Dually Noted: Layout-Aware Annotations with Smartphone Augmented Reality

Jing Qian
jing_qian@brown.edu
Brown University
Providence RI, USA

Qi Sun
qisun0@nyu.edu
New York University
Brooklyn, NY, USA

Curtis Wigington
wigingto@adobe.com
Adobe Research
San Jose, CA, USA

Han L. Han
han.han@lri.fr
Université Paris-Saclay
Inria, France

Tong Sun
tsun@adobe.com
Adobe Research
San Jose, CA, USA

Jennifer Healey
jennifer.a.healey@gmail.com
Adobe Research
San Jose, CA, USA

James Tompkin
james_tompkin@brown.edu
Brown University
Providence, RI, USA

Jeff Huang
jeff_huang@brown.edu
Brown University
Providence, RI, USA

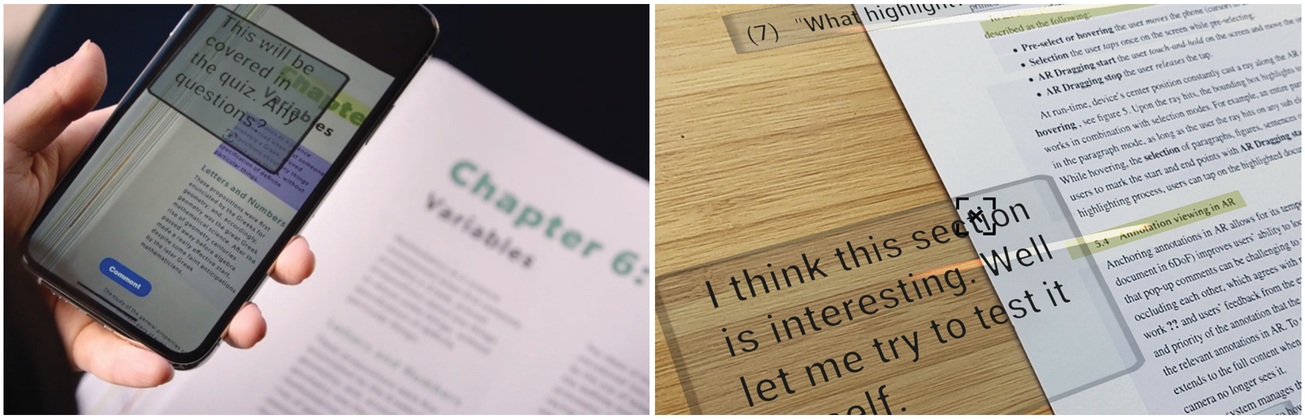

This will be covered in the quiz. Any questions?
Comment
Chapter 6: Variables
(7) "What highlight
I think this section is interesting. Well let me try to test it myself.

Figure 1: *Dually Noted* is a document annotation system for smartphone augmented reality (AR). It significantly improves user precision and efficiency in selecting and annotating text and images. Users in remote collaboration can freely share comments without permanently marking a physical book. *Left:* Our interface. *Right:* The annotated document on the smartphone screen.

ABSTRACT

Sharing annotations encourages feedback, discussion, and knowledge passing among readers and can be beneficial for personal and public use. Prior augmented reality (AR) systems have expanded these benefits to both digital and printed documents. However, despite smartphone AR now being widely available, there is a lack of research about how to use AR effectively for interactive document annotation. We propose Dually Noted, a smartphone-based AR annotation system that recognizes the layout of structural elements in a printed document for real-time authoring and viewing of annotations. We conducted experience prototyping with eight users to elicit potential benefits and challenges within smartphone AR, and this informed the resulting Dually Noted system and annotation interactions with the document elements. AR annotation is often unwieldy, but during a 12-user empirical study our novel structural understanding component allows Dually Noted to improve precise highlighting and annotation interaction accuracy by 13%, increase interaction speed by 42%, and significantly lower cognitive load over a baseline method without document layout understanding. Qualitatively, participants commented that Dually Noted was a swift and portable annotation experience. Overall, our research provides new methods and insights for how to improve AR annotations for physical documents.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502026

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; **Smartphones**.

KEYWORDS

Annotation; augmented reality; document interaction; smartphone; paper; text; layout structure

ACM Reference Format:
Jing Qian, Qi Sun, Curtis Wigington, Han L. Han, Tong Sun, Jennifer Healey, James Tompkin, and Jeff Huang. 2022. Dually Noted: Layout-Aware Annotations with Smartphone Augmented Reality . In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3502026

1 INTRODUCTION

Annotating documents is a form of sensemaking that helps us develop mental models [2]. Once a document has been annotated with comments and insights from multiple parties in collaboration, sharing these annotations can serve as a platform for feedback and discussion [17, 36, 49, 51, 53, 59, 61]. Although digital annotation is easily shared, printed media is still prevalent and often preferred due to the tactile sensation and physical navigation experience [62, 65]. However, sharing annotations on printed documents is inconvenient, especially during collaborative scenarios and with updates to and from digital documents.

