Participatory Art History: Exploring the Potential of Crowdsourcing for Interpretive Task

Xiaoyu Zeng
Department of Art and Art History
University of Texas at Austin
Austin, TX
edith.xiaoyu.zeng@utexas.edu

Abstract

This research experiments with art historical and crowdsourcing methodologies in exploring whether crowd workers, as opposed to trained experts, can be recruited as active participants and contributors in interpreting artworks. With collected data from micro-task crowdsourcing on Amazon Mechanical Turk, this paper compares crowdsourced analyses with art historical writings to reveal certain commonalities in human perceptions and interpretations of visual art. It also evaluates the experience of using online commercial crowdsourcing marketplace to collect these non-expert analyses, and argues for the crowd's potential to accomplish relatively subjective and open-ended interpretive tasks.

Introduction

The long tradition of connoisseurship in art history emphasizes specialized intellectual training, aesthetic sensibility, and the development of an empirical intuition to analyze artworks based on available stylistic and technical evidence. Ever since the mid-twentieth century, connoisseurship has been constantly challenged by contemporary critics for being subjective and elitist. New approaches, most notably poststructuralism and reception theory, have empowered the receptor as an active co-creator of meaning (Sontag 1966). The artist of a work no longer has the authority over its connotation, and viewers are encouraged to construct their own interpretations.

The democratization of information in the past decades further established the general public as an anonymous "crowd." Web communications allowed book readers or movie audience to share their thoughts that shape the collective experience with a work. In spite of art museums' increased emphasis on visitor participation and the academia's advocacy for diverse perspectives, the collective crowd is still largely perceived as a passive body of recipient and educatee (Simon 2010).

Conflicting philosophies of traditional high culture connoisseurship and contemporary egalitarian information technologies open up new possibilities for non-experts to participate in unlocking and co-creating meanings. In this paper, "expertise" indicates specialized training with a set of predesignated standards and classifications, as similar to supervised machine learning. An "expert" referred hereafter is not necessarily a human agent, but any operator trained in such fashion.

Current methods to achieve more objective interpretations can be roughly classified into two major trends: one is to recruit more human judgements to correct homogeneity-generated biases, and the other is to utilize non-human computational methods to make new observations. Research on historical artifacts have utilized both human and machine computations while still drawing heavily on expert opinions. Existing practices can be evaluated with three main aspects in terms of the task's agency and goal. The analytical agent can be 1) human or non-human of an 2) expert or non-expert nature, and the task can be 3) interpretive or non-interpretive in terms of its goal.

One common approach is to utilize machine computation in place of human connoisseurs to process and analyze artworks. In many studies of this type, the training data are labeled by human experts and classified into a set of predefined rubrics. The trained machine computational agent thereby remains expertise-based. In past decades, traditional fine art has been studied with scientific methods in areas such as conservation, information retrieval, computer vision, and machine learning. In a paper published in 2008, Stork provided an overview of the computer analysis of paintings and drawings (2008). He argued that computer methods had many advantages in their scalability, level of abstraction, and objectivity when compared to human judgments. Lev Manovish's recent proposal to use predictive analytics and deep learning to better understand contemporary art further elaborated a similar belief that automating methods will complement human connoisseurs' inevitably subjective understanding of cultural artifacts (2015). Automated algorithms developed to analyze digitized images ranged from tasks with objective answers, such as image identification and annotation, to applying these objective standards to make more complex and subjective aesthetic judgments, such as stylistic analysis (Hughes, Graham, and Rockmore) and image comparison based on different extracted features (Polatkan et al. 2009).

While many of these computer vision studies of artworks used a limited amount of expert-labeled images as their training data, crowdsourcing has become another popular method for data aggregation and labeling in other types of image analysis tasks. Sorokin and Forsyth have pointed out crowdsourcings potential for computer vision image annotation tasks in terms of its quality, speed, and cost-efficiency (2008). Davis et al.'s paper argued that human micro-task crowdsourcing could be a co-processing unit to accomplish tasks unsuitable for machine computation due to complexity or expense (2010). Researchers have utilized online crowdsourcing to collect human-generated data for various interests, such as behavioral researches (Mason and Suri 2012), image annotations (Vondrick, Ramanan, and Patterson 2010), and surveys (Behrend et al. 2011).

This use of crowdsourcing recruits non-expert human agents to perform non-interpretive tasks. Existing crowdsourcing tasks processing artworks have mostly delegated close-ended, non-creative, and segmented tasks to the crowd. These projects' task designs usually follow the assumption that, since crowdsourced human labor complements machine computation, the collective human co-processor does similar tasks as those accomplished by machines: handling questions with objective standards and easily-verifiable answers. Annotation and transcription tasks, as seen in Ancient Lives project on Zooniverse or Tate Gallery's AnnoTate project,[1] are the most common types of crowdsourcing projects initiated by museums, libraries, and universities to transcribe scans of primary-source materials into labeled digital contents for archival inventory and preservation. Crowdsourcing such tasks online helped institutions to manage their collections affordably, while providing volunteers an opportunity to learn and contribute to topics they are passionate about. However, solely imitating machine computation in crowdsourcing leaves out the potential of the human brain's many unique abilities, such as perception of emotions, associative thinking, and aesthetic sensibilities.

