

# Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts

Anbang Xu, Shih-Wen Huang, Brian P. Bailey

Department of Computer Science

University of Illinois

Urbana, IL 61801

{xu26, shuang51, bpbailey}@illinois.edu

## ABSTRACT

Feedback on designs is critical for helping users iterate toward effective solutions. This paper presents Voyant, a novel system giving users access to a non-expert crowd to receive perception-oriented feedback on their designs from a selected audience. Based on a formative study, the system generates the elements seen in a design, the order in which elements are noticed, impressions formed when the design is first viewed, and interpretation of the design relative to guidelines in the domain and the user's stated goals. An evaluation of the system was conducted with users and their designs. Users reported the feedback about impressions and interpretation of their goals was most helpful, though the other feedback types were also valued. Users found the coordinated views in Voyant useful for analyzing relations between the crowd's perception of a design and the visual elements within it. The cost of generating the feedback was considered a reasonable tradeoff for not having to organize critiques or interrupt peers.

## Author Keywords

Feedback; crowdsourcing; creativity; critique; design.

## ACM Classification Keywords

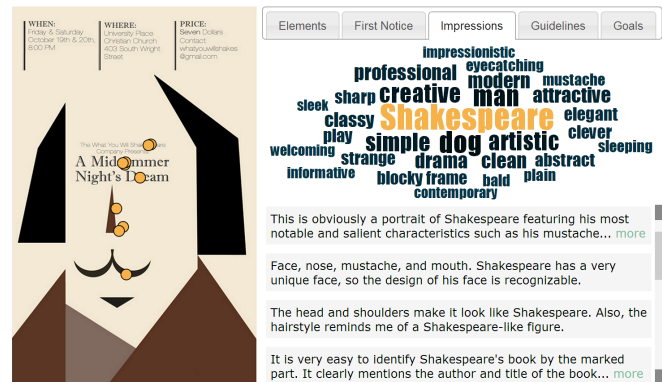
H.5.3 [Information Interface and Presentation]: Group and Organization Interfaces – Collaborative computing.

## INTRODUCTION

Feedback is a critical aspect of the creative design process [14, 18]. One key benefit of feedback is that it reveals gaps between what the user intends and what others perceive in a design [12]. For example, Figure 1 shows a poster that a user experienced in design intended to be perceived as Shakespeare, but was surprised to learn of an unintended interpretation (see word cloud, Figure 1). Gaps between intentions and interpretations are not uncommon in design and, if not addressed, can have adverse consequences [40]. Knowing where the gaps exist is therefore critical for helping users iterate toward more effective solutions [9].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CSCW'14, February 15–19, 2014, Baltimore, Maryland, USA.  
Copyright © 2014 ACM 978-1-4503-2540-0/14/02...\$15.00.



**Figure 1.** The feedback generated by Voyant on a poster. The interface is showing a word cloud representing the impressions formed when the crowd first viewed the design. Selecting an impression (in orange) shows the corresponding explanations and visual markers associating specific regions of the design.

Feedback also helps remove creative blocks [24], inform design choices [7], and gain insight into the design [31].

A user currently has several options for receiving feedback on designs. One option is organizing a critique [12], where the comprehensive feedback fosters insight into the design problem, solution approach, and strategies utilized. Though organizing a critique is warranted at key checkpoints (e.g. as scheduled by clients or instructors), it can be burdensome otherwise. Also, users without a background in design may not have the desire or peer network to organize a critique.

A second option is to informally ask peers for feedback, which incurs minimal overhead and can yield quick results. However, peer feedback can have an overly positive bias [31]. The user may also feel uncomfortable showing her design to others due to the fear of criticism [6] or may be concerned about interrupting their work [16]. A third option is to submit a design to an online community (e.g. Core77) where diverse feedback is possible. Unfortunately, studies report that little feedback is generated and the quality does not typically exceed judgments such as “I like it” [33, 37].

An emerging option for receiving feedback on designs is harnessing the wisdom of crowds. For example, researchers have recently enabled users to pose design alternatives to a crowd to collect their preferences [8]. The advantage is that crowd workers offer a potentially diverse audience, but the drawback is that the workers do not typically have expertise

in design. To enable crowds to provide additional feedback on designs (beyond preferences), it is critical to know what types of feedback users would desire from non-experts and explore how that feedback could be generated by a system and effectively aggregated and presented to users.

In this paper we present *Voyant*, a novel system giving users access to a non-expert crowd to receive structured feedback on the perceptions of their designs from a selected audience. The term *non-expert crowd* is used to refer to a crowd without expertise in design. *Voyant* generates five types of feedback: a list of elements seen in the design, the order in which the elements are noticed, the impressions formed when the design is first viewed (see Figure 1), and the interpretation of the design relative to guidelines in the domain and the user's stated communicative goals.

For each feedback type, *Voyant* provides coordinated views that consist of a visual overview of the crowd's perceptions and visual markers overlaid on the design. This interactive model enables analysis of the association between the perceptions of a design and the visual elements within it. To use *Voyant*, the user imports a design image and configures the crowd demographics. Once generated, the feedback can be utilized to help iterate toward an effective solution.

Any user creating visual designs can use *Voyant* to assess how non-experts perceive the designs. The audience of our system therefore includes novice and expert designers and non-designers who are creating posters, logos, illustrations, or Web pages for courses, research, or personal interests. We use the term *user* to refer to this broad audience.

Compared to other options for receiving feedback, *Voyant* does not require scheduling or use of social capital. It reduces the fear of criticism since the user and crowd workers are unlikely to be able to identify each other [26]. It eliminates concern about interrupting other's work by employing paid workers and produces feedback beyond simple judgments by structuring crowd workflows.

The contributions of this paper are:

- Results of a formative study identifying several types of feedback on a design that are desired from non-experts. Our results reveal specific types of feedback (e.g. on the user's goals) that are desired but not directly provided by prior work [8] or existing systems [43-45].
- The design and implementation of a system (*Voyant*) for generating the desired feedback on designs from a crowd of non-experts. A novel feature is that the user interface of the system provides coordinated views to help users analyze associations between the crowd's perception of a design and the visual elements within it.
- Results of an evaluation demonstrating the utility of the feedback generated by a crowd-based system with users and their designs. The results also demonstrate the utility of the coordinated views and interactions in *Voyant*.

## RELATED WORK

We discuss how our work is original relative to prior efforts that use crowdsourcing in design and existing methods for receiving design feedback. We also describe the creative need for receiving feedback during the design process.

### Crowdsourcing in Design

Design is a complex task and researchers have investigated many directions for how a non-expert crowd can aid design. For example, Dow et al. studied how crowd technologies can aid each phase of the design process such as collecting preferences on variations of a design (AB testing) [8]. In the CvC design method, the crowd works with the designer to form a team as part of an open design competition [25]. The designer can leverage these team members to collect their preferences on proposed design solutions. Yu and Nickerson showed how crowds and genetic algorithms can be combined to generate concept sketches [38]. Others have utilized crowds for rating ideas in innovation competitions [36] and testing interfaces based on task performance [21].

Our research is original relative to this prior work because it generates *structured feedback* on designs. For example, our approach generates the crowd's impressions when viewing a design and interpretation of the design according to the user's goals. This feedback extends prior work which solicited only individual preferences (e.g. "I like this one"). The feedback we provide is important since preferences are regarded as the least helpful aspect of design feedback [13].

