Expense Control: A Gamified, Semi-Automated, Crowd-Based Approach For Receipt Capturing

Maximilian Altmeyer[1,2], **Pascal Lessel**[2,3], **Antonio Krüger**[2,1]
[1]Saarland University, [2]DFKI GmbH, [3]Saarbrücken Graduate School of Computer Science
Saarbrücken, Germany
{first name}.{last name}@dfki.de

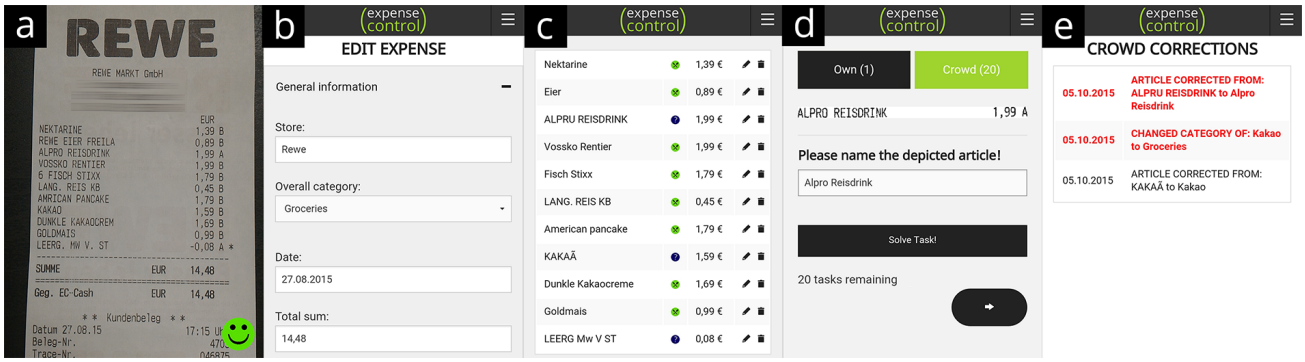

Figure 1: Workflow of our app. a) Photographing the receipt. b) Recognized general information. c) Extracted articles and categories. d) Microtask to be solved by the crowd. e) Corrections performed by the crowd

ABSTRACT

We investigate a crowd-based approach to enhance the outcome of optical character recognition in the domain of receipt capturing to keep track of expenses. In contrast to existing work, our approach is capable of extracting single products and provides categorizations for both articles and expenses, through the use of microtasks which are delegated to an unpaid crowd. To evaluate our approach, we developed a smartphone application based on a receipt analysis and an online questionnaire in which users are able to track expenses by taking photos of receipts, and solve microtasks to enhance the recognition. To provide additional motivation to solve these tasks, we make use of gamification. In a three-week-long user study (N=12), we found that our system is appreciated, that our approach reduces the error rate of captured receipts significantly, and that the gamification provided additional motivation to contribute more and thereby enrich the database.

Author Keywords

Gamification; Crowdsourcing; Wisdom of Crowds; OCR; Digital Household Accounting Book

IUI'16, March 07–10, 2016, Sonoma, CA, USA.

DOI: http://dx.doi.org/10.1145/2856767.2856790

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Optical character recognition (OCR) has improved significantly in recent years [27] and is used in various domains, for example to digitize printed documents (e.g. books) [16], to translate text in real time [35] or to provide assistive technologies for blind and visually impaired users [3]. However, OCR results are still error-prone and heavily depend on the quality of both the picture taken and the printed text [11]: the higher the quality of the picture and text, the better the results, and inversely. This makes it hard to achieve reliable OCR results in domains that are confronted with e.g. text fonts that are hard to recognize, pale ink, or crumpled paper [37,40] and gets even worse when taking pictures with a smartphone camera because of bad lighting and distortion of the picture [11].

One approach to fix spelling errors is using crowdsourcing which relies on the concept of the wisdom of crowds [34]. This concept states that a group of people is able to come to a better decision than an individual [34] and is often used to overcome problems that cannot adequately be solved by computers [7, 18] or to enhance algorithms by combining methods from artificial intelligence and crowdsourcing [4]. In this paper we investigate a crowd-based approach to enhance the outcome of OCR. Our goal is not solely to correct spelling errors, as done in existing work [1, 6, 12, 26] but also semantically enhance OCR results and attach meaningful meta-

data by the use of designated microtasks solved by an unpaid crowd. We evaluated our approach in the domain of receipt capturing to keep track of expenses with a special focus on extracting single articles, corresponding prices and an appropriate categorization. Existing attempts in this field using rule-based mechanisms [32] or machine learning methods [40] did not provide this information, probably because extracting those entities adds substantial difficulty to the OCR problem space. The increasing interest in self-tracking [24, 39], and the desire to track expenditures, additionally strengthen the need for a solution to this problem.

We developed a budgeting application (cf. Figure 1) for smartphones that allows for tracking expenses by taking pictures of receipts to extract relevant entities such as the total sum, store name, single articles and their corresponding prices, and a categorization of both the receipt and each article. To enhance the recognition algorithm, we used the outcome of different microtasks that were solved by users of our app. Solving these microtasks is not motivated by monetary rewards because it may negatively influence the quality of the generated solutions [25]. Instead, we use gamification - the use of game elements in non-game contexts [8], as it has already been successfully used in the field of crowdsourcing [6, 10, 21, 36]. The presented system is able to run in a self-sustained manner without using any external crowdsouring platforms. In a three-week-long user study, we found that the error rate when extracting entities from receipts can be significantly reduced with the help of crowd-solved microtasks and that the outcomes of these microtasks additionally improve entity extraction in future receipts. Moreover, we were able to confirm positive effects of the gamification elements on the willingness of subjects to participate.

The paper is structured as follows: we first review the challenges in performing OCR in the receipt analysis domain by inspecting German receipts from different stores, and report the results of an online study. We then consider related work and introduce how we integrated the findings into a smartphone prototype. Finally, we report and discuss the results of a user study using our prototype.

DOMAIN-SPECIFIC CHALLENGES AND REQUIREMENTS

To identify technical challenges, we inspected receipts to draw conclusions about their content and structure. We furthermore utilized an online questionnaire and reviewed popular budgeting applications informally to establish requirements for our prototype. More information about the online questionnaire and the receipt analysis can be found in [20].

Receipts Analysis

One major problem when extracting entities from receipts automatically is the absence of a uniform format [32]. We therefore inspected 117 German receipts from 85 different shops to deduce an abstract model for German receipts. We analyzed all receipts using two reviewers, each one inspecting every receipt, and thereby identified three sections (see Figure 2a):

1. **Header**: In 84.7% of all receipts, the store's name was provided as plain text. In addition, this region contains information about the store's address (95.3%), its telephone number (83.5%), the date of purchase (100%), or the website (41%). We found that the name of the store has no consistent format (since it is of arbitrary length and content); nor could be a fixed location identified for the store's name within the header region.