Utilizing the human aspect of crowdsourcing, multiple studies have experimented with crowdsourcing in carrying out more complicated and subjective tasks. Extending the practice from querying close-ended questions to stylistic analysis, Kovashka and Lease's paper combined human inputs with machine computation to detect stylistic similarity of a group of nineteenth-century paintings (2010). A similar experiment in natural language processing used crowdsourced responses to analyze narrative similarity in Dutch folklore (Nguyen, Trieschnigg, and Theune 2014).

Natural language processing research has utilized online crowdsourcing for sentiment analysis (Nakov et al. 2013), yet few similar works have been done on the human perception of images. Admittedly, compared to written sentences, visual compositions contain fewer objective clues to establish a ground truth that holds true to most viewers. However, such a ground truth is perhaps no more subjective than determining the meaning of verbal language. The lack of similar studies of visual expressions poses a series of question. Can we predict the emotion of a given image in a way similar to detecting irony in a written paragraph? Compared to ordinary images, artworks present more deliberately composed human expressions. Do diverse interpretations of the same artwork share some commonalities, even forming a consensus or schools of thought, to reveal certain universal experiences with the visual syntax? Recruiting the human non-experts, online crowdsourcing provides an cost-effecient and non-traditional way to collect data for this interpretive task.

Research Goals

This study experiments with three main areas: the crowd's participation in the art historical tradition of visual analysis and interpretation, a broader application of online crowdsourcing for interpretive tasks, and data collection for future research in related areas.

Participatory Art History

The first research goal involves the aforementioned hesitation of art history to incorporate non-experts as active participants and co-creators of meaning. Through collecting responses from people who are not necessarily interested in fine art, this research explores whether the untrained viewers would provide interpretations similar to scholarly opinions, and whether the same image evokes similar emotions among these two groups and individual viewers. Shared interpretations of a work's sentiment and expression may reveal certain universal experiences evoked by certain artistic expressions, and such universality can become subjects of further study to understand human perception and cognition of art as artificially and purposefully composed signs.

Crowdsourcing

This research also experiments with a broader application of online crowdsourcing practices. The goal is to find out whether this form of paid crowd work can be used for interpretive tasks with relatively subjective and unpredictable answers. Due to the prevalent micro-task design, which divides a large project into simple and repetitive tasks, online crowdsourcing platforms have been used mostly for close-ended questions with objective answers. This research's task design and collected worker feedbacks provide insights for future crowdsourcing tasks of similar nature.

Visual Analysis Data

Data collected in this research could be useful for text-image relation analysis in museum studies, art education, and computer vision. As the middle ground between fully automated machine recognition and the empirically intuitive expertise of the human connoisseur, crowd-generated interpretations could potentially create a more diverse set of syntax and connection for better understanding and modeling the relation between texts and images.

Hypotheses and Risk

Based on the research goals, the following risks and hypotheses that will be tested with collected results:

- The project may not attract a large and diverse pool of participant to collect meaningful interpretations of artworks;
- The participating crowd may not provide high-quality interpretations;

[1] See http://www.ancientlives.org/ and https://anno.tate.org.uk/.

- Crowd-generated analysis of artworks share commonalities with expert opinions;
- Based on last hypothesis, these two groups' analyses would have more overlap for figurative images than abstract ones;
- Online crowdsourcing can be utilized to accomplish relatively subjective and open-ended interpretive tasks.

Research Design

Platform

This project used Amazon Mechanical Turk (AMT), one of the most popular and well-studied online commercial crowdsourcing marketplace, to collect the crowd workers' non-expert interpretations of artworks.

AMT has many advantages when compared to other platforms. It allows individuals to register with their existing Amazon accounts and post or work on tasks without significant time or financial commitment. Existing user demographic studies have also revealed a worker pool that contains mostly English-speaking U.S. and Indian populations that constituted a pool "arguably closer to the U.S. population as a whole than subjects recruited from traditional university subject pools" (Paolacci, Chandler, and Ipeirotis 2010). Such representability and diversity were more suitable for this study's purpose than other platforms that provided high-quality but specialized labor. Although tasks in this project required a certain amount of crowdwork experience to ensure work quality, the qualification requirements were lowered than most other AMT tasks to include as many participant as possible. All assignments asked for workers who have completed 50 or more HITs and have approval rates of 95% or higher.

In terms of usability, AMT has a relatively simple interface for requesters to create and manage tasks in groups. It also provides developer documentation for more sophisticated task designs. The ready-made templates reflect some of the most frequently-performed task types on the platform: image classification, web-search-based data collection, image content moderation, natural language sentiment analysis, survey, image/audio/video annotaiton and transcription, or writing website descriptions.[2] Although this project's tasks did not fall strictly into any of these existing categories, the required visual analysis and writing skills were probably familiar to most AMT crowd workers.