Crowd-based usability Web sites can also be used to collect feedback on designs [43-45]. For example, when using Fivesecondtest [44], a user can pose free-form questions to the crowd about a design and the site returns a word cloud from the responses. In comparison, our system is original because it provides coordinated views for analyzing the association between the crowd's perception of a design and the visual elements within it. Our system also generates specific feedback (e.g. on the user's goals) not directly provided by these sites. Though the question format of the sites could be used to approximate the questions targeted by our system, the user would need to conceptualize and phrase the questions in a way that yields desired and consistent responses. Finally, we report results of the first study testing the utility of a crowd-based feedback system.

### Methods for Receiving Design Feedback

There are at least two approaches for computationally generating feedback on designs. One approach, software critics, is to encode and apply domain-specific knowledge in the form of rules to generate feedback on designs [15]. The limitation of this approach is that it cannot consider the unique goals of the user or nuances of the design problem [29]. In contrast, our approach overcomes this limitation because it is based on the coordination of *human input*. A second approach is to build computational models to predict aesthetics and affect [39], complexity and colorfulness [27], or perceptual groupings [28] using the visual features of a design. Our work generates a broader range of feedback

such as feedback on the user’s communicative goals and established guidelines in the visual domain. Also, because the feedback is synthesized from human input, it can include explanations and inspiration for improvement.

Besides software tools, a user can use social approaches to receive feedback on designs such as organizing critiques, informally asking peers, and participating in online design communities. Critique is the traditional venue for receiving feedback in the creative design domains [12, 14]. However, the critique imposes the burden of organizing discussions, which are typically needed only at scheduled checkpoints. Users without a background in design may also not have the peer network or desire to organize a critique. A user can ask peers for feedback but may be uncomfortable due to the fear of criticism [6] or concerned about interrupting their work [16]. A design can also be posted to an online community (e.g. Core77 [41] or Dribbble [42]) to receive feedback. Studies of these types of communities, however, report that little feedback is generated and the quality does not typically exceed statements such as “I like it” [33, 37].

In contrast to these social approaches, the feedback in our system can be generated on-demand by any user performing visual design work. It reduces the fear of criticism since the user and crowd workers are unlikely to identify each other and tracing a design to the user should be hard, especially if anonymized, thereby offering anonymity [26]. It removes concern about interrupting other’s work because it uses paid workers and it uses structured workflows to generate feedback that surpasses simplistic statements.

**The Creative Need for Design Feedback**

From a creative cognition perspective, feedback can foster insight or unblock creativity because the feedback serves as retrieval cues that activate new memory items and thought production [24]. This, in turn, can spark new understanding of a design, new solution approaches, or new perspectives on the design problem. For instance, prior work has shown that exposure to external visual stimuli aids brainstorming groups in producing more diverse ideas [24, 32]. Dow et al. showed that designers who receive feedback during iterative design produce higher quality outcomes than those who do not [9] and sharing multiple designs with others, due in part to increased communication, improves design exploration and outcomes [7]. Our system strives to allow these benefits to be realized by enabling users to receive feedback as often as desired during the design process.

**FORMATIVE STUDY**

A critical challenge for generating design feedback via crowdsourcing is to identify what type of feedback is desirable to the user, yet can be generated by non-experts. To address this challenge, we conducted interviews with twelve participants (six female). Eight were experienced graphic designers (experience ranged from 1 to 8 years; median was 3.5) while the other four occasionally create visual designs (e.g. data graphics, posters for conferences, or personal Web sites) as part of their work or interests but

<b>Characterizing Design Feedback</b>
Could you describe a recent or ongoing design project for which you received feedback? How did this feedback influence the design or your approach to the design?
How often do you receive feedback for a design project? How would you characterize effective and ineffective feedback?
<b>Types of Feedback Desired</b>
What type of feedback is most desirable during your process?
How do you usually receive that feedback (informal discussion, formal critique, online community)? What do you see as the key strengths and weaknesses of these methods?
<b>Desirability of Feedback from Non-Experts</b>
Imagine you could interact with a crowd of non-experts during the design process. What questions would you want to ask the crowd? What type of feedback would you want to receive?

**Table 1. The main topics and a sample of questions asked within each topic for the semi-structured interviews.**

had no formal training in design. Inclusion of these two user categories was intended to tap diverse perspectives on what type of feedback is desired during the design process. Interviews lasted about one hour and were conducted face-to-face in the user’s workspace. Participants received \$15.

The interviews were semi-structured. The participant was first asked to describe a recent or ongoing design project that served as an initial warm-up and offered context for the discussion. Twelve prepared questions were asked probing the importance of feedback in the design process, what methods are used to receive feedback and from whom, what makes feedback effective, and what type of feedback might be desired from a non-expert crowd (posed as an audience without design knowledge). Additional questions pursued interesting points raised by the participant. Table 1 shows the topics covered and a sample of questions in each one. Following qualitative analysis methods [30], the interview data was coded using a bottom-up approach. The data was segmented into logical units and a first pass was made to assign categories to the units. Subsequent passes were then performed to organize the categories into broader themes.

We had three main findings from the interviews. First, all of the participants emphasized the importance of receiving feedback during the design process. Echoing prior work [4, 29, 31], participants reported that feedback helps them better understand what is (not) being communicated in the design, inform design choices, and remove creative blocks. One participant stated: *“When I get feedback, I start thinking about things in a different way, or I start realizing something that I was overlooking because I spent so much time on the project ... It helps me come to a final thing.”* [P1]

Second, we found that participants may struggle to receive the feedback desired using existing methods. For example, similar to results reported in [31], participants reported that feedback received from peers is often overly optimistic and lacks clear direction for improvement: *“They [peers] know what went into it and how hard I worked. Maybe they don’t like*

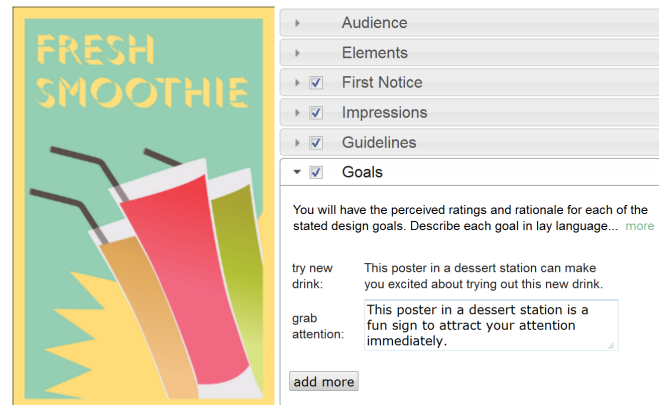
my project but they don't want to hurt my feelings. So those will be sugar-coated" [P2]. Participants who leveraged online communities such as Dribbble complained the feedback received was too little and lacked depth (echoing findings in [33, 37]). As one participant stated: "I post my work and I will get feedback occasionally. But it's usually very shallow and very positive. 'that looks nice or good work'. It is not very often that you get very true feedback on this site. It's also a follower-based community. So I have like 50 followers and some people have thousands. So they get a lot of feedback. Since I have so few, I don't get a lot of feedback." [P3]. Feedback from others knowledgeable about design was deemed useful but not typically available outside of scheduled critiques or studio work, or if one's peer network does not include designers. "There is no easy way to access feedback... I don't have many friends who are designers. For graphic design stuff, I haven't received extensive feedback... I usually just pick my favorite layout and color" [P4]. Requesting feedback is therefore considered a limited resource that must be used judiciously.