Augmented reality (AR) is a promising solution that offers advantages from both the physical and virtual domains by rendering dynamic digital content over printed documents [41, 42, 64, 65]. Annotations shared through AR have greater social presence than pure text [52]; enhance discussions of off-line materials [68]; and facilitate knowledge acquisition [77]. AR further enables sharing annotations on a printed document and accumulating collective knowledge when direct marking is impossible (e.g., a library book or public poster) or impractical (e.g., a textbook that needs to be sold later). Currently, users benefiting from AR annotation through hardware like headsets [41], projectors [75], desktop apparatus [25], or fabricated digital paper [38, 54] that can be expensive or non-portable [26, 74, 75], and thus are of limited device accessibility, convenience, and portability for everyday use.

Smartphone AR is now widely accessible. Its relative small size and portability is crucial to the increasing need to transfer information between the physical and digital worlds in social and portable context. Smartphone AR enables physical-digital sharing, such as updating printed documents with new digital information [24] beyond printed content and pen-written annotations. Smartphone AR can also exploit spatial context, such as displaying digital information as situated visualizations [13], and allowing proxemic interactions [39] via a magic-lens metaphor for natural and embodied interaction experience [14]. The potential summarized above motivates us to build a working, portable system that bridges the gap between multi-user digital and physical annotation. This allowed us to further explore the benefits and challenges of smartphone AR annotation interaction (e.g., efficacy and usability). These challenges are likely because the smartphone's form factor demands high precision and responsiveness due to the small screen and compact document layouts, especially for word-level annotations.

To identify challenges and benefits from users' accounts, we adopted Buchenau's experience prototyping approach [16]. For benefits, we found that smartphone AR adds digital convenience and portability to the physical reading experience and enables users to seamlessly share annotations. For challenges, the main difficulties are accurately highlighting text and viewing multiple annotations. These challenges were aggravated by the tracking limitations of the physical document and the exacerbated error from traditional ray casting selection.

To address these challenges, we developed *Dually Noted* to leverage document layout understanding for improved precision and efficiency (Figure 1). Dually Noted automatically identifies the layout structure of the document to determine the class and position of images, tables, headings, paragraphs, sentences, and words. This allowed us improve ray casting selection efficiency, increase robustness to user hand movement errors, lower cognitive workload, and provide annotation view management. Our technique improves interaction between printed and digital documents and advances content-oriented collaborative experiences.

Explicitly, we make the following contributions:

- Insights from eight users through an experience prototyping protocol that elicits potential benefits and challenges with a smartphone AR annotation prototype;
- A novel AR interaction technique that automatically interprets a document's structural elements to significantly improve users' view satisfaction, interaction precision/efficiency, and cognitive workload with real-world annotation tasks; and
- A prototype system that enables smartphone users to seamlessly create, view, and synchronize digital annotations insitu on a printed document to its digital copy.

2 RELATED WORK

2.1 Benefits of Augmenting Printed Documents

The value of augmenting physical documents with digital content has been well established. Previous work has found that digital augmentation enables new interaction possibilities, including hyperlinking dynamic content, searching, copying text, and annotating [9, 41, 65, 74]. Digital augmentation can also reduce the reader's cognitive load on consumption [22], enhance workflows [41], improve engagement [44], improve learning efficacy [23, 70], and enhance collaboration [46]. A wide spectrum of techniques to enable augmentation include projecting the digital content [26, 32, 34, 35, 58, 65, 74, 75], using headworn devices [41, 68] and user-centric transparent displays [10, 30, 31], or printing circuits and thermochromic inks to augment paper documents (interactive paper) [38, 54, 71]. However, these solutions are often highly specialized and hinder portability; they are either limited to a small user population (headworn devices) or require additional devices or a surface to setup (projections). Dually Noted instead brings the advantages of these solutions to everyday users using their smartphone devices and exploits document layout structure to facilitate AR annotation in real-time.

2.1.1 Handheld AR For Document Consumption. Handheld electronic devices such as smartphones bring lightweight and portable experiences to AR document interactions. Earlier work [43, 56, 67] looked at potential benefits of augmenting document annotation, but lacked the adequate computation power to perform real-time AR tracking. Others explored benefits of overlaying digital information on a handheld device with different applications [11, 14, 24]. For example, the Chameleon system [24] uses a handheld device to show situated digital information and discusses how such a device brings new functionalities to physical documents, such as retrieving detailed information digitally on a printed map or displaying levels of authorized information for different users. Brown et al. explore a magic-lens metaphor to visualize human anatomy when a user poses the handheld devices over a human body [14]. Others have explored the MagicBook interface on interactive book reading applications [11, 29, 37]. Later work found that these interactive books help students to learn positively with lower cognitive load [22] and better spatial visualization [63] and knowledge retention [23, 77]. In exploring augmenting printed documents with digital visualizations, these prior works primarily focused on the browsing experience rather than the interactive annotation experience. A recent work explored AR annotation with a smartphone attached to a bracket [77], allowing users to add "clickable" annotations. The system, however, has limited mobility with neither inline text support nor figure selection, and did not focus on improving the efficacy of the annotation experience. Our work supports real-time inline text and figure selection, as well as the capacity to synchronize printed and digital reading materials, and dives deeper into improving the overall efficacy of the AR annotation experience.