Human Intelligence Task and Evaluation Criteria

To publish crowdsourcing tasks online, a requester needs to first create a "project" specifying the task's descriptions, financial reward, worker qualification, as well as contents and instructions. Each time the requester publish the project, AMT creates a "batch" that has multiple Human Intelligence Tasks (HITs) with an expected number of total input and effective time period. AMT charges the requester a 20% fee on top of rewards and bonuses paid to workers.[3] The total projected cost has to be prepaid to the requester's account to successfully publish a batch. Workers get paid after the requester's approval of their submitted answers, and this pending period can range from 8 hours to multiple days depending on each individual task setup.

Eight projects were created for this research. Seven of them contained three images; one contained six in pairs for comparative analysis. For each image, the participant is asked to first identify the main subject, and then write down a brief paragraph interpreting the image's expression and sentiment. Table 1 shows a sample HIT with task instructions, examples, and an image with two questions.

The first identification question functioned as a "buffer question" by imitating annotation tasks commonly found on AMT. This should be easy to complete before the participant continue to the interpretation task. It was also a verifiable question in an otherwise gold-standard-free task. One could identify any main object in the image and the answer still requied manual verification. Nevertheless, this step filtered unsatisfactory and bot-generated inputs by rejecting responses labelling nonexistent objects. It saved the time spent reading the longer interpretive answer submitted by spam workers. More importantly, this annotation question encouraged the participant to carefully look at the image before rushing into the analytical task. Allowing people to identify any main object in the image, instead of a specified item, helped them to work quickly through the first question and spend the most time on the interpretive task.

The second question asked the person to analyze what they see in the image. Task instructions provided a basic example without prescribing a rubric. One participant's feedback mentioned that the instruction helped them start thinking before they actually began working on the images. Several comments from the first batch mentioned the anxiety of being rejected for providing the "wrong" answer. Subsequent task instructions therefore explicit stated that any judgment showing an honest effort would not be rejected.

To ensure input quality, task instructions specified that purely personal preference was unacceptable; participants should support their interpretations with visual evidence reflecting active meaning-making. It was important to determine the level of subjectivity: on a spectrum ranging from the most objective (reCAPTCHA image transcription, for example) to the most subjective tasks ("What is your favorite ice cream flavor?"), this project focused on the relatively subjective task that still required observations of external clues. Participants were asked to support their interpretations based on the image and thereby should maintain a degree of objectivity. Answers such as "I don't know" or "I like/dislike this image" were considered insubstantial because they failed to analyze the image, whereas "I think this image shows winter or coldness because all trees are bare" constituted a quality response for this task.

Another means of quality control was to restrict the same participant's multiple inputs of the same image. Task instructions highlighted in red font that each person should not complete a HIT for more than once, but they were welcomed to work on other HITs containing different image sets. As collected results discussed later in this paper would show, participants who were interested in the task were the most

[2] https://requester.mturk.com/create/projects/new

[3] For AMT's pricing, see https://requester.mturk.com/pricing.

motivated interpreters and consistently provided high quality responses for different artworks. AMT's "qualification type" allowing requesters to assign workers pre-approved qualifications to work on certain tasks was utilized to label workers who had completed the given HIT and ensure unique inputs. Each batch of HITs published was complemented with a new qualification type that specified HIT identification number. Each worker ID was added to the corresponding qualification type when manually approving their submitted results. Workers whose approval rate showed a previous experience with other HITs in this project were checked for pre-assigned qualification type before their new submissions get approved. If a worker had already completed the HIT before, their answer would be rejected with their worker ID blocked. Of the total 144 submissions, there have only been one duplicate submission for the same HIT. Most crowd workers were cauticous to avoid being blocked; three participants sent emails after completing a HIT to ascertain that they could work on other batches after the current one.

Payment and Bonus

Submitted crowd inputs were evaluated for validity and quality. All valid responses were paid the full compensation, which was adjusted at different points in the project timeline into $0.06, $0.10, $0.12, $0.20, and $0.00 to see the changes of worker behaviors and input quality. High-quality inputs were paid a bonus that ranged from $0.03 to $0.05.

A valid response should answer both questions properly following the instructions. The answer quality was evaluated with several additional considerations. For tasks in this project, a high-quality response does not necessarily provide the "correct" answer by agreeing with expert analysis from the textbook. Instead, it should show more effort, engagement, and closer observation of the image to draw the conclusion. Collected data also showed that longer responses, in general, contained more analytical information and reflections than the brief ones. Answers that noticed multiple visual elements also reflected more efforts spent on the task.

Three responses to Franz Marc's *Deer in the Forest* (1913) reflected different levels of engagement and analytical skills asked for this task. One participant wrote: "This is a very fun image to study as the play of color and the structure allows you to see something different at each glance. The animals depicted are interesting to see." This valid response analyzed the painting in vague descriptions without providing actual visual evidence nor reaching a clear conclusion about the image's mood or syntax. Two responses rewarded with bonus provided more grounded interpretations of the artwork. One viewer focused on the color symbolism: "It seems that the image depicts the cycle of nature. It presents a sense of calm because the deer[s] are at rest. The yellow and blue colors evoke day/night feeling, while the orange and green suggest spring/fall. The yellow looks like sunlight in the forest." Another worker emphasized the artwork's sentiment: "The deers lying in the forest seem to silently communicate with one another, gauging a sentiment, almost as if it might be time to go, or one noticed something. They're all looking around. The image is [a] relatively tranquil and peaceful scene." Although quality assessment in this research remained inevitably subjective, the rule of thumb was to pay bonus for participants' time, effort, and grounded interpretations of artworks, instead of rewarding answers that overlapped with expert opinions. For paid tasks in this research, around half of valid responses received bonuses. All valid inputs for the unpaid HITs were paid bonus as compensation.