Finally, all of the participants expressed that feedback from a non-expert crowd would be useful for comparing the perceptions of the crowd to their own expectations. For example, P5 stated "We are trained designers to see certain things in certain ways. But if I can get non-trained people to understand my concept. Really see it. That means I did something really well ... There are a lot of people that have designed artifacts that can be only understood by designers. And that's a problem. Because the general public should be able to grasp your concept and understand it." P3 made a similar argument: "I'd like to get feedback from people outside of the design world. Because it's so easy to get caught inside the design bubble. A lot of times, if you are surrounded by designers, you only design for designers."

From the responses, we also identified four types of design feedback that would be desirable from non-experts. One type of feedback relates to assessing the visual hierarchy of a design: "What is the first thing people see in a design?" [P1]; "Does it stand out? What do they first read? What do they first see? What do they think they are gonna see?" [P6]. A second type of feedback mentioned is capturing first impressions when the design is viewed: "What are people's initial impressions of it?" [P4]; "What's your reaction to this or what does this make you think of? ... So I can see if that's the reaction that I want them to perceive" [P7]. A third type of feedback is whether the crowd understands the communicative goals of the design: "Does my design convey the right message to the audience?" [P8]; "Here is my goal.... How is this coming across to you?" [P9]; "How well do my concepts read?" [P10].

A fourth type of feedback tacitly mentioned is assessing a design relative to design guidelines, e.g. contrast, proximity, and alignment [34]. This feedback was inferred from many responses expressing the need for technical insights about a design and clear direction for how to improve it: "I want highly technical feedback and constructive feedback...what's wrong and how to fix it?" [P11]; "Do the colors clash or complement each other? Are there any distractions?" [P4].

Based on findings from the interviews and prior work [35, 37], we designed and implemented a novel Web application



**Figure 2.** The interface for configuring the feedback generation process. From this interface, Alice uploads her poster image, configures the desired audience (M/F, USA, 18-35) and selects which feedback is desired. For example, after selecting Goals, Alice enters her communicative goals for the poster.

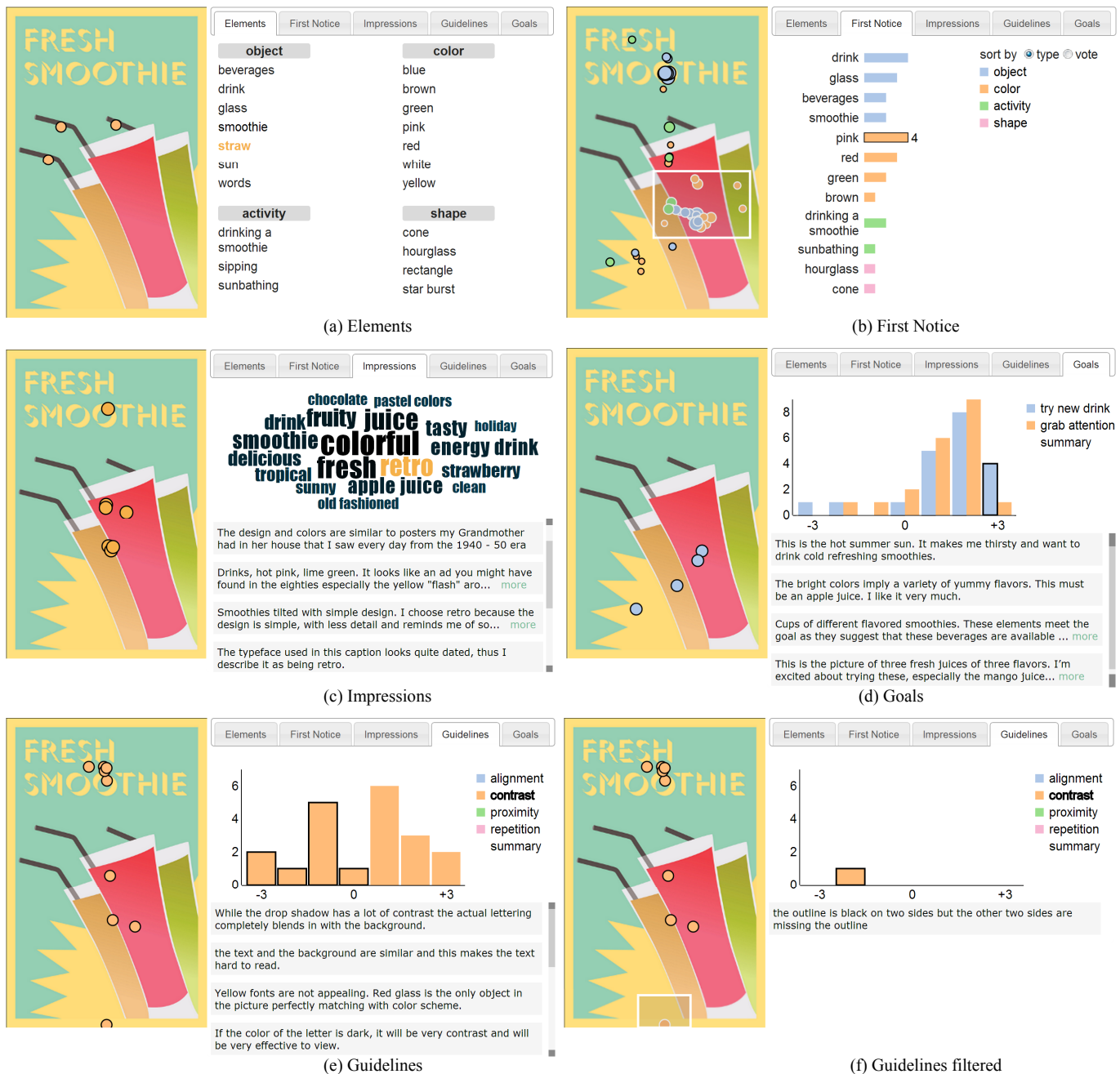
called *Voyant* that generates design feedback desired from non-experts. For the non-expert crowd, the system currently uses paid workers from an online labor market. The system accepts a visual design as input, decomposes the feedback generation process into sub-tasks suitable for non-experts, aggregates the results, and presents them to the user.

#### USER EXPERIENCE OF VOYANT

To demonstrate the utility of our system, we first describe the design feedback that can be generated, interactions for exploring the feedback, and how the feedback benefits the user. Consider a real scenario adopted from our evaluation. Alice, a novice designer, is designing a poster for a dessert station as part of a design project (Figure 2). The goal of the design is to attract attention and create excitement about trying the new service in a restaurant. Alice wants to assess how the poster is perceived by her intended audience and how these perceptions align with her expectations. Figures 2, 3, and 4 illustrate our system in context of this scenario.

To initiate feedback on a design, the user navigates to the main page of our Web application (currently restricted to our institution). From the page, the user uploads a visual design (e.g. poster, logo, or Web page) as an image, selects the feedback desired, and configures any related parameters. Five types of feedback can be selected or configured:

- *Elements* are the individual elements that are visible or otherwise perceived ("seen") in the design including colors, shapes, objects, and activities.
- *First Notice* refers to the visual order in which elements are first noticed in the design.
- *Impressions* are the perceptions formed in one's mind upon first viewing the design.
- *Goals* refer to how well the design is perceived to meet its communicative goals. If selected the user is prompted to briefly state each of her goals for the design.



**Figure 3.** The user interface offers coordinated views for each type of feedback generated on Alice’s poster. In (a), she reviews the list of elements seen by the crowd. In (b), she views which elements and corresponding regions of the design are first noticed and creates a filter to determine which elements relate to the center markers. In (c), a word cloud represents the impressions and their frequency that the crowd had upon first viewing the design. In (d), a set of interactive charts show the distributions of the crowd ratings for each communicative goal. (e) A similar set of charts shows the distributions of the ratings for each design guideline. In (e), the user selects the non-positive ratings to inspect the corresponding explanations and visual markers. In (f), the user further explores the feedback by selecting a subset of the visual markers (bottom center), which filters the ratings shown in the chart and the explanations.

