Dr. Alessandro Bozzon is an Assistant Professor at the Delft University of Technology, and had completed his PhD from Politecnico di Milano, Italy in 2009. His current research interests are in the fields of Web data management; human computation, crowdsourcing, and games with a purpose; and information retrieval; with applications to crowdsourced knowledge creation, urban computing, and smart enterprise workforce.

Lora Aroyo is associate professor Intelligent Information Systems Web and Media Department of Computer Science, at the Free University Amsterdam, in the Netherlands. Her specialities are all explicitly 'user centred' crowdsourcing, user and context modelling, recommender systems, user centred design for semantic systems on e-learning. In the recent past Aroyo was scientific coordinator of the EU Integrated Project: NoTube: integration of Web and TV data with the help of semantics. Aroyo also coordinated the NWO project: CHIP: Cultural Heritage Information Personalization. This project was awarded a Semantic Web Challenge prize.

Geert-Jan Houben is full professor of Web Information Systems at the Software & Computer Technology department at TU Delft. His main research interests are in Web Engineering, Web Science, and User Modeling, Adaptation and Personalization. He is managing editor of the Journal of Web Engineering, editorial board member for Journal of Web Science, International Journal of Web Science, and ACM Transactions on the Web. He is scientific director of Delft Data Science, TU Delft's coordinating initiative in Data Science, leading TU Delft's research program on Open & Online Education, and principal investigator in AMS, Amsterdam Institute for Advanced Metropolitan Solutions.

Please cite this article as: J. Oosterman et al., On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.008