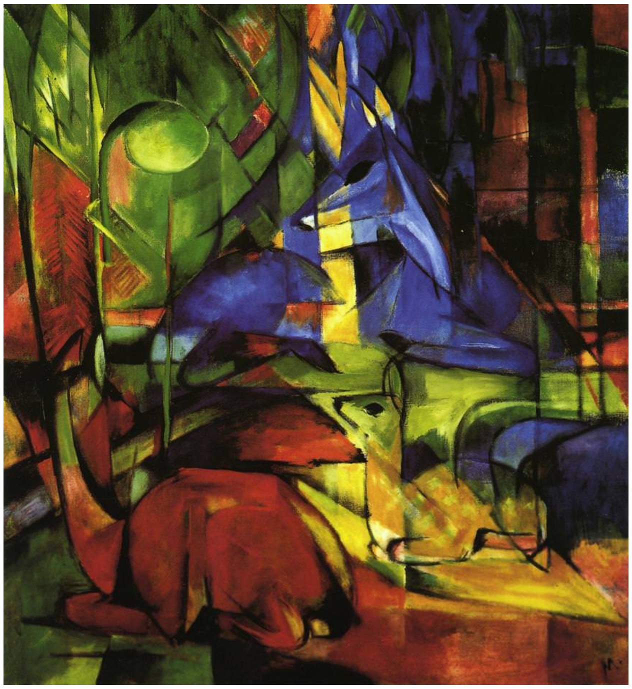


Figure 1: Franz Marc, *Deer in the Forest*, 1913

Image Selection

Table 2 lists the 30 artworks whose digital reproductions were used in this project. Considering AMT participant pool's cultural background, artworks included come from various historical periods, styles, and individuals of European and American traditions. They range from ancient Roman floor mosaic depicting domestic animals to contemporary artist Jeff Koon's porcelain sculpture of Michael Jackson. Most artworks used are figurative, with the assumption that people are generally more confident in interpreting figurative images than abstract ones. All crowdsourced responses were compared to passages from *Janson's History of Art: The Western Tradition* (Davies et al. 2010), an introductory textbook providing a general survey on Western art history with highlights from each period.

For this project, the ideal image choices were the less famous yet stylistically representative works from well-studied artists, schools, regions, or historical periods. Well-known works, such as Michelangelo's *David* (1501-1504) or Edvard Munch's *The Scream* (1893) were intentionally ommitted to avoid participants' pre-existing impressions affecting their judgments. Instead of *The Scream*, the image set included Munch's *Vampire* (1895), another painting by

Table 1: A Sample HIT

INSTRUCTIONS

You are invited to participate in a research project. You will be asked to respond to three (3) images. As compensation, you will receive $0.10 for each completed assignment. All answers will be stored anonymously.

Responses showing active engagement and contribution will be rewarded with bonus and invitations to participate in upcoming surveys.

Spam, fake, or sloppy respones will be rejected. Please provide your own interpretation and analysis; web-searched results will be rejected.

Please do not attempt to complete this HIT more than once. Don't accept this task if you have completed another HIT with the same images. You are welcomed to work on other HITs with different image sets. To ensure research data collection quality, duplicate submissions for the same task will be rejected, and your worker ID may be blocked as the result.

Task Description

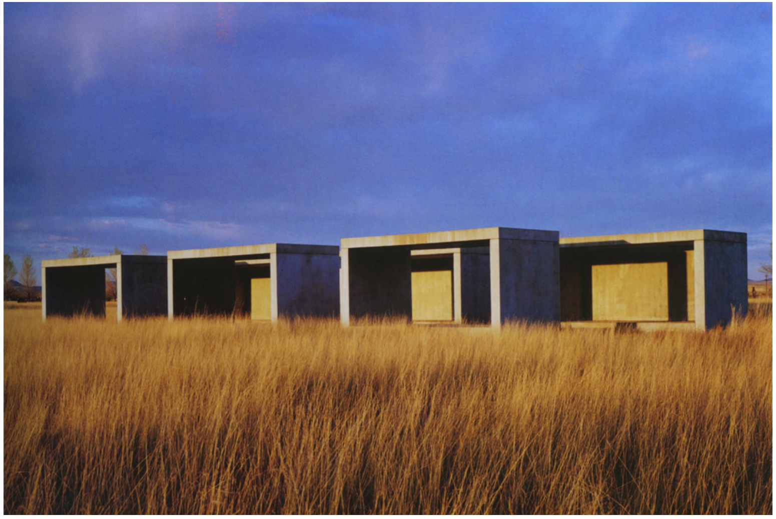
1) Write down a word to describe the main object (environment, animal, human figure, etc.) you see in the image.