- *Guidelines* refer to how well the design is perceived to meet known guidelines in the domain. Our set includes proximity, alignment, repetition, and contrast [34].

In early prototypes, the *Elements* were produced as part of the workflow for generating *First Notice*. However, user testing revealed that knowing the set of elements seen in the

design was useful on its own and was therefore extracted into its own type of feedback. The user may also configure the geography, gender, and age of the audience.

Once the feedback is configured, the system instantiates the crowd workflows, decomposes the complex process of feedback generation into micro-tasks, and submits them to

an online labor market (Mechanical Turk). Analogous to [33], the tasks are designed such that they can be performed by crowd workers with little or no design knowledge. The individual task outcomes are then aggregated and presented to the user. Each type of feedback typically requires a few hours to generate and costs a few US dollars. To ensure that the crowd workers match the audience specified by the user, prospective workers first complete a demographic survey and only those workers whose responses match the user's specification are allowed to progress to the feedback tasks.

For each feedback type, Voyant provides coordinated views that consist of a visual overview of the crowd's perceptions (e.g. in the form of a list, word cloud, or bar chart) and visual markers overlaid on the design (see Figure 3). Though coordinated views is an established visualization technique for exploratory analysis [1], we apply this technique to enable analysis of the association between the perceptions of a design and the visual elements within it. Once generated, the feedback is presented as a tabbed interface where each tab contains one type of feedback.

For *Elements*, the user receives a list of the elements "seen" by the crowd. To help the user parse the elements, they are grouped into four categories including object, color, shape, and activity, which were derived from [14]. The set of categories could be easily modified in future work. Each element can be selected to view the regions of the design associated with it. Marked regions appear as visual markers (circles) overlaid on the design. For example, in Figure 3a, selecting the element "straw" shows the corresponding visual markers placed by the workers. The benefit is the user can compare what was "seen" to what was expected. For example, in Figure 3a, some crowd workers perceived the yellow background shape in the bottom right as the "sun." The user can assess whether this perception aligns with her expectation and, if not, can use this discrepancy to seed the next iteration on the design.

The second tab *First Notice* shows the visual order in which the elements are seen. The list can be ordered globally based on the number of workers who reported seeing the element first or ordered separately in each category. Assessing the visual hierarchy is an important aspect of visual design [22]. For example, from the feedback shown in Figure 3b, the user learns that the glass with the red-colored content is the element most frequently seen first as indicated by the concentration of visual markers and the corresponding elements "red," "pink," "drink," "glass," and "beverage." Interestingly, "fresh smoothie" attracts far less attention despite its prominent size and location.

The *Impressions* tab shows a word cloud of the crowd's first impressions of the design (Figure 3c). A word cloud is used as it offers a compact representation and efficiently communicates the more and less frequent impressions. The word size is proportional to the number of workers who shared the impression. If the user selects a word, the system shows the visual markers on the design corresponding to

that impression. Conversely, the user can select an individual or group of markers on the design image to view the corresponding impressions in the word cloud and read the explanations offered by the workers. For example, from Figure 3c, one of the most common perceptions was "retro" which was associated with the title and the red and brown drink glasses in the poster. Though not shown in the figure, the user could also explore the markers placed on these elements to identify other words associated, e.g., the glasses were associated with "colorful", "fresh", and "juice".

The two remaining tabs for *Guidelines* and *Goals* provide similar forms of feedback and interaction. For each of these, Voyant presents an interactive bar chart summarizing the ratings from the crowd. The ratings were made on a seven-point scale from strongly disagree (-3) to strongly agree (+3) with whether the design met the given guideline or goal. Also, each worker could place a marker to indicate the region of the design associated with the rating and enter rationale. The user can select subsets of the guidelines or goals to inspect and compare the results (Figure 3d-f).

In addition, the user can select a subset of the bars in a chart to inspect only those markers and explanations associated with the selected ratings. For example, in Figure 3e, the user selects all the non-positive ratings for *contrast* in the chart and is then able to review the corresponding visual markers and explanations. Likewise, as shown in Figure 3f, the user can select a set of visual markers on the design to view the corresponding ratings and explanations.

From the ratings of the goals shown in Figure 3d, the user can have increased confidence that the goals are being reasonably communicated given the positive ratings. But, from the ratings of *Guidelines* in Figure 3e, the user learns that the design could use improved contrast. Inspecting the explanations reveals that many crowd workers felt the color and style of the "fresh smoothie" text did not adequately stand out relative to the background and other colors. This and other feedback can seed additional design iterations.

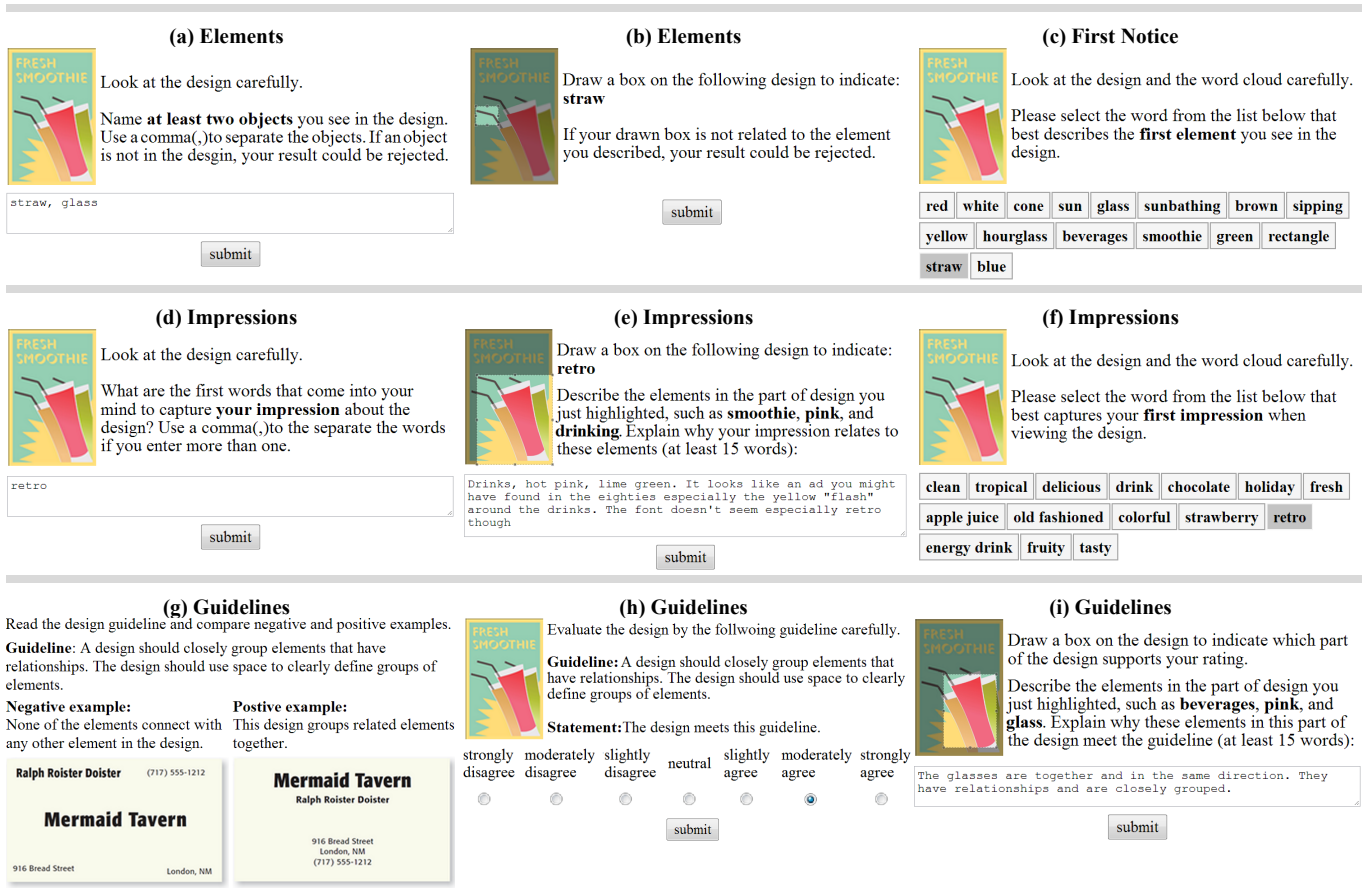
### CROWD WORKFLOWS FOR FEEDBACK GENERATION