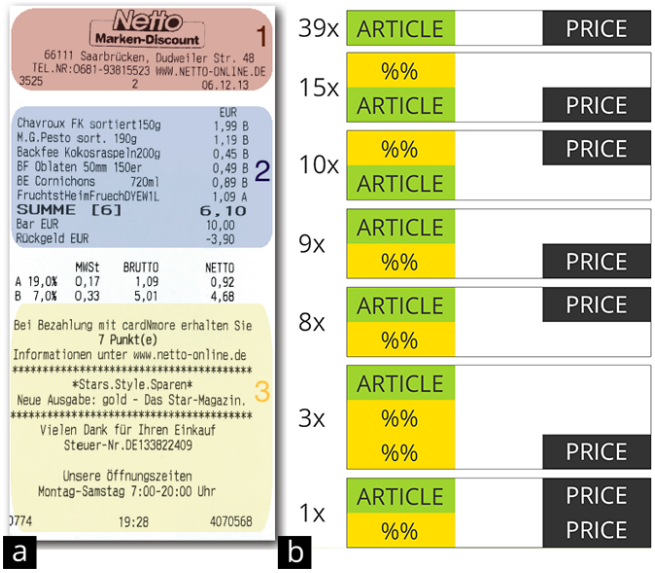

Figure 2: a) A typical receipt from a German supermarket with the identified regions: header (1), body (2), additional information (3). b) Arrangement patterns of articles and prices; % denotes additional information

2. **Body**: The body contains listed articles together with their prices as well as the total sum of the purchase. Since we aim for extracting single articles together with corresponding prices, we further examined how prices, article names and additional information (like article numbers, quantity, etc.) are aligned, and determined seven different patterns within our sample, as depicted in Figure 2b). We also analyzed the chosen wordings for the total sum of a receipt and found that there are 18 different representations used in our sample: most often the German word for "sum" was used (39), followed by "total" (31), and 16 remaining words that were distributed with high variance.

3. **Additional information**: This section contains entries not directly necessary for the capturing process. However, we found that there may be information (e.g. a web address) that could be helpful if the header is not recognized.

Based on these findings, we deduced four challenges when extracting relevant entities from receipts:

C1*: Identification and extraction of the store's name*
The analysis showed that it is hard to identify and extract the store's name because of its arbitrary length, content and format. Matters were complicated further due to the fact that no fixed location could be determined.

C2*: Identification of articles and extraction of their prices*
The layout examination of articles and corresponding prices revealed that it is not possible to simply match prices with

Figure 3: Article name (red rectangle on the left) and corresponding price (red rectangle on the right)

words of the same line to extract the article name because there is no consistent format or structure given (cf. Figure 2b). Moreover, article names are also of arbitrary length, making a purely programmatic approach more or less impossible. Figure 3 illustrates the problem of matching an article name to the corresponding price.

C3*: Categorization of single articles and the whole expense*
We consider deducing categories for single articles and the whole expense hard, since no information allowing inference of a category can be found on the receipt.

C4*: Identification and extraction of the total sum*
As our analysis reveals, although there are a lot of receipts using the same wording, which could be used to identify the total sum in most cases, there is still an incomplete set of further wordings that need to be considered.

Online Questionnaire

We setup an online questionnaire to obtain information about participants shopping behavior, their interest in keeping track of expenses, and their attitude towards automatic receipt capturing to establish requirements for our prototype. The questionnaire was available for six weeks.

Participants
We recruited 238 participants (Female: 101). Concerning age, the data is skewed young (<18: 2.73%, 18-25: 51.26%, 26-30: 13.03%, 31-40: 10.92%, 41-60: 19.33%, >60: 2.73%). This can be explained by the way the questionnaire was promoted (social media and student mailing lists).

Results
Although 82.35% of the participants were not keeping track of their expenses, only 12.18% claimed to be uninterested in doing so. Asked for causes, participants stated that the main reason is the high amount of effort to track expenses manually (stated by 57.98%). We therefore conclude that our prototype needs to ease the process of tracking expenses (**R1**). Participants reported visiting weekly farmer's markets and other specialized markets at least infrequently (5.88%), which makes a manual way to record expenses necessary (**R2**), since there are not always receipts provided. We moreover learned that the majority of people who own a smartphone (83.61%) always have it available (70.85%) while shopping or at least most of the time (21.11%). Additionally, a majority (64.71%) of participants stated that for receipt capturing they would prefer to take a photo with their smartphone. These two findings suggest that an application for mobile devices is the most suitable platform for our system (**R3**). 46.64% of the participants prefer to categorize the whole purchase instead of single items. Still, 16.38% would want every single item categorized, and 27.31% want both (single items and the whole purchase). Therefore, both options should be offered (**R4**). Concerning the evaluation period, 82.35% prefer monthly records of their purchases, which requires statistics and filters to allow for monthly aggregation (**R5**).

Informal Budgeting App Review

After establishing requirements based on the online questionnaire, we informally reviewed 10 budgeting apps, with a special focus on popular apps as reported in [2], to establish a core set of requirements for our prototype. We discovered that all reviewed apps contained statistics allowing aggregation of expenses at least by month and by category. Additionally, the amount of money spent was visualized using different types of charts. We therefore require the prototype to offer the possibility to visualize expenses in a statistical manner (**R6**). We furthermore learned that there should be a way of seeing all expenses together with the possibility to filter for different time intervals (**R7**). Other core functionalities that should be offered are editing and deleting expenses (**R8**). Based on the review, we additionally identified 10 different categories by taking the most provided categories within the applications into account. This information was used to build a basis set of categories (groceries, electronics, personal hygiene, fashion, household, pet needs, gardening, freetime, other) to be used for the categorization of expenses.

RELATED WORK

We inspected related work in the domain of crowdsourcing in general, approaches using crowdsourcing to correct OCR results, and related work in the domain of automatic receipt capturing as these domains are relevant for our work.

Crowdsourcing Picture Classification

There exist many approaches using crowdsourcing to solve problems that cannot be solved by machines in a simple way. For example, Von Ahn and Dabbish [36] developed the ESP Game, in which randomly paired players see images they need to tag. Every time they agree on a word, they receive points and the word is accepted as a valid tag. The approach confirms that a crowd can be used to annotate pictures and generate meaningful metadata, which is what we also do, but in another domain. Another example showing that a crowd not only succeeds in categorizing entities based on images, but outperforms single users in this task, can be seen in Lessel et al. [21]. The authors make use of crowdsourcing to generate classifications of waste that was inserted into an augmented recycling bin. Based on these classifications, feedback was provided to both the crowd as well as people standing in front of the trash bin, thereby educating them towards sustainable recycling behavior. This approach is also interesting for us since the crowd was not paid; instead, users were motivated through game elements, which we also aim for.

Crowdsourcing and OCR

One prominent approach is the reCAPTCHA system [37], displaying pictures of words extracted from scanned texts to people visiting websites. The extracted words are those

that are unrecognizable by OCR. By typing the corresponding word, users can confirm they are human and indirectly fix OCR errors. The approach showed that a crowd is able to correct errors even when the crowd size is low (six users are sufficient) and additionally proved that a post-correction of OCR errors by human beings increases the accuracy compared to standard OCR. These findings support our idea of using a crowd to correct errors in extracted entities. Additionally, the fact that the crowd does not need to be large is an important information for the validation process in our prototype. Nevertheless, the reCAPTCHA system is used to correct OCR errors without regarding semantic coherences, as we do in order to identify and match articles and prices.