2) Write a couple sentences, or a brief paragraph, to describe the mood of this image and your interpretation: how does this image make you feel? What do you think the scene is about? Why?

Example

Q: Describe the main subject in this image in one or more words:
A: one tree; a green tree; a man in red hat; a hat, etc.
Q: What is going on in this picture? In your opinion, what is this image trying to convey? How does it make you feel? Why?
A: Please answer the prompt questions by analyzing the image based on what you see in the picture as evidence. Support your interpretation with what you see in the image. We will not reject any judgment showing an honest effort. For example, ”I think it represents winter/coldness because the image shows bare tress in a blue tone” constitutes a quality response, while simply answering “I like/dislike this picture” or “this is a tree” is insubstantial because it fails to actually analyze the picture.

- Image 1 -



Please describe the main subject in this image in 1 or more words (this can be the human figure, animals, natural environment, or other objects of your first impression):

Use this section to describe the mood / sentiment / your analysis of this image. What is going on in this picture? In your opinion, what is this image trying to convey? How does it make you feel? Why?

Thank you for your time and effort!
(Optional) feedback/suggestion for us to improve this HIT:

Table 2: Image Set

Artist	**Title**	**Date**
Unknown	*Mosaic from House of Faun, Pompeii*	c. 200 - 79 B.C.
Unknown	*Fresco from Villa of Publius Fannius Synistor, Pompeii*	c. 90 - 79 B.C.
Unknown	*Book of Kells, folio 34r*	c. 800
Jean de la Huerta	*Mourner*	c. 1443-1457
Albrecht Dürer	*Praying Hand*	1508
Raphaello Sanzio da Urbino	*Portrait of Bindo Altoviti*	1512-1515
Michelangelo Buonarroti	*The Risen Christ (Drawing)*	1522-1534
Caravaggio	*Boy with a Basket of Fruit*	c. 1592
Rembrandt Harmenszoon van Rijn	*The Anatomy Lecture of Dr. Nicolaes Tulp*	1632
Jan Steen	*Beware of Luxury*	1663
Edouard Manet	*Spring*	1881
Edvard Munch	*Vampire*	1895
Paul Gauguin	*Nafea Faa Ipoipo (When Will You Marry)*	1892
Odilon Redon	*Decorative Panel*	c. 1902
Pablo Picasso	*La vie*	1903
Wassily Kandinsky	*Blue Mountain (Der blaue Berg)*	1908-1909
Odilon Redon	*Woman among the Flowers*	c. 1910
Henri Mattise	*Dance*	1910
Franz Marc	*Deer in the Forest*	1913
Marcel Duchamp	*L.H.O.O.Q.*	1919
Pierre-Auguste Renoir	*Woman among the Flowers*	c. 1910
Joan Miró	*Harlequins Carnival*	1924-1925
Francis Picabia	*Transparence - Tłte et Cheval*	c. 1930
Piet Mondrian	*Broadway Boogie Woogie*	1942-1943
Joseph Cornell	*Soap Bubble Set*	c. 1949
Joan Miró	*Composition au visage*	1965
Donald Judd	*Chinati Foundation*	1970s
Jeff Koons	*Michael Jackson and Bubbles*	1988
Gerhart Richter	*Betty*	1988
Jeff Wall	*A Sudden Gust of Wind (after Hokusai)*	1993

the same artist showcasing the sense of isolation, anxiety, and despair commonly found in Munch's oeuvre and widely recognized as a personal style. Other works were selected based on the same criterion.

In order to establish a degree of neutrality in collected interpretations, artworks selected for analysis aimed to maintain a distance between viewers and the subject matter depicted. Works that focus solely on religious or recognizable propagandistic themes, such as medieval prayer books and contemporary political satire cartoons, were excluded in this project to prevent idiosyncratic preferences and opinions as much as possible. On the other hand, artworks containing nudity and violence were still included, although none of the listed works intentionally focus on such content. Works such as Michelangelo's *Risen Christ (Drawing)* (1522-1534) and Rembrandt's *The Anatomy Lecture of Dr. Nicolaes Tulp* represent the periodic styles of Renaissance depiction of human body and seventeenth-century development of medical science. Most responses to these two images, while noticing the subject matter immediately, did not treat the adult content as their main themes. AMT allows requesters to check a box for such content and display "WARNING: This HIT may contain adult content. Worker discretion is advised" in the HIT title. To avoid confusion and frustration, fragmentary works was also excluded in this process. This criterion eliminated many ancient works that have survived in less than ideal conditions. Without providing any factual information, images in each task are designed to be interpreted with the least influences from authoritative opinions and personal prejudice as possible.