Voyant decomposes feedback generation into a description and interpretation phase, inspired by how critique is taught in design education [14]. In each phase, the tasks focus worker attention on specific aspects of a design rather than soliciting holistic evaluations to improve outcomes [33].

#### Description

The purpose of the first phase is to enumerate what can be "seen" in a design such as objects, colors, shapes, and activities. The inclusion of this step was motivated by how novices are taught to critique visual designs by first describing what elements are seen [14]. Once identified, one can consider how the elements relate to each other, the goals of the designer, or guidelines in the domain.

To generate elements for a design, micro-tasks are created and submitted to the online labor market. The worker task screen shows the design and prompts the worker to enter at



**Figure 4.** The workflows and related task screens for generating design feedback in Voyant. For *Elements*, (a) the worker enters at least two elements seen or perceived in the design and then (b) sizes and locates a rectangle to locate each element. For *First Notice*, (c) a worker selects which element is first noticed in the design. For *Impressions*, (d) a worker enters the first word(s) that come to mind after viewing the design. For each impression entered, (e) the worker locates the most related area of the design and enters rationale. In (f), additional workers vote on the collected impressions and enter rationale. For *Guidelines*, (g) the worker learns about the guideline, (h) rates the design on the guideline, and (i) locates the area of the design that best supports the rating and enters rationale. The workflow for *Goals* is the same as that for *Guidelines*.

least two elements seen in the design relating to a given category (color, object, shape, or activity). See Figure 4a. The category prompts provide a means for focusing worker attention on the task. After a worker lists two or more elements, s/he is asked to mark regions of the design corresponding to each element listed, if applicable (Figure 4b). The markers provide an efficient mapping from what is described to what is seen. Five tasks are created for each category by default. However, if elements collected for a category are all unique, additional tasks are created for that category until a new element repeats an earlier one (a proxy for saturation). The repeated element is removed and the final set is piped into the second phase - interpretation.

**Interpretation**

In the second phase, the design is interpreted relative to the visual hierarchy, first impressions, design guidelines, and any communicative goals entered by the end user.

*First Notice Workflow*

The workflow for *First Notice* creates task screens showing the design and all the elements listed during Description. The ordering of the elements on the screen is randomized to avoid bias (Figure 4c). The worker is prompted to select which element s/he notices first in the design. The worker is then instructed to mark on the design (by placing and sizing a rectangle) where the selected element is located. From experimentation, the number of workers currently recruited is twice the number of elements listed, up to a limit of thirty. The number of times each element is selected is returned and presented to the user (Figure 3b, right panel) along with the corresponding visual markers (Figure 3b, left panel).

*Impressions Workflow*

The *Impressions* workflow generates task screens showing the design and prompts the worker to enter one or more words that come to mind when first viewing it (Figure 4d). For each word entered, the same worker is prompted to mark the design (place and size a rectangle) to indicate the

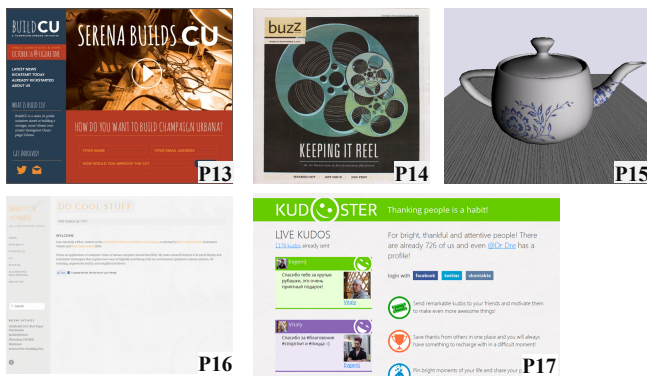


Figure 5. The designs collected for the evaluation included Web designs, a magazine cover, and an illustration. P18, P19, and P20’s designs were shown in Figures 1, 2, and 6 respectively.

region related to the impression and enter rationale (Figure 4e). As an aid, the system inserts up to three elements from the *Description* phase whose markers are nearest to the one just entered. This workflow ends once twenty tasks have been completed or thirty unique impressions are collected.

From the impression words collected, the system extracts the unique words and their frequency. Additional workers are then recruited to view the design, select which words from the existing set best captures their impression (Figure 4f), and enter rationale. This latter step is performed to further elicit the frequency of impressions from the crowd.

**Guidelines Workflow**

For the *Guidelines* workflow, a series of tasks are generated where workers rate how well the design adheres to the set of guidelines available in the system. A worker rates his or her perception of how well the design adheres to a guideline on a 7-point scale (Figure 4h). The worker is then prompted to mark a region on the design and describe how the content in that region relates to the rating (Figure 4i). This task is repeated several times per guideline, and then repeated for all guidelines selected by the user during the configuration phase. Because workers are not assumed to have knowledge of visual design, the first part of each task screen explains the guideline to be rated in lay language and shows both a positive and negative example (Figure 4g).

**Communicative Goals Workflow**

The workflow for *Goals* is similar to *Guidelines*. A series of tasks are created for each goal entered by the user during the configuration phase. A worker views the design and one of the goals and rates how well the design is perceived to

Workflow	Number of Workers	\$US
Elements	20 (5 for each element type)	1.0
First Notice	30	1.5
Impressions	50 (20 for listing; 30 for voting)	2.5
Guidelines	80 (20 for each guideline)	4.0
Goals	40 (e.g. 2 goals; 20 for each goal)	2.0

Table 2. The typical number of workers recruited and cost for each workflow in the current implementation.

Question	Avg (SD)
Q1. The feedback would help me improve my design	5.8 (1.3)
Q2. The “Impressions” feedback was helpful.	6.0 (1.2)
Q3. The “Goals” feedback was helpful.	5.4 (1.2)
Q4. The “First Notice” feedback was helpful.	5.1 (1.0)
Q5. The “Elements” feedback was helpful	4.5 (0.9)
Q6. The “Guidelines” feedback was helpful.	4.1 (1.6)
Q7. The user interface was easy to learn and use	5.2 (1.6)
Q8. The overview for each feedback type was helpful	6.2 (0.7)
Q9. It was easy to inspect details of the feedback	5.5 (1.1)
Q10. The use of the visual markers was helpful	5.4 (1.4)
Q11. I would request feedback more often with Voyant	5.1 (1.1)

Table 3. Participant ratings for the different types of feedback generated and the overall system.

meet the goal. As with *Guidelines*, the worker then marks a region on the design that best relates to the rating and enters rationale. This task is repeated several times per goal.