The Australian Newspapers Digitization Project (ANDP) [16] followed the idea to use OCR corrections to improve their digitized historic newspaper articles. They created a web service that volunteers could use to correct text passages. More than 9,000 users corrected over 12.5 million lines of text, and the number of volunteers participating in this platform is growing further [31]. This approach also shows the success of a crowd correcting erroneous text regions. However, in line with the reCAPTCHA approach, the focus of ANDP is not to provide microtasks to gain semantic coherences or to use collected information to improve an extraction algorithm, as covered by our work.

Turning crowd-based OCR correction into a game was investigated in the Digitalkoot system [6]. The system aimed to improve the digitalization of old newspaper articles from the National Library of Finland. Two games were developed, both based on pictures of words already extracted by an OCR engine. Over 4,768 people were recruited in 51 days, completing more than 2.5 million tasks by using a gamified and crowd-based approach. This supports our idea of using a crowd-based approach and gamification as a motivator.

Automatic receipt capturing

Considering the digitalization of receipts, there are several attempts to be found in literature. One is the work of Zhu et al. [40] which describes an approach for automatic expense reimbursement using OCR to digitize receipts with the help of conditional random fields and regular expressions. The latter were used to extract entities with limited variation, such as phone numbers or the transaction amount. Conditional random fields were used to extract entities having large variation (e.g. store names). The results show that this system is more robust to recognition errors, but still is far from being accurate, which motivates us to investigate a crowd-based approach in this context.

Receipts2Go [19] is another system that also targets the digitalization of single-sided documents, by capturing them with a cell phone, which is in line with our idea. They also used regular expressions to extract entities with low variety. However, there is no evaluation of this concept, and the question of how well this approach performs remains open. The fact that the authors suggest qualifying an image before processing it, post-processing of OCR results and using results from extractions in the past to improve OCR results, supports our concept, in which we provide live feedback when capturing

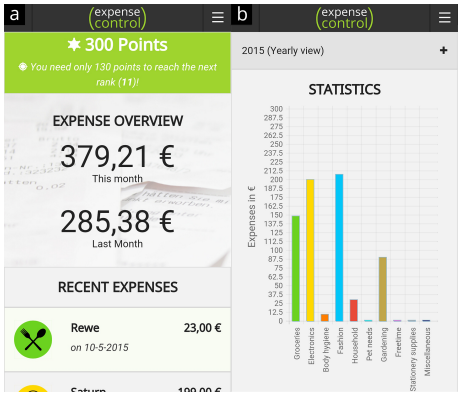

Figure 4: a) Main screen and b) statistics view of our app.

a receipt and use crowd-generated content from past extractions to infer relevant information for upcoming analysis.

In contrast to our work, both approaches considered only certain elements to be extracted: Individual articles together with corresponding prices, and a proper categorization, were not considered. Since both approaches used regular expressions for the extraction of entities with low variety, we adapted this technique in our algorithm.

The work of Shen and Tijerino [32] relies on ontologies to extract entities from receipts and uses an Object-Relationship Model, which provides information about sets of objects and their relationship, as well as constraints over object and relationship sets. As this approach relies on perfect, flawless OCR results, it may not be directly applicable in our setting, in which receipts are captured with a smartphone camera. Therefore, we additionally use crowdsourcing to correct and classify entities and try to combine information gained through the crowd with static methods described in this work.

To our knowledge, our approach is the first investigating crowdsourcing to reduce the error rate when extracting relevant entities for expense tracking. Using gamification to motivate the crowd enables our system to run in a self-sustained manner and furthermore allows to investigate the effects and the perception of gamification in this domain. This is important since there exists work demonstrating that gamification is not always successful [14] and depends on many different factors [38]. To our knowledge, there seems to be no other work considering the identification and extraction of single articles together with respective prices and categories based on a photographed receipt. In contrast to existing approaches making use of crowdsourcing only to correct text errors, we go a step further and use microtasks to obtain semantically relevant information to enhance our recognition algorithm.

SYSTEM DESIGN

To evaluate our approach, we developed a budgeting application that allows for tracking expenses by taking pictures of receipts. The design of this prototype is based on the requirements we have established, and implements solutions to the aforementioned challenges. To enhance the recognition algorithm, we used the outcome of different gamified microtasks that were solved by an unpaid crowd.

Concept

We designed the concept of our prototype based on the results presented in the last sections. To accomplish **R3**, we decided to target mobile devices and implemented our prototype as an Android application. This also seems to be beneficial regarding the ease of tracking expenses (cf. **R1**) since it offers the possibility to add expenses by taking pictures of receipts with the smartphone's camera. During the conceptualization we also performed a usability test as suggested by Nielsen [29] with 5 participants (3 female), aged 33 on average. We asked them to accomplish tasks within the app and used a think-aloud approach [28]. Participants were also asked to solve 25 microtasks including all task types. Impacts of the usability test on our concept are stated in the following sections.

Budgeting Features

On the home screen of our application (cf. Figure 4a) we show the overall amount of money spent in the current month as well as in the last month (cf. **R5**). Clicking on the monthly expenses directly takes the user to a view showing all expenses from the corresponding month, as this functionality was considered important in the usability study. Furthermore, the last expenses were shown together with categories, total sum, and the store name. The prototype furthermore provides a view in which expenses are visualized in a statistical manner (cf. Figure 4b) based on categories of single articles or the category of the overall expense, which can be set by the user (cf. **R4**, **R6**), and custom time intervals (cf. **R7**). Moreover, the app can show all expenses and filter them by year or by month (cf. **R7**, **R5**). Expenses can always be edited or deleted (cf. **R8**) and may also be added manually (cf. **R2**).

Automated Receipt Capturing

The workflow (cf. Figure 1) of adding expenses by taking a picture of the receipt starts before the user actually takes the photo. As suggested in several related works [11, 19], live feedback on the quality of the picture is given (cf. Figure 5c and Figure 5d) in form of a smiley in the camera view to enhance the OCR outcome. As soon as a picture is taken, it gets preprocessed and text is extracted by an OCR engine on the user's smartphone. We decided to outsource the whole extraction to the user's smartphone to keep the traffic as low as possible for the user and reduce the workload on the webserver, to which the result containing recognized text and corresponding line numbers is sent. On the webserver, we extract all relevant information based on the received OCR result using the outcomes of microtasks solved by the crowd.

In the following sections we describe in more detail how we realized the concept.

Receipt Capturing and Image Preprocessing

As the quality of the picture taken greatly affects the quality of the OCR result [11], we decided to implement different approaches to enhance the picture taken by the user. In a first step, we provided live feedback on the quality of the picture in form of a smiley in the camera view. Whether the smiley is green and smiling (cf. Figure 5c) or red and frowning (cf. Figure 5b) depends on lighting and the orientation of the mobile device, since both attributes have been identified as crucial for good OCR results in the literature [11, 15].