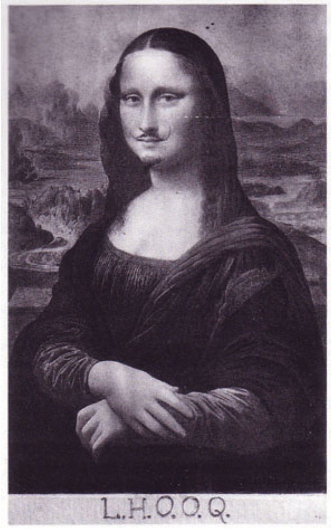


Figure 2: Marcel Duchamp, *L.H.O.O.Q.*, 1919. 19.7 x 12.4 cm. Private collection, Paris.

Judging by collected written responses, these selection criteria helped to maintain the overall quality of analysis. People appeared more cautious and neutral when encountering a new image, and more biased when responding to familiar subjects. Marcel Duchamp's *L. H. O. O. Q.* (1919) (Figure 2), an altered image of Leonardo da Vinci's famous *Mona Lisa* (1503-1517), ignited many participants' strong tendency to defend the established canon. Their word choices became less objective; many mentioned "mad," "angry," "annoying," "funny," and "made me laugh" to describe their emotional responses. Several people wrote that although *Mona Lisa* was not their favorite, "it is an important work in art history" and it "deserves to be treated seriously;" one participant further claimed that they found the image very irritating because they "can't stand graffiti."

In the case of Edouard Manet's *Spring* (1881), a painting included in the same HIT, responses collected from the same group of participants showed a clearly different disposition. Rather than expressing strongly emotional responses, the same viewers paid more attention to Manet's Impressionist color, composition, and portraiture. Most people used phrases such as "somehow not happy," "pleasant weather...looks like spring time," and "possibly from the wealthy class" to analyze the image while describing their responses as "ambivalent," "curious," or "I can't tell if she's enjoying the day." Most people made closer observations of the unfamiliar Manet painting in order to figure out its meaning, while none of them spent any word to describe the landscape or the sitting figure in Duchamp's reappropriated masterpiece—they have all focused on the added mustache and letters instead.

Results and Analysis

The project have received 144 valid and 11 invalid written responses. The latter included 1 duplicate submission from the same participant, 3 spam answers, and 7 unsatisfactory submissions that failed to complete both questions, misunderstood the HIT as an annotation task, or turning in web-search results instead of independent interpretations. These erroneous answers have been rejected with explanations; the spam worker IDs have been blocked from subsequent tasks. On average, each image of artwork collected 25 to 30 valid responses in an effective period between 3 to 5 days.

Collected answers reflect participants' different levels of familiarity with Western art history. Overall, non-expert participants recruited agreed more with expert interpretations on contemporary subjects than historical artworks, while the majority of both groups appeared more sensitive and responsive to certain visual elements than others when analyzing an image. Worker feedbacks and participation also revealed AMT crowd workers' evident interest in performing interpretive, open-ended tasks.

Crowdsourced Non-Expert Interpretations

This section addresses the performance quality of crowdsourced non-expert interpretive tasks, as raised in the research hypotheses, with an analysis of collected responses.

An overview of collected inputs showed that each published batch has received at least one analysis that largely agree with expert interpretations from the textbook. The crowdwork's quality, as suggested in the research hypotheses, can therefore be improved with the change of quantity: by collecting sufficient amounts of response, the requester is more likely to receive the "correct" answer from the crowd.

This solution cannot be combined with the consensus approach popular in crowdsourcing task design, for the "correct" answer may not received the highest worker votes. The requester would need to review and evaluate collected responses to find the desired input. Meanwhile, if we define "high-quality response" as those showing more effort, engagement, and better visual analytical skills rather than factual knowledge, a better approach to collect these high-quality responses would be quickly identifying the small portion of participants, who are interested in the task and good at visual analysis, as early as possible in the project, and recruit them as domain experts for subsequent tasks. As will be discussed in the financial incentive section below, a relatively low pay for the majority of participants while paying bonus can help requesters in identifying these domain experts and control budgets.

The hypotheses regarding crowd-generated interpretations' overlap with expert analysis have been verified with collected data. Crowdsourced interpretations showed different degrees of agreement with expert analysis for different types of artworks. The most overlap between the opinions of expert and crowd workers, as well as among different crowd workers, happened in the analysis of modern and contemporary art. Figurative contemporary art with recognizable subject matter received the most consensus and formed clear schools of thought. Abstract art received more divergent interpretations of their meanings, yet each image had a prevalent interpretation of its mood and sentiment that most crowd workers and experts agreed upon.

For *Michael Jackson and Bubbles* (1988), all 30 participants successfully identified the main subject as a man with a chimpanzee (or in some cases recognized as a monkey) and 27 of them mentioned the sculpture's golden tone in their written responses. Some recognized Michael Jackson and commented on contemporary popular culture. When analyzing the overall mood of the work, three responses used words that were also found in scholarly reception of the artist's personal style: "a feeling of false luxury or opulence;" "intentionally gaudy... extreme narcissism" and "insincere." Though participants provided different conclusions about what the sculpture tried to express, almost all of them analyzed the golden tone's color symbolism. In this case, the artist's intentional composition and color have been recognized by both professional art critics and contemporary non-expert viewers. Even for the purely abstract painting *Broadway Boogie Woogie* (1942 - 1943), without knowing the artist's love for Jazz music in the mid-century Manhattan, many responses have associated the image with positive energy, busy city life, and music.