**IMPLEMENTATION**

Voyant was built as a Web application and can be accessed from most popular Web browsers. The functionality described in this paper has been fully implemented. The user interface was built using JavaScript, JQuery, and D3.js. The micro-tasks generated by the application are submitted to Amazon.com’s Mechanical Turk, a popular online labor market. Only workers with  $\geq 95\%$  task approval rate are considered for our tasks. Table 2 shows the default number of workers and total payment made for each feedback type. These values were determined through experimentation and could be easily adjusted or made configurable by the user.

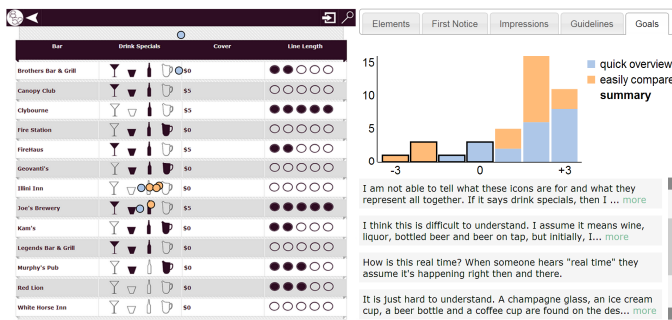
**EVALUATION**

The purpose of the evaluation was to gauge the helpfulness of the different types of feedback generated by our system, assess the usability of the visual feedback representations and interactions in the interface, and understand how the use of the system may be integrated into the design process.

Eight users (P13-20) were recruited for the study (4 female) and none had prior knowledge of our system. Four of the participants were practicing graphic designers (2-5 years of experience) who worked on print, Web, and interaction design. The other four created visual designs for personal use such as Web pages and illustrations as part of course or research projects. They had no training in visual design.

We collected one in-progress or recently completed visual design from each participant for which feedback was desired. Figure 5 shows the designs collected. For example, P14 was designing a magazine cover and wanted to visually communicate it is about a film festival. P17 was designing a Web site showing positive social work in the community and wanted to convey a warm, inviting feeling. Relevant information such as the goals and intended audience was collected for each design. Our system was then configured and used to generate the full set of feedback on each design. The time to generate the feedback on each design ranged





**Figure 6.** Part of the feedback (Goals) generated by Voyant for a Web design collected from P20. The participant explored the non-positive ratings and learned that the icons were not being perceived the way he intended and needed refinement.

from 22 to 70 hours with a median of 40 hours. The cost ranged from \$10.70 to \$13, with a median of \$11. Interviews were then scheduled with the participants. An interview lasted one hour and remuneration was \$15.

The interview consisted of a warm-up (15 minutes), system walkthrough (30 minutes), and post discussion / survey (15 minutes). In the warm-up, the participant described the feedback previously received on the design, where it came from, and how useful it was. For the walkthrough, we first introduced the system and then the participant used it to explore each feedback type at their own pace while thinking aloud. We prompted discussion about the helpfulness of the feedback, usability of the interface, and how the system can be used in practice. Afterward, the participant completed a survey. It contained eleven questions stated in a neutral manner regarding the helpfulness of the system and each type of feedback. Responses were given using a 7-point Likert-scale ranging from strongly disagree (1) to strongly agree (7). Table 3 summarizes the questions and responses.

### Helpfulness of the Feedback Overall

Participants reported that the feedback generated by our system was helpful overall and would help them improve their designs (Q1,  $\mu=5.8$ ). For example, participants stated:

*"There is some misunderstanding of my poster [see Figure 1]. Many people said that the portrait is not clear enough to show the Shakespeare's figure. They think it's a dog. That's the most important part I learn from the feedback. I will make new sketches based on the feedback."* [P18]

*"It is useful to see the number of people that have certain issues. I keep getting people's misunderstanding of the icons of the drink specials [see Figure 6]. I definitely get right away that was a major issue because there are so many people that talked about it"* [P20]

*"This is very useful feedback and something that I couldn't get in any other way."* [P17]

Specifically, the feedback helped users identify problems with their designs. For example, one of P15's design goals was to create a rendering of a teapot in Chinese style. However, she learned from the *Goals* feedback that this goal was not well supported. After exploring the visual

markers and explanations linked to the lower ratings for this goal, she gained a better understanding of the problem. To address it, she stated she would add tea cups beside the teapot to reinforce the perception of a teapot rather than a coffeepot and add a dragon pattern to communicate Chinese style. Likewise, P20 discovered from *First Notice* that "not a lot attention is given to the main navigation bar at the top... We need to do something to highlight this area more." Due to the low ratings and the corresponding explanations about contrast in *Guidelines*, P16 said that he would now increase the contrast between the foreground and background colors.

A second way the feedback was helpful is that it allowed participants to gain insight into their designs. For example, in the poster created by P19 (see Figure 3d), her color selections were intended only to make the poster "colorful." But from the feedback, she discovered that people were associating different flavors with her color selections (e.g. mango or apple to the yellow-orange smoothie) and that she could manipulate the color to cause people to literally 'taste' her design.

A third way the feedback helped is that it allowed participants to consider their designs from different perspectives, especially when the feedback contained conflicting viewpoints. For example, P13 said "Some of these positive things are contrasting what negative people said. I find myself comparing "+3" to "-3" if they are talking about the same thing... I can see why it worked for one person and see why it didn't work for another person. So it's easy for me to find a compromise between these two."

Finally, participants felt the feedback was a useful complement to prior feedback received on their designs:

*"I think the quality of the feedback is really good. Because many parts of the feedback are actually the same as the feedback I got from my professors and my classmates."* [P15]

*"It is really interesting to get people's feedback, especially if it's coming from non-designers. I'm so used to focusing so much on type. The most surprising thing is that people [non-designers] don't care much about the type as they care about the imagery."* [P14]

*"I got a lot more critical feedback. That's very hard to get from peers."* [P16]

### Helpfulness of the Different Types of Feedback

The *Impressions* feedback was reported as the most helpful among the five types of feedback generated (Q2,  $\mu=6.0$ ). From the perspectives of the participants:

*"I like this word map [Impressions] and I think this word map is incredibly helpful... Because I can think about what I was trying to achieve with this design and reflect it to the words people are using when describing this. Then I can see if I really did well enough by portraying something in my design."* [P13]

*"Impressions were the most useful because they brought up more unexpected details.... I'm looking for unusual words or data. I can see the less common terms associated with the design."* [P19]

Participants also agreed that the *Goals* (Q3,  $\mu=5.4$ ) and *First Notice* (Q4,  $\mu=5.1$ ) feedback was helpful:

*“Goals is helpful for verifying the level achievement of goals that are actionable by the audience. Goals sort of remind me of AB testing if I had multiple iterations of the design and compared the feedback on the goals.”* [P17]

*“What you see first is important. This [First Notice] will give me information on how to make things more attractive and memorable in short time frame.”* [P19]

There was moderate agreement the *Elements* feedback was helpful (Q5,  $\mu=4.5$ ). P15 felt the feedback was “a good check” on whether design elements are interpreted correctly. He reasoned that “At a basic level, those things [Elements] are very important to understand. Like if someone says that object looks like an elephant, and it’s not supposed to. That’s something that could be really helpful to know.” [P15]. Other participants felt the feedback may be too low-level.

The Guidelines feedback received lower ratings, though the average was still on the positive side of neutral (Q6,  $\mu=4.1$ ). Positive reactions included: “This is useful, especially for me. I’m bad at following common design principles” [P20]; “because I’m not a professional designer. I might forget about color, positioning, or grouping elements. If they mentioned these, it’s useful for me to keep these in mind.” [P17].