Figure 5: a) Identified horizontally aligned regions, b) Unfavorable position of the camera, c) Suitable picture, d) Region Of Interest (ROI)

To measure whether there is too much distortion or the image is skewed, we implemented an algorithm that identifies horizontally aligned regions on the given picture by using edge detection methods together with morphological closing transformations. In a further step, we calculate bounding boxes for these regions, as shown in Figure 5a). If the boxes are higher than they are wide, we can conclude that the smartphone is in an unfavorable position and distortion might be too high, since we considered words to be wider than high. Based on our receipt analysis, we can also conclude that there is too much noise in the image (e.g. because of other text elements in the background) when the x-coordinates of the bounding boxes are too heterogeneous, since we found that articles on receipts are always in alignment. After the picture was taken, we preprocessed the picture, since this further enhances OCR results [12]: In a first step we again used the algorithm described above to identify horizontally aligned regions. Based on these regions, we identified the region of interest (ROI), i.e. where the receipt is located in the picture. This was done by discarding all horizontally aligned regions where the x-coordinate of the corresponding bounding box was either lower than the calculated mean of all x-coordinates, or higher (with a certain threshold). The remaining regions were consolidated to form another bounding box which represents the region where the receipt is located (represented by the red line in Figure 5d). This area was extracted from the picture to obtain better results when thresholding it. Afterwards, we used similar approaches (thresholding, Gaussian blur, deskewing) as reported in related work [15, 19].

Entity Extraction And Consideration Of Posed Challenges

After preprocessing we used tesseract[1] to extract text regions from the picture on the user's smartphone. The extracted text is uploaded to a webserver, where three steps to extract all relevant entities from the receipt are performed:

1: Receipt Segmentation

At first, we divide the receipt into three regions: header, body and additional information (cf. Figure 2a). To identify the header, we search for the first occurrence of a price, using regular expressions. Once a price is found, the line above is considered to represent the end of the head section. Afterwards, we search for words indicating the total sum in our

[1] https://github.com/tesseract-ocr, last accessed January 5, 2016

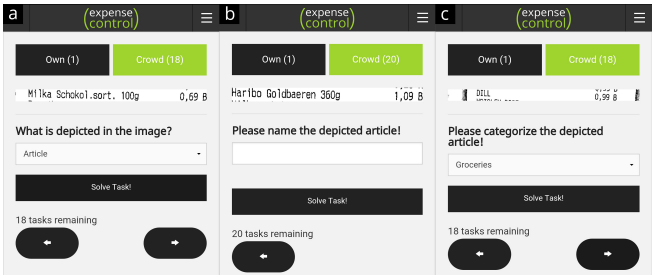

Figure 6: Crowd microtasks: a) Classification, b) Article naming, c) Article categorization.

database, that were collected by solving microtasks, to find the end of the body region. Once we find such a word, we assume the respective line to contain the total price, which can be extracted using regular expressions, targeting **C4**. Moreover the line containing the total sum represents the end of the body region. The remaining part of the receipt is therefore considered to be the additional information region.

2: Extracting the Store's Name

After dividing the receipt in the three parts, we assume to find the store's name in the header region. Moreover, as our analysis revealed, the header may contain entities like the store's address, its URL and the phone number, which serve as store identifiers and can be determined using regular expressions. Once store identifiers are extracted, they can be looked up in the database to find a matching store and thereby overcome **C1**. Every time a user corrects an extracted store in the app, we save the correction in our database and link it to corresponding store identifiers, which allows to deduce the correct merchant in future analysis.

3: Extracting Article Names and Corresponding Prices

As we learned by the receipts analysis, article names and their prices can be found in the body. Therefore, we identify all prices within the body region, using regular expressions. Next, we iterate over all lines in the body and perform a fulltext search as well as use Levenshteins distance [23] to find the content of this line in our database, which contains the outcome of all crowd-solved microtasks. We search in the table containing the corrected versions of entities but also in the table containing the raw, possibly erroneous OCR version of entities to compensate for spelling errors. Once we find a match, a classification (article, additional information) for this line can be made, based on the outcome of the associated classification microtask (cf. Figure 6a). Using these classifications, we can match one of the patterns we found in the receipts analysis (cf. Figure 2b), which allows to infer a match for an article and the corresponding price solving **C2**. Once a line is classified as an article, we can also receive its corrected name by the outcome of the respective article correction microtask (cf. Figure 6b) and its corresponding category through the result of the article categorization microtask (cf. Figure 6c) and thereby also provide a solution for **C3**.

Crowd Microtasks

Whenever a line cannot be classified properly, i.e. the entity cannot be found in our database, a microtask is generated to obtain missing information.

We decided to divide all problems or unknown entities of a receipt into microtasks, instead of creating larger tasks like correcting and classifying a whole receipt at once. This strategy is considered to lead to fewer mistakes, a greater stability in the face of interruption, and provides an easier experience for the user [5]. In the app, each microtask is shown isolated and consist of an image of the unknown receipt line and a short task description, following the suggestions given for crowd user interfaces in [30]. Users can decide to either solve own tasks, which means that these tasks correspond to problems that occurred when analyzing own receipts, or crowd tasks that were generated when analyzing receipts of other users. Every microtask can be skipped by the user, because the contained picture may be of low quality or ambiguous in some cases (e.g. if more than one line was extracted accidentally). If a microtask is skipped by at least six users, we discard the picture, similar to [37], as it cannot be used to infer proper information. However, the owner still can manually update the receipt and thereby provide a solution for the respective microtask. We decided to use three different task types, which we assume to be helpful to match articles and prices, to extract the total sum, and to categorize articles and the overall expense (cf. Figure 6):

Classification Microtasks

Based on the receipt analysis we identified three different types for entities: **article names**, **additional information** (e.g. article numbers or quantity indications) and **total sum**. Given the classification of a line (whether it contains an article, additional information or a total sum), we are able to match articles and prices and furthermore extract an overall sum for the purchase. Therefore, this task is the first task that is generated when an entity cannot be classified. The user is asked to identify the entity to be an article, additional information or total sum, as depicted in Figure 6a). In the usability test, participants had no problems classifying entities as articles or total sums but struggled to classify entities as additional information. As a reason, they stated that the wording seemed too generic for them. We therefore added more specific classification options that are internally mapped to the additional information option.

Article Correction Microtasks

This microtask is created once an unknown entity is identified as an article by the crowd. The user is asked to name a depicted article as shown in Figure 6b). The outcome of this microtask is used to correct OCR errors (spelling errors), identify and distinguish articles and provide a meaningful article name, since the articles are often abbreviated on the receipt. This task also makes it possible to store a relation between the raw text obtained by the OCR engine and the corrected version. This relation is very useful since it can compensate for typical OCR errors (e.g. confusing a zero with the letter "O" or spelling errors) and makes it possible to match abbreviated article names to their corrected versions.

Article Categorization Microtasks

Again, this microtask is generated after an entity was classified as an article, to obtain a meaningful category for it.