Historical artworks received much more dissimilar interpretations based on commentators' personal preferences, experiences, and expectations. The untrained general public could not identify visual themes requiring knowledge of historical context, specialized references, and regional styles. Modern scholarship has interpreted Caravaggio's *Boy with a Basket of Fruit* (c. 1592) as a symbol of sexualized androgynous youth presented for consumption (Gregory 2011). Some crowd participants analyzed stereotypical Baroque elements in the painting, such as the intense colors and dramatic lighting, as well as the implied sexuality, but concluded that they were "confused" about what the artist tried to convey. The crowd's confusion in interpreting this naturalistically rendered human figure creates an interesting contrast with their confidence about the abstract *Broadway Boogie Woogie*. The non-expert viewers seemed particularly sensitive to cultural and temporal elements in the artworks; simply identifying the subject matter of an image appeared insufficient for people to determine meanings of the visual syntax.

As mentioned before, responses showed varied degrees of familiarity with Western art history. Some participants commented that they had seen the original or reproductions of the displayed artwork. One answer identified the relatively well-known painting *Dance* (1910) by Henri Mattise with its title, artist, and main theme, while failed to recognize other two images from the same HIT. Familiarity with periodic style also helped one to determine the overall expression and sentiment of an artwork. Another response analyzing a page from the *Book of Kells* (c. 800), a medieval religious manuscript, showed the participant's familiarity with medieval culture by describing the image as "medieval... probably from the Crusader era. It reminds me of Islamic design, but seems Western." For another artwork in the same HIT, this participant again mentioned that the image was evocative of "the long journey undertaken by the travelers in Chaucer's Canterbury Tales." Artists who had acclaimed personal styles could also be identified without knowing the artwork's factual background information. One participant recognized *La vie* (1903) as a work from Picasso's blue period, without identifying the painting's title or studied symbolism.

Visual analytical abilities, instead of factual knowledge, seem to be one of the most critical skills for accomplishing interpretive tasks in this project. Several recurring participants who completed multiple HITs with different artworks have provided high-quality analyses. Based on the experiences from this project, one way to harness the crowdsourcing of more complicated and subjective tasks is to identify and recruit desired workers as domain experts using effective task design, timely communications, and proper incentives.

Incentives

As Mason and Watts have pointed out in their study, increased financial incentives in social science crowdsourcing researches had increased quantity and not necessarily quality of the work (2010). The first two trial HITs of this research paid $0.06 for annotating and analyzing three images. They received 6 responses in 12 hours, and one participant complained in the feedback that the task took longer than they had expected and the pay should be raised. Subsequent HITs with were set to $0.10, $0.12, and $0.20. The $0.20 group received more responses in a shorter period of time compared to the other two, yet there was no significant difference in their overall qualities. When the price was adjusted from $0.20 back to $0.10 for the last 4 published batches, the result quality was also not affected significantly.

The unpaid HIT, which had collected 52 responses, provided new insights to crowd workers' motivations. The task took significantly longer to complete and received more low-quality responses than its paid counterpart, which contained the same images and was published a week after the unpaid one. However, the unpaid HIT also received slightly more amount of long written responses analyzing the images in detail than the paid HIT. 36 of all 52 participants had never worked on any other HIT from this project. Two have completed two other HITs before, and forteen had completed one. One person commented that the task "[was] the first non-payment [HIT] I've done and have nothing negative to say about it. I wish there were more paid HITs like this one." Another comment mentioned "too bad it was unpaid–I accepted the HIT because the images were interesting." Three other comments expressed similar opinions. All participants who provided valid responses to the unpaid batch were paid $0.10 bonus as compensation to encourage future participation.

Whereas financial reward is certainly a major component of worker incentives in completing a HIT, psychological factors also influences pricing and work quality in paid crowdsourcing. As the survey by Kaufmann, Schulze, and Veit has indicated, for many AMT workers. the intrinsic motivation aspects—especially "enjoyment-based motivations"—are more important than extrinsic categories such as payoffs and social motivations (Kaufmann, Schulze, and Veit 2011). Therefore, an optimized design to make the task enjoyable and encourage participation would likely help requesters to provide additional non-monetary compensations. In this research, participants have constantly complained that the pay was too low for so much writing, but they still wanted to do the task because they found the topic interesting; one commented that they enjoyed the task and felt "[this] doesn't feel like work." A potential solution to balance cost and work quality is to pay a higher bonus to engaged participants while keeping the standard reward low.

Worker-Requester Communication

AMT's design has been criticized for its lack of worker-requester interaction supports (Brawley and Pury 2016), and this project's task design aims to overcome this platform disadvantage by encouraging participants to provide constructive feedbacks within each HIT submission.