In terms of less positive reactions, some participants had strong pre-conceptions and “don’t necessarily trust non-experts to give technical feedback” [P18]. However, after reviewing the Guidelines feedback on their designs, they became less skeptical: “I like this comment. This line is not parallel to that line. It doesn’t satisfy the alignment design guideline. That’s a good point.” [P18]. Finally, two participants commented that some explanations about the given ratings were “arbitrary” or “not related” and they had to expend effort to filter these irrelevant explanations.

### Feedback Representations and Interactions

The coordinated views provided by Voyant enable exploration of the feedback from the perspective of either the overview representation or the markers on the design. Participants reported that this interactive coupling was helpful for analyzing the association between the crowd’s perception of a design and the elements within it (Q7-Q10,  $\mu \geq 5.2$ ). For example, when exploring *Impressions*, we observed participants selecting specific words to identify the locations on the design associated with them and selecting a subset of the markers to identify the associated impressions. This was supported by several statements:

*“I wasn’t expecting the word cloud to be so helpful. Because in general when I see a word cloud I usually just assume people put words together randomly. There is no data behind it. But being able to click on each word and find individual things they cite, like the actual areas in the photo [markers] they think, enforces that. That’s really valuable.”* [P13]

*“The most useful part of the interface is that I can check specific parts of the design [using visual filters]. For example, I want to see if my title in the poster works well. So I*

*just draw a box to select this part and look at the ratings and the comments.”* [P18]

*“It is the most efficient way to quickly highlight like that [draw visual filter on the design]. Oh, that part I was concerned about that. Let’s see what people thought about that part. It can help me find comments about the areas I am most interested or concerned about. I think that’s probably the best feature.”* [P14]

In addition, several participants suggested that more ways to explore the feedback would be desirable. For example, some wanted to interactively explore the feedback based on the demographics of the crowd (e.g. age and gender). This may help reveal potential patterns in the responses, aid interpretation, and lead to further insights into the design.

### Potential Use in Practice

We shared with participants the costs and response times needed to generate feedback on their designs and discussed how this might affect the use of the system in practice. The participants did not view this as a major limitation, rather they viewed it as a fair tradeoff for not having to organize a face-to-face critique or interrupt others to receive feedback:

*“No efforts from my side at all. Because I don’t need to use my social capital to do this [request feedback]. I have to call my friends at [company] and ask how my website works. They work at [company] and they are busy. But here it’s just 10 dollars”* [P17]

*“I can get feedback at any time I want. For example, I did this project at 10 or 11PM at night. I’m really excited about this. I will upload it to the tool to get feedback. But it’s very inappropriate to request feedback from my professor and classmates at that time. Also, when I do freelance projects, I don’t have chance to show my designs to them. This tool allows me to work by myself.”* [P15]

Participants reported they would use a system like Voyant to receive feedback more often in the design process (Q11,  $\mu=5.1$ ). Participants felt the system would be particularly useful after they had begun iterating on a selected design:

*“I could see myself using it in the mid to the late phase of the project. Say I have a concrete idea for the way I’m gonna approach the project. From there, after I have done a few iterations on something I was fond of. Then I may send the first or second draft down and get some feedback.”* [P14]

*“When I was doing this, I was maybe about half way through the project. I think that’s a pretty good time to take data like this and look at it.”* [P16]

Some participants also wanted to use the system to generate and compare feedback on different versions of a design. For example, P13 stated “I could see this being incredibly useful for having multiple concepts at first. Before this, I did a quite few different sketches and concepts. I could see using the impressions on these would be really useful.... I would do three different versions of this site layout and then compare the impressions.” This and other extensions are described next in Discussion.

## DISCUSSION AND FUTURE WORK

Our system helps users identify and understand problems with their designs, but it is up to the users to determine how to best modify the designs to address those problems. The fact that the feedback generated by our system can surface problems with a design is itself a significant outcome, e.g., as iterating on a design requires knowing what problems to address [9]. The user can also inspect the overviews, visual markers, and explanations to not only identify problems but also understand why they occurred. The feasibility of using crowds to brainstorm solutions to the problems identified could be explored in future work. It would also be interesting to observe how Voyant is utilized throughout a design cycle and how the generated feedback prompts iterations on a design. Such a study could reveal how the feedback benefits different design phases and types of designs, and users with different levels of design expertise.

Participants in our study appreciated exploring the feedback through the visual markers placed on their designs. Each marker was placed by a crowd worker to associate an element of the design with her perception during a task. Placing a marker was typically straightforward, but became problematic when the element could not be easily localized, e.g., background color, white space, or relations between multiple elements. System designers could improve this aspect by offering the worker different types of markers, e.g., different shapes, free-form annotation, and selectable areas or groups defined by image segmentation techniques.

The *Guidelines* feedback received ratings lower than the other types. One cause could be that the worker tasks for *Guidelines* were more complex than the others. For instance, when rating a design relative to a guideline, a worker had to comprehend the guideline, review the examples, and apply it to the design. Outcome quality may therefore have been more variable on these tasks. To improve quality, additional scaffolding for workers such as apprenticeship or showing worked examples for the tasks could be explored. For the other feedback, only minimal quality control was performed and participants in our evaluation did notice that some comments were irrelevant, which cast doubt on the corresponding input. Voyant and similar systems should therefore consider adopting additional quality controls such as use of self-assessment [10], find-fix-verify [3], validation questions [19], or task designs [20] to improve feedback. Similarly, future work should explore how to optimize the workflows in terms of the workers, costs, and time required.

Our current system generates feedback on one design at a time. However, if desired, a user could generate feedback on multiple variations of a design or competing concepts. The feedback could be used to conduct deeper comparison between the designs and complement simple AB tests (e.g. as offered in [2, 8]). A related extension is to visualize the feedback history for the evolution of a design. For example, the feedback could be shown and made accessible through an interactive, visual timeline [11]. The system also could

allow users to bookmark informative views of the feedback as a way to help guide and explain later design choices [17].

Five types of feedback are generated in Voyant, but other types may also be beneficial. For example, the system could enable dialog between the user and the crowd, e.g., by allowing questions to be posed for each other about the design. Other feedback such as the analysis of a design relative to precedents or contemporary trends [5] would also be useful, but cannot yet be generated. To the degree that a crowdsourcing or other computational approach could generate these types of feedback is an open question.

Our system recruits a crowd of workers from an online labor market, but other crowds such as the employees of a large organization, members of an online community, or students in a large design course could also be used. In the latter case, the perception-oriented feedback generated by our system could complement other forms of peer critique already integrated into the course [23]. A related future study would be to compare the feedback generated by crowds with different levels of design knowledge.

Voyant targets the visual design domain. This domain was selected because it is challenging and more people are performing visual design work due to the increased demand for visual communication skills and the wide availability of graphic design software. But, the concept of receiving structured feedback from non-experts may be applicable to other domains. For example, in architecture, a system like Voyant could enable architects to receive impressions of the form of a building or interpretation of its design goals.

## CONCLUSION

Crowdsourcing offers an emerging opportunity for users to receive rapid feedback on their designs. Our work has made three contributions in this direction. First, we contributed results from interviews identifying several types of feedback on a visual design that are desired from non-experts. A second contribution was the design and implementation of a system that generates the desired feedback on designs from a crowd of non-experts. A novel aspect of the system is that it provides coordinated views to help users analyze associations between the crowd's perceptions of a design and the elements within it. Finally, we contributed the results of an evaluation demonstrating the utility of the feedback generated by a crowd-based system for users and their own designs. Users reported that the impressions of the design and the interpretations of their goals were the most helpful, though the other feedback types provided were also valued. Users also appreciated the coordinated views and the interactions for exploring and understanding the feedback generated. Voyant exemplifies how crowdsourcing can be harnessed to enable new forms of computational tools that support design activity.