Figure 7: Different game elements used in the prototype. a) Points are awarded. b) An achievement is unlocked. c) Personal profile. d) Leaderboard.

The user can decide between ten different categories (cf. Figure 6c) we established in the budgeting apps reviewing process. The categorizations provided by this task are used not only for single articles but also to categorize a whole expense based on the categorizations of each purchased article, as this is required by users (cf. **R7**).

Solving Microtasks

As soon as a minimum amount of users (in our prototype we required at least six participants since [37] indicates that this amount is sufficient) participated in a microtask, a solution is generated for this task. Depending on the type of the task, the method to acquire this solution differs. A classification needs to reach at least least 60% of all votes. Once this is achieved, the raw OCR result gets stored in our database, together with the determined classification. This allows to recognize similar entities in the future, provide input for the entity extraction algorithm, and thus decrease the amount of errors. Once an entity is voted to be an article, both Article Correction and Article Categorization Microtasks are delegated to the crowd. The outcome of these tasks are determined by selecting the option with the most votes. Once an option is chosen, the corresponding article name gets corrected or respectively gets enriched by a category. The owner of the associated receipt gets notified about all changes and corrections that were performed by the crowd (cf. Figure 1e). The textual representations of these changes were significantly shortened after participants considered them as too long in the usability test.

Gamification Elements

To motivate users to solve microtasks and use our app, we integrated different gamification elements and thereby followed the design implications given in [33], stating that a budgeting app should use methods to engage users so that they keep tracking expenses over longer timescales. Gamification has been successfully used in many crowdsourcing systems to motivate and engage users [21, 37]. Furthermore, the use of gamification was able to improve crowd participation and the number of solved tasks in different domains [13]. We decided to integrate points, badges and a leaderboard together with a personal profile to provide attributes to measure fame or reputation, which is considered to motivate and retain users [9]. Points (cf. Figure 7a) can be collected by solving microtasks or receiving badges. A badge is awarded if the user solves an extraordinary amount of microtasks or regularly uses the app (e.g. by scanning receipts; an example is depicted in Figure 7b). However, we did not want to give points without any meaning, but instead decided to grant users advantages based on their score. Since solving microtasks is a service of one user for a whole crowd, we wanted to make sure that microtasks of users working a lot for the crowd are delegated with higher priority than those of users not solving many microtasks to other users. This means that problems that occur when analyzing a receipt of a user who solves many microtasks of the crowd are delegated and potentially solved faster than problems of users solving few microtasks for others. Additionally, we integrated a leaderboard (cf. Figure 7d) since competition is perceived as positive and motivating by many users [25] and is considered a solid retention scheme [9]. To further motivate users, we also showed their current rank and how many points are needed to get to the next rank directly on the home screen of our app. Unlocked badges, the amount of points and the username were summarized in a personal profile (cf. Figure 7c). Users could also see in their profile what they need to do to receive a certain badge and how their points, their rank on the leaderboard and the priorization of their problems are related to each other.

EVALUATION

By the evaluation of our approach, we tried to find evidence for the following hypotheses:

H1 Our prototype subjectively eases keeping track of expenses.

H2 The outcome of designated microtasks solved by a crowd can be used to reduce the error rate of captured receipts.

H3 The outcome of microtasks solved by a crowd can be used to reduce the error rate of new receipts that are unknown to the system.

H4 Gamification motivates users to solve microtasks.

H1 is motivated by findings from our online study suggesting that users are interested in keeping track of expenses but shy away from the huge effort involved. Since our system allows capturing receipts by photographing them, we assume that we ease the capture of expenses. **H2** builds on the assumption

that our approach using the outcomes of crowd-solved microtasks lowers the error rate, compared to a baseline without any crowd-based data, when extracting relevant information from captured receipts. **H3** is based on the assumption that solving a certain microtask not only affects a single receipt, but helps to improve other receipts to be captured in the future as well. If we can show that our approach works, i.e. that a crowd solving microtasks indeed leads to a lower error rate when extracting entities from receipts, the question remains open how to motivate a crowd to solve microtasks without using monetary incentives. This leads to **H4** which is motivated by the related work [6, 13, 21, 36] showing that gamification has positive effects in crowdsourcing.

Method

We let participants use our prototype for three weeks and asked them to capture their expenses. At the beginning of the study, the app was locked and automatically generated a password. This password needed to be used to fill out an online questionnaire assessing buying behavior as well as interest in and experiences with tracking expenses. Only after finishing this questionnaire was the app unlocked.

To investigate **H4**, gamification elements were not visible in the first week, so as to acquire a baseline. We decided against a counterbalanced measures design (i.e. reversing the order of half the participants) since deactivating game elements later could have detrimental effects on participants [14]. After the first week, the app was automatically locked again and could be unlocked by finishing the mid-session online questionnaire. In this questionnaire we provided questions about the app usage, how tracking expenses was perceived by the participants, how motivated they were towards solving microtasks, and what could be done to increase their motivation. After this questionnaire was finished, the app was unlocked again and all gamification elements activated. Two weeks later, participants were asked to take part in the post-session questionnaire, which had similar questions as the mid-questionnaire for purposes of comparison. We asked questions about how tracking expenses was perceived and whether participants were motivated to solve microtasks. Furthermore, we investigated whether certain game elements were considered motivational or not, and provided questions related to how far our system eases keeping track of expenses. Both the mid-session as well as the post-session questionnaire, were used to investigate **H1** by posing questions about how the approach was perceived subjectively concerning ease of expense tracking. We used 5-point Likert scales to measure agreement with statements participants were shown to. During the study, we logged all receipts that were added by the participants together with all solved microtasks to investigate **H2** and **H3**. In order to calculate an error rate, we needed to provide a ground truth for comparison with results of our approach. Therefore, we went through all receipts and classified, categorized and corrected all lines of these receipts manually. This was done in two steps: One person provided a ground truth for each line of every receipt that was added by the participants and the other one checked for errors (e.g. spelling mistakes or other interpretation options).

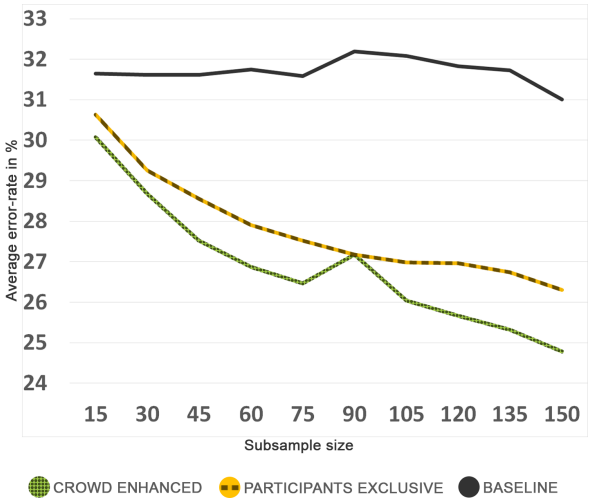

Figure 8: The error rate of new receipts for the CE algorithm (green) and the CEPE algorithm (orange) compared with the baseline algorithm (gray) for different subsample sizes.