Participant feedbacks have greatly informed revisions of task design. To facilitate the feedback process, an optional final question asks for comments or suggestions. A one-sentence message sent along with bonuses also thank participants for their quality work or useful suggestions. When running the first couple trial batches, a participant sent in detailed suggestion saying that the task would cost less and receive more responses faster if the default worker qualification was changed from "master" to specific working experiences and approval rates, because the task did not require AMT's designated annotation and transcription domain expertise. Payment adjustments, typographical error checks, and the new layouts to minimize scrolling and standardize image sizes were also results of constructive participant comments.

Besides suggestions for improvement, people have also used the comment section to share their positive experience with the task. Although the research was conducted on a commercial platform instead of a citizen science project website, some participants showed much interest in it despite the low pay. Many have commented that they have never completed a task like this and were concerned that they did not provide the "correct" answers. Two people commented that they were curious to learn more about the task. Several other comments said that an artwork included in this HIT reminded them of their own experiences and they enjoyed doing the task.

Conclusion and Future Work

The main concern in the research hypotheses was the project's potential inability to attract a large and diverse pool of participants to generate meaningful results that go beyond the commonly low expectation for an untrained crowd. The collected responses have shown that paid crowd workers on AMT have both the motivation and potential to accomplish relatively more subjective interpretive tasks. Optimized designs and effective communications helped to incentivize workers' non-monetary motivations and encourage high-quality inputs.

Due to the limited scope of this project, future work can be done with various methodologies to fully explore the potential of crowdsourcing interpretive tasks. At the practical level, the automation of 1) quality control and spam filtering for subjective writing tasks that do not have a pre-established golden standard and 2) effectively comparing crowdsourced written responses to existing scholarly writings to determine similarities and differences. In this research, both steps have been done by manual screening and comparison, especially the second task in which crowdsourced answers and scholarly writings contained very different words and expressions and could become an interpretive crowdsourcing task itself ("Do these two pieces of writing agree with each other?"). Facilitating these two steps in the research would enable human requesters to handle large scale crowdsourcing projects.

This project's original plan for a comparative study with the same HITs on AMT and a citizen science project site running simultaneously failed due to the limited resources and existing restrictions. However, judging from collected analysis and feedbacks, there are perhaps more commonalities between paid and volunteer crowd workers' incentives than expected. Future comparative study would yield valuable insights to the crowd participants' motivation and the optimization of task design for different online crowdsourcing platforms.

In terms of the crowd-generated content, it would also be illuminating to compare online crowdsourced interpretations with non-expert analysis collected from other resources. For example, how does results from a university or museum-goer pool differ from those collected on AMT or another paid crowdsourcing platform? Although the "crowd" has been treated as a collective web-based human co-processor in this paper, looking into the differences between various subgroups can probably guide future requesters by understanding the characteristics that make each group truly unique.

Lastly, this paper only provides a limited experimentation with online crowdsourcing's performance of relatively subjective and open-eneded interpretive tasks. Future projects disseminating different tasks of similar nature would contribute to a more nuanced understanding of crowdsourcing's full potential.

References

Behrend, T. S.; Sharek, D. J.; Meade, A. W.; and Wiebe, E. N. 2011. The viability of crowdsourcing for survey research. *Behavior research methods* 43(3):800–813.

Brawley, A. M., and Pury, C. L. 2016. Work experiences on mturk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54:531–546.

Davies, P. J. E.; Denny, W. B.; Hofrichter, F. F.; Jacobs, J. F.; Roberts, A. S.; and Simon, D. L. 2010. *Janson's History of Art: The Western Tradition*. 8th edition. Upper Saddle River, N.J. : Pearson Prentice Hall.

Gregory, S. 2011. Caravaggio and vasari's "lives". *Artibus et Historiae, No. 64: Special issue on the quincentennial of Giorgio Vasari's birth (1511-2011)* 32:167–191.

Hughes, J. M.; Graham, D. J.; and Rockmore, D. N. Stylometrics of artwork: uses and limitations. In David G. Stork, Jim Coddington, A. B.-K., ed., *Computer Vision and Image Analysis of Art. Proceeding of SPIE-IS&T Electronic Imaging. Last access December 2, 2015. http://people.hws.edu/graham/SPIE2010_print_JH.pdf.*

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, 1–11.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* 44(1):1–23.

Nakov, P.; Kozareva, Z.; Ritter, A.; Rosenthal, S.; Stoyanov, V.; and Wilson, T. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013). http://www.aclweb.org/anthology/S13-2052* 312–320.

Nguyen, D.; Trieschnigg, D.; and Theune, M. 2014. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 321–330. ACM.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5(5):411–419.

Polatkan, G.; Jafarpour, S.; Brasoveanu, A.; Hughes, S.; and Daubechies, I. 2009. Detection of forgery in paintings using supervised learning. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2921–2924. IEEE.

Simon, N. 2010. *The participatory museum*. Museum 2.0. Accessed December 1, 2015. http//:www.participatorymuseum.org/read/.

Sontag, S. 1966. *Against Interpretation*. New York: Farrar, Straus and Giroux.

Vondrick, C.; Ramanan, D.; and Patterson, D. 2010. Efficiently scaling up video annotation with crowdsourced marketplaces. In *Computer Vision–ECCV 2010*. Springer. 610–623.