## REFERENCES

1. Baldonado, M.Q.W., Woodruff, A., & Kuchinsky, A., Guidelines for Using Multiple Views in Information Visualization. In *AVI*, (2000), 110-119.

2. Bernstein, M.S., Brandt, J., Miller, R.C., & Karger, D.R., Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *UIST*, (2011), 33-42.
3. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., et al., Soylent: A Word Processor with a Crowd Inside. In *UIST*, (2010), 313-322.
4. Cross, N. Designerly Ways of Knowing. *Design Studies*, 3, 5, (1982), 221-227.
5. Dannels, D.P. and Martin, K.N. Critiquing Critiques: A Genre Analysis of Feedback Across Novice to Expert Design Studios. *Journal of Business and Technical Communication*, 22, 2, (2008), 135-159.
6. Diehl, M. and Stroebe, W. Productivity Loss in Brainstorming Groups: Towards the Solution of a Riddle. *Journal of Personality and Social Psychology*, 53, 3, (1987), 497-509.
7. Dow, S.P., Fortuna, J., Schwartz, D., Altringer, B., & Klemmer, S., Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *CHI*, (2011), 2807-2816.
8. Dow, S.P., Gerber, E., & Wong, A., A Preliminary Study of Using Crowds in the Classroom. In *CHI*, (2013), 227-236.
9. Dow, S.P., Hedderley, K., & Klemmer, S.R., The Efficacy of Prototyping Under Time Constraints. In *Creativity & Cognition*, (2009), 165-174.
10. Dow, S.P., Kulkarni, A., Klemmer, S., & Hartmann, B., Shepherding the Crowd Yields Better Work. In *CSCW*, (2012), 1013-1022.
11. Dunne, C., Riche, N.H., Lee, B., Metoyer, R.A., & Robertson, G.G., GraphTrail: Analyzing Large Multivariate, Heterogeneous Networks while Supporting Exploration History. In *CHI*, (2012), 1663-1672.
12. Elkins, J. *Art Critiques: A Guide*. New Academia Publishing, Washington DC, 2012.
13. Feldman, E.B. The Teacher as Model Critic. *Journal of Aesthetic Education*, 7, 1, (1973), 50-57.
14. Feldman, E.B. *Varieties of Visual Experience: Art as Image and Idea*. Prentice-Hall, Englewood Cliffs, N.J., 1981.
15. Fischer, G., Nakakoji, K., Ostwald, J., Stahl, G., & Sumner, T., Embedding Computer-Based Critics in the Contexts of Design. In *CHI*, (1993), 157-164.
16. Fogarty, J., Hudson, S.E., Atkeson, C.G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J.C., et al. Predicting Human Interruptibility with Sensors. *TOCHI*, 12, 1, (2005), 119-146.
17. Heer, J., Viégas, F.B., & Wattenberg, M., Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In *CHI*, (2007), 1029-1038.
18. Hundhausen, C.D., Fairbrother, D., & Petre, M. An Empirical Study of the "Prototype Walkthrough": A Studio-Based Activity for HCI Education. *TOCHI*, 19, 4, (2012), 1-36.
19. Kittur, A., Chi, E.H., & Suh, B., Crowdsourcing User Studies With Mechanical Turk. In *CHI*, (2008), 453-456.
20. Kittur, A., Nickerson, J.V., Bernstein, M.S., Gerber, E.M., Shaw, A., Zimmerman, J., Lease, M., et al., The Future of Crowd Work. In *CSCW*, (2013), 1301-1318.
21. Komarov, S., Reinecke, K., & Gajos, K.Z., Crowdsourcing Performance Evaluations of User Interfaces. In *CHI*, (2013), 207-216.
22. Kress, G.R. and Leeuwen, T.v. *Reading Images: the Grammar of Visual Design*. Routledge, 1996.
23. Kulkarni, C. and Klemmer, S. Learning Design Wisdom by Augmenting Physical Studio Critique With Online Self-Assessment, Stanford University technical report, 2012.
24. Nijstad, B.A. and Stroebe, W. How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups. *Personality and Social Psychology Review*, 10, 3, (2006), 186-213.
25. Park, C.H., Son, K., Lee, J.H., & Bae, S.-H., Crowd vs. Crowd: Large-Scale Cooperative Design through Open Team Competition. In *CSCW*, (2013), 1275-1284.
26. Paulus, P. and Yang, H.-C. Idea Generation in Groups: A Basis for Creativity in Organizations. *Organizational Behavior and Human Decision Processes*, 82, 1, (2000), 76-87.
27. Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K.Z., Predicting Users' First Impressions of Website Aesthetics With a Quantification of Perceived Visual Complexity and Colorfulness. In *CHI*, (2013), 2049-2058.
28. Rosenholtz, R., Twarog, N.R., Schinkel-Bielefeld, N., & Wattenberg, M., An Intuitive Model of Perceptual Grouping for HCI Design. In *CHI*, (2009), 1331-1340.
29. Schön, D.A. Designing as Reflective Conversation with the Materials of a Design Situation. *Knowledge-Based Systems*, 5, 1, (1992), 3-14.
30. Strauss, A.L. *Qualitative Analysis for Social Scientists*. Cambridge University Press, 1987.
31. Tohid, M., Buxton, W., Baecker, R., & Sellen, A., Getting the Right Design and the Design Right. In *CHI*, (2006), 1243-1252.
32. Wang, H.C., Fussell, S.R., & Cosley, D., From Diversity to Creativity: Stimulating Group Brainstorming with Cultural Differences and Conversationally-Retrieved Pictures. In *CSCW*, (2011), 265-274.
33. Willett, W., Heer, J., & Agrawala, M., Strategies for Crowdsourcing Social Data Analysis. In *CHI*, (2012), 227-236.
34. Williams, R. *The Non-designer's Design Book*. Peachpit Press, 2008.
35. Xu, A. and Bailey, B.P., A Crowdsourcing Model for Receiving Design Critique. In *CHI Extended Abstracts*, (2011), 1183-1188.
36. Xu, A. and Bailey, B.P., A Reference-Based Scoring Model for Increasing the Findability of Promising Ideas in Innovation Pipelines. In *CSCW*, (2012), 1183-1186.
37. Xu, A. and Bailey, B.P., What Do You Think? A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *CSCW*, (2012), 295-304.
38. Yu, L. and Nickerson, J.V., Cooks or Cobblers? Crowd Creativity through Combination. In *CHI*, (2011), 1393-1402.
39. Zheng, X.S., Chakraborty, I., Lin, J.J., & Rauschenberger, R., Correlating Low-level Image Statistics with Users' Rapid Aesthetic and Affective Judgments of Web Pages. In *CHI*, (2009), 1-10.
40. Does This JC Penney Tea Kettle Look Like Hitler?, NBC's Today.com, <http://www.today.com/news/does-j-c-penney-tea-kettle-look-hitler-6C10100642>, Retrieved on May 28 2013.
41. Core77. <http://www.core77.com>.
42. Dribbble. <http://dribbble.com>.
43. Feedbackarmy. <http://www.feedbackarmy.com>.
44. Fivesecondtest. <http://fivesecondtest.com>.
45. Usabilla. <http://www.usabilla.com>.