Results

In three weeks, 191 receipts were added by 12 participants (5 female, 7 male). The age distribution was skewed young (21-30: 9, 31-40: 1, >40: 2). On average, participants added 15.92 receipts during the study (SD=14.35, Mdn=8). Before the study, only 3 subjects were keeping track of their expenses although 10 participants claimed to be interested in doing so. Three-fourths of the participants go shopping 3-4 times a week, one-eighth go shopping twice a week and one-eighth claimed to go shopping 5-6 times per week. Concerning their buying behavior, participants visit same stores (M=4.42, SD=0.52, Mdn=4) and tend to buy the same products (M=4.33, SD=0.49, Mdn=4).

Perception of the Prototype

Participants stated in the post-session questionnaire that the app eases tracking expenses (M=3.91, SD=1.04, Mdn=4) and that they would rather use our system than manually track their expenses (M=4, SD=1.18, Mdn=4). Moreover, they considered capturing expenses by taking pictures of receipts to be easy (M=4.45, SD=0.68, Mdn=5) and perceived the smiley as helpful for taking a good picture (M=4.27, SD=1.35, Mdn=5). These findings provide evidence supporting **H1**: the prototype eases keeping track of expenses. Moreover, crowd corrections were perceived positively (M=4.38, SD=0.74, Mdn=4.5) and considered meaningful (M=4.5, SD=0.54, Mdn=4.5). Participants also had the feeling that they used the app often (M=3.73, SD=1.35, Mdn=4).

Entity Extraction and Crowd Performance

During the evaluation, 15393 microtask solutions were generated by 12 participants, 1282.75 solutions per participant on average (SD=1053.54, Mdn=1101). To obtain an error ratio for each receipt, we analyzed every captured receipt again, using the outcome of all microtasks (classifications and corrections) solved by the crowd (crowd-enhanced algorithm, CE), and compared the result with an approach that solely

relies on assumptions concluded by the receipt analysis without considering the crowd (baseline algorithm). Since it is impossible to deduce categories of articles for the baseline algorithm (that does not utilize any crowd data), we excluded the category in the error rate to obtain comparable results. Thus, the error rate is calculated by the ratio of the amount of wrong entities (wrong classification and/or wrong value of a line) to the overall amount of entities (right classification of a receipt line and correct value).

The baseline algorithm produced an error rate of 31.8% (SD=22.02, Mdn=32%) whereas the CE algorithm reached an error rate of only 10.36% (SD=14.68, Mdn=5). A paired t-test showed a significant effect between these error rates ($t(190)=12.47$, $p<0.01$) supporting evidence for **H2**: the outcome of designated microtasks solved by a crowd can be used to reduce the error rate of captured receipts.

To evaluate whether microtasks of one receipt can be used to enhance newly added receipts, we picked a random subsample with 15 to 150 receipts (10 iterations, increasing the subsample size by 15 each time) and used this as a training sample. We then iterated over all receipts that were not contained in this subsample (the test sample) and applied both the crowd-enhanced algorithm as well as the baseline algorithm. In the CE algorithm, we only considered solutions of microtasks that were related to unknown entities of receipts within the selected training sample to enhance receipts in the test sample. Since participants subjectively claimed to buy the same products, we additionally performed a crowd-enhanced participants exclusive (CEPE) algorithm in which we used receipts of one user for the test set and the receipts of all other users as training sample to avoid having receipts of the same user in the test set and in the training sample. To receive more reliable results, we repeated the subsample selection 50 times for each sample size. Figure 8 shows the error rates for each subsample for all three algorithms. The results support **H3**: the outcome of microtasks solved by a crowd can be used to reduce the error rate of new receipts that are unknown to the system, since both the CE and the CEPE algorithm outperform the baseline algorithm in all sample sizes. We conducted a repeated measurements ANOVA and found a significant effect between the algorithms ($p<0.05$). Pairwise comparisons using the Bonferroni method revealed that the difference between the CE algorithm and the baseline is significant ($p<0.01$), as well as the difference between the CEPE algorithm and the baseline ($p<0.01$). Moreover, the CE algorithm performed significantly better than the CEPE algorithm ($p<0.01$). The fact that the error rate is decreasing with increasing number of receipts further suggests that our approach improves over time (with increasing data retrieved by the outcome of microtasks).

Effects and Perception of Gamification

To obtain insight about how game elements were perceived subjectively by participants, we asked questions concerning fun and engagement in the mid- and post-session questionnaire and compared answers before and after game elements were active in the app. Additionally, all used game elements were specifically addressed in the post-session questionnaire. In the mid-session questionnaire, participants subjectively had the feeling that they solved many of their own (cf. Table 1a) and many crowd microtasks (cf. Table 1b). In the post-session questionnaire the feeling of solving many of their own tasks did not change signifantly. The same was true for the feeling of having solved many crowd tasks. However, participants tended to disagree with the statement that solving their own microtasks was fun or engaging (cf. Table 1c), as well as solving crowd tasks (cf. Table 1d) in the mid-session questionnaire (before gamification was active). In the post-questionnaire for their own tasks, this perception did not change significantly, but improved for crowd tasks significantly (cf. Table 1d), suggesting gamification had an effect. Concerning the game elements used, the highscore was considered most motivating (M=3.75, SD=1.56, Mdn=4.5), followed by points (M=3.67, SD=1.50, Mdn=4) and badges (M=3.42, SD=1.31, Mdn=3.5).

Question	Mid-Session	Post-Session	sig.
a: Solved many own tasks	M=3.58, SD=1.44, Mdn=4	M=3.67, SD=1.56, Mdn=4	p=0.723
b: Solved many crowd tasks	M=4.17, SD=0.93, Mdn=4.5	M=4.17, SD=1.34, Mdn=5	p=1.0
c: Solving own tasks was fun	M=2.58, SD=1.31, Mdn=2.5	M=4.17, SD=1.34, Mdn=5	p=0.131
d: Solving crowd tasks was fun	M=2.58, SD=1.31, Mdn=2.5	M=3.42, SD=1.51, Mdn=3.5	t(11)=2.49, p<0.05

Table 1: Questions and respective answers concerning fun and engagement in the mid- and post-session online questionnaire.

Week 1	Week 2	Week 3	sig.
M=238.25, SD=209.87, Mdn=173.5, Min=27, Max=676	M=569.75, SD=462.05, Mdn=522, Min=0, Max=1258	M=474.75, SD=426.67, Mdn=415.5, Min=0, Max=1169	p<0.05

Table 2: Overview of solved tasks per user/week.

Considering the amount of solved tasks per user in each week, we performed a repeated measurements ANOVA and found a significant effect between the three weeks (cf. Table 2). Pairwise comparisons using the Bonferroni method showed that the difference between weeks 1 and 2 is significant ($p<0.05$) as well as the difference between weeks 1 and 3 ($p<0.05$), whereby week 1 was the baseline phase without any gamification elements. These results show evidence for **H4**: the use of gamification additionally motivates users to solve microtasks. However, the differences between the number of solved microtasks for each participant in every week (cf. Figure 9) and the ordinary perception of gamification, indicate that although gamification overall lead to a higher number of solved microtasks, there seems to be differences in how gamification is perceived and considered motivating. Figure 9 indicates that some participants were not affected by gamification at all. It also shows that the number of solved tasks decreases

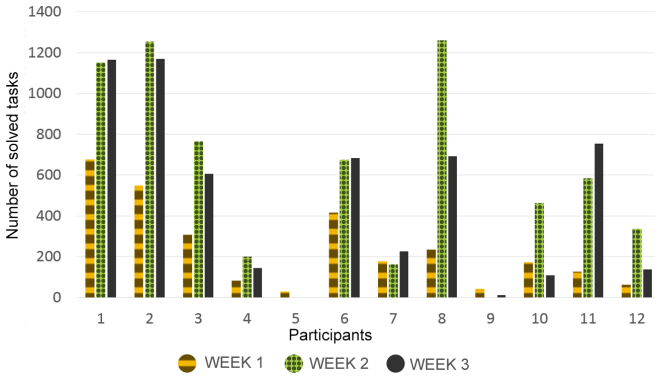

Figure 9: Solved microtasks for each participant/week.

for many participants in the last week again. Although this effect was not significant, it poses the question in how far the used game elements are able to motivate users in the long run.

Discussion

The study showed that participants are interested in keeping track of expenses and appreciate an automated approach to capture receipts. Our prototype was considered to ease the tracking process of expenses, supporting **H1**, and thereby mitigating the main reason for not keeping track of expenses: the high effort involved. The conducted analysis to measure the error ratio of receipts revealed that the crowd-based approach is able to significantly enhance the extraction of relevant information, as it produces roughly only a third as many errors as the baseline approach, showing evidence for **H2**.

The examination of how far the outcome of microtasks solved by the crowd can be used to reduce the error rate of new receipts that are unknown to the system, revealed evidence for **H3** since the error rate of both the CE and the CEPE algorithm performed better than the baseline approach for all training samples. However, the differences to the baseline are not very high, which most probably is explainable by the relatively low size of the crowd (12 participants) and the limited time of our study (3 weeks) leading to a relatively low amount of data. Further investigation is needed to find out in how far the error rate of new, unknown receipts keeps on decreasing with increasing crowd size and duration of the study. An explanation why newly added, system unknown receipts can be enhanced, lies in the increasing amount of collected data which raises the chance of finding receipt entities in the database. The reason why the CE approach performs better than the CEPE approach might be because of the fact that participants tend to buy the same articles and visit same stores (as they have reported), which increases the chance to have same articles in the test and in the training set. This, on the other hand, is beneficial for an approach as done by us, since a user that is solving her own tasks can thereby improve the algorithm to better recognize her products in the future. The number of solved microtasks was significantly influenced by the use of gamification: It increased dramatically in the second week after gamification was introduced and subjectively, solving crowd-tasks appeared to be more fun. However, we also found that the number of solved microtasks got lower in the last week of the investigation, but still was significantly higher than in week one. These results show supporting evidence for **H4**. Nonetheless, as we only investigated the prototype short-term, it is questionable whether this is only a novelty effect. Moreover, the high variance of solved tasks per user suggests that the gamification elements are not motivating for all users. An explanation for this can be found in [17] showing that competition, as was used in our system, can also be demotivating. Figure 9 also shows that two participants were not interested in solving microtasks, independent of whether gamification was active or not, thus necessitating further incentive mechanisms. Again, a long-term study with more participants seems suitable to explore the impacts of gamification on motivation further. A larger amount of participants also allows to have a control group without any game elements throughout the study. We decided against a control group in this work as we wanted to investigate whether the number of solved microtasks increases with gamification for each subject, which demands a within-subject design, especially considering our low participant count.

CONCLUSION

In this paper we investigated a crowd-based approach to enhance the outcome of optical character recognition in the domain of receipt capturing to keep track of expenses. We developed a prototype which made it possible to track expenses by taking photos of receipts and not only recognizes entities, but also enhances them with semantic information. At the same time the presented system does not rely on external crowdsourcing platforms (e.g. Amazon Mechanical Turk) but is able to run in a self-sustained manner without using monetary incentives. We based our approach on an online questionnaire to gain information about participants shopping behavior, their interest in keeping track of expenses and their attitude towards automatic receipt capturing, as well as on a receipt analysis to obtain insights about technical challenges in this field. In a three-week-long user study, we were able to show that our approach was appreciated by the users and eases keeping track of expenses subjectively. This mitigates the main reason for not keeping track of expenses: the high effort involved. We furthermore were able to show that our approach significantly reduces the error rate for captured receipts, and that the outcome of crowd-solved microtasks can be used to enhance recognition for new receipts as well. Moreover, we found evidence that gamification provided additional motivation to users to contribute more, solve a higher amount of microtasks and thereby enrich the database.

In future work we plan to investigate the long-term effects of our approach to test in how far the error-rate further decreases when accumulating more data and whether gamfication can be used to keep users engaged over a longer timespan. Therefore we aim to conduct an in-the-wild study and release the app in the market. Before, we plan to revise our gamification concept in order to motivate a broader range of users and investigate how we can change the design and concept of the microtasks to be more fun and engaging. One promising approach to accomplish this is described in [22], in which users are able to decide about the game elements they want to use, thus leading to a more tailored gamification design.

REFERENCES

1. Hend Al-Rouqi and Hend S. Al-Khalifa. 2014. Making Arabic PDF Books Accessible Using Gamification. *Proceedings of the 11th Web for All Conference. ACM, 2014.* (2014), 1–4. DOI: http://dx.doi.org/10.1145/2596695.2596712
2. Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. 2011. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM, 2011.* (2011), 47. DOI: http://dx.doi.org/10.1145/2037373.2037383
3. Erin Brady, Meredith Morris, Yu Zhong, Samuel White, and Jeffrey Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.* (2013), 2117–2126. DOI: http://dx.doi.org/10.1145/2470654.2481291
4. Justin Cheng and Michael S Bernstein. 2015. Flock : Hybrid Crowd-Machine Learning Classifiers. *CSCW '15 Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015).
5. Justin Cheng, Jaime Teevan, Shamsi Iqbal, and Michael Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.* (2015), 4061–4064.
6. Otto Chrons and Sami Sundell. 2011. Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. *Human Computation. 2011.* (2011), 20–25. http://cdn.microtask.com/research/Digitalkoot-HCOMP2011-Chrons-Sundell.pdf
7. Gabriel de la Cruz, Bei Peng, Walter Lasecki, and Matthew Taylor. 2015. Towards Integrating Real-Time Crowd Advice with Reinforcement Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion. ACM, 2015.* (2015), 17–20. DOI: http://dx.doi.org/10.1145/2732158.2732180
8. Sebastian Deterding and Dan Dixon. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". *Proceedings of the 15th International Academic MindTrek Conference. ACM, 2011.* (2011), 9–15. DOI: http://dx.doi.org/10.1145/2181037.2181040
9. Anhai Doan, Raghu Ramakrishnan, and Alon Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM 54.4 (2011).* 54, 4 (2011), 86–96. DOI: http://dx.doi.org/10.1145/1924421.1924442
10. Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 2012. Quality Through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012.* (2012), 871. DOI: http://dx.doi.org/10.1145/2348283.2348400
11. Daniel Esser, Klemens Muthmann, and Daniel Schuster. 2013. Information Extraction Efficiency of Business Documents Captured with Smartphones and Tablets. *Proceedings of the 2013 ACM Symposium on Document Engineering. ACM, 2013.* (2013), 111–114. DOI: http://dx.doi.org/10.1145/2494266.2494302
12. Paula Estrella and Pablo Paliza. 2014. OCR Correction of Documents Generated during Argentina's National Reorganization Process. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM, 2014.* (2014), 119–123.
13. Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives. *Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015.* (2015), 333–343.
14. Juho Hamari and Harri Sarsa. 2014. Does Gamification Work ? A Literature Review of Empirical Studies on Gamification. *47th Hawaii International Conference on System Sciences. IEEE, 2014* (2014), 3025–3034. DOI: http://dx.doi.org/10.1109/HICSS.2014.377
15. Rose Holley. 2009a. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine 15.3/4 (2009)* (2009). DOI: http://dx.doi.org/10.1017/CBO9781107415324.004
16. Rose Holley. 2009b. Many Hands Make Light Work : Public Collaborative OCR Text Correction in Australian Historic Newspapers. *National Library of Australia Staff Papers (2009)* March (2009), 1–28.
17. Chin-Lung Hsu and Hsi-Peng Lu. 2004. Why do People Play On-Line Games? An Extended TAM with Social Influences and Flow Experience. *Information & Management 41.7 (2004).* 41, 7 (2004), 853–868. DOI: http://dx.doi.org/10.1016/j.im.2003.08.014
18. Shih-Wen Huang, Pei-Fen Tu, Wai-Tat Fu, and Mohammad Amanzadeh. 2013. Leveraging the Crowd to Improve Feature-Sentiment Analysis of User Reviews. *Proceedings of the 2013 International Conference on Intelligent User Interfaces. ACM, 2013.* March 1922 (2013), 3–14. DOI: http://dx.doi.org/DOI=10.1145/2449396.2449400
19. Bill Janssen, Eric Saund, Eric Bier, Patricia Wall, and Mary Ann Sprague. 2012. Receipts2Go: The Big World of Small Documents. *Proceedings of the 2012 ACM Symposium on Document Engineering. ACM, 2012.* (2012), 121—-124. http://dl.acm.org/citation.cfm?id=2361381

20. Frederic Kerber, Pascal Lessel, Maximilian Altmeyer, Annika Kaltenhauser, Christian Neurohr, and Antonio Krüger. 2014. Towards a Novel Digital Household Account Book. *CHI'14 Extended Abstracts on Human Factors in Computing Systems. ACM, 2014.* (2014), 1921–1926. DOI: http://dx.doi.org/10.1145/2559206.2581288
21. Pascal Lessel, Maximilian Altmeyer, and Antonio Krüger. 2015. Analysis of Recycling Capabilities of Individuals and Crowds to Encourage and Educate People to Separate Their Garbage Playfully. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2015.* (2015), 1095–1104. DOI: http://dx.doi.org/10.1145/2702123.2702309
22. Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2016. Don't Whip Me With Your Games - Investigating Bottom-Up Gamification. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2016.* (2016), (to appear). DOI: http://dx.doi.org/10.1145/2858036.2858463
23. Vladimir I. Levenshtein. 1966. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*. Vol. 10.
24. Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2010.* (2010), 557. DOI:http://dx.doi.org/10.1145/1753326.1753409
25. Elaine Massung, David Coyle, Kirsten F. Cater, Marc Jay, and Chris Preist. 2013. Using Crowdsourcing to Support Pro-Environmental Community Activism. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.* (2013), 371–380. DOI: http://dx.doi.org/10.1145/2470654.2470708
26. Günter Mühlberger, Johannes Zelger, and David Sagmeister. 2014. User-Driven Correction of OCR Errors: Combining Crowdsourcing and Information Retrieval Technology. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM, 2014.* 52, 1 (2014), 53–56.
27. M.P. Nevetha and A. Baskar. 2015. Applications of Text Detection and its Challenges : A Review. *Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015.* (2015), 712–721.
28. Jakob Nielsen. 1992. Evaluating the Thinking-Aloud Technique for Use by Computer Scientists. (1992).
29. Jakob Nielsen. 2000. Why You Only Need to Test with 5 Users. (2000).
30. Bahareh Rahmanian and Joseph G. Davis. 2014. User Interface Design for Crowdsourcing Systems. *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14* (2014), 405–408. DOI: http://dx.doi.org/10.1145/2598153.2602248
31. Charalampos Saitis, Andrew Hankinson, and Ichiro Fujinaga. 2014. Correcting Large-Scale OMR Data with Crowdsourcing Categories and Subject Descriptors. *Proceedings of the 1st International Workshop on Digital Libraries for Musicology* (2014), 1–3.
32. Zhinian Shen and Yuri Tijerino. 2012. Ontology-Based Automatic Receipt Accounting System. *International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM. Vol. 3. IEEE, 2012* (2012), 236–239. DOI: http://dx.doi.org/10.1109/WI-IAT.2012.265
33. Stephen Snow and Vyas Dhaval. 2015. Fixing the Alignment: An Exploration of Budgeting Practices in the Home. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. ACM, 2015.* (2015), 2271–2276.
34. James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.
35. Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise. 2014. A Mixed Reality Head-Mounted Text Translation System Using Eye Gaze Input. *Proceedings of the 19th International Conference on Intelligent User Interfaces. ACM, 2014.* (2014), 329–334. DOI: http://dx.doi.org/10.1145/2557500.2557528
36. Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2004.* (2004), 319 – 326. DOI: http://dx.doi.org/10.1145/985692.985733
37. Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science 321.5895* 321, 5895 (2008), 1465–1468. DOI: http://dx.doi.org/10.1126/science.1160379
38. Erika Noll Webb. 2013. Gamification : When It Works, When It Doesnt. *Design, User Experience, and Usability. Health, Learning, Playing, Cultural, and Cross-Cultural User Experience. Springer Berlin Heidelberg, 2013* (2013), 608–614.
39. Mark Whooley, Bernd Ploderer, and Kathleen Gray. 2014. On the Integration of Self-Tracking Data Amongst Quantified Self Members. *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014-Sand, Sea and Sky-Holiday HCI. BCS, 2014.* February (2014), 151–160. DOI: http://dx.doi.org/10.14236/ewic/hci2014.16
40. Guangyu Zhu, Timothy J. Bethea, and Vikas Krishna. 2007. Extracting Relevant Named Entities for Automated Expense Reimbursement. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007.* (2007), 1004–1012. DOI: http://dx.doi.org/10.1145/1281192.1281300