2.2 Synchronizing Annotation from Printed Documents

Digitizing annotations from physical documents ensures easy annotation sharing and preservation. Using a digital pen is a well-established method for seamlessly digitizing printed documents, especially when pens can both mark and scan a document [65]. Guimbretière [27] has presented an infrastructure to digitize, store, and manage physical annotations in a database. The author used a stroke collector to digitize annotations created with a digital pen. Later systems such as Coscribe [65] expanded the availability of digitized annotations to multiple users, and PapierCraft [42] further added "digital functions" such as copying a text paragraph to the digital environment. Recently, Holodoc [41] has combined a digital pen and a head-mounted display (HMD) to project digital data back into pen-based document augmentation systems, establishing a closed-loop ecosystem for synchronizing, displaying, and authoring in both digital and physical environments.

Compared to the pen-based system, our work provides an alternative approach that enables an augmented annotating experience that does not require marking directly on the document. This strategy provides an opportunity for users to annotate documents when a digital pen is not available or when directly marking on documents is not appropriate (e.g., on borrowed books, conference posters, or public documents).

2.3 Layout Structures for Annotation

Prior systems that leverage a document's layout structure focus mainly on purely digital content, with goals such as improving the viewing experience on mobile devices [55] and extracting layout features to improve productivity [3, 72, 73]. One benefit of knowing the layout structure is the subsequent capacity to dynamically change the arrangement of a document. Chang et al. explored four strategies to create greater space for annotations by dynamically modifying the layout of a document, including moving paragraph blocks, overlaying text, or allocating annotations to nearby margin space [18]. SpaceInk [60] creates extra white spaces to support annotation by rearranging a document's text and figures without losing its original content. Similarly, Adobe Acrobat's Reflow functions on mobile devices rearrange the page using the layout structure to fit the content to the mobile device's screen such that users do not need to manually adjust zoom levels to read. While prior work leverages layout structure to support document editing and viewing in the digital environment, our system uses the document's layout structure in the AR context and explores corresponding implications for annotation authoring, anchoring, and viewing.

2.4 Viewing Annotations in AR

Annotations displayed in AR can be difficult for users to read clearly due to their variations in visibility, position, size, and transparency [8]. One common challenge is to avoid overlaps between the AR annotation themselves [5, 48] and the underlying information [8]. Bell et al. [8] used rectangular area features to determine whether an annotation overlapped with another in the image space; similarly, Makita et al. [48] used a probabilistic model to avoid overlaps. Temporal coherence can also affect the experience of viewing annotations in AR because the user is often in motion [69]; other methods such as dynamically rearranging the annotations based on user position and viewing angle have been beneficial to the user's ability to locate annotations [47]. Changing distance in 3D space can also dynamically update the displayed content . For example, projecting different text menus on a user's hand as the hand moves closer to the user allows for natural navigation [76].

On a device with a small screen, like a smartphone, optimizing the annotation viewing experience is more challenging than on larger devices. We use document layout structure to place annotations without collisions, and we also use the smartphone's distance relative to the document to show and hide layers of annotations without content-occluding overlap.

3 EXPERIENCE PROTOTYPING

Our goal is to identify the challenges and potential benefits of smartphone AR annotation with real users. To understand the first-hand experience of using smartphone AR to annotate printed documents, we used Buchenau's experience prototyping [16] protocol with an initial prototype and conducted a formative study. Experience prototyping helps identify usability issues and elements of user experience by presenting users with early prototypes. The core of experience prototyping is engaging users with functional systems to obtain first-hand account data rather than surveyed opinions. This is important in AR interaction systems since AR experiences are difficult to imagine before experiencing them. As the prototype's fidelity impacts user feedback, we iterate our initial prototype to ensure it works as intended before sharing it with users. We do not compare the AR annotation with digital annotation because users are already familiar with digital experience; instead, we interview them for their feedback on digital experience.

3.1 Initial Prototype

We implemented the initial prototype using a common AR selection technique for object annotation, ray casting [20, 57]. Users can swipe on the screen to create half-transparent strokes to fill the background color of a text span or mark figures (i.e., highlighting). Users can also tap on the stroke to key-in comments, which appear virtually as editable, movable sticky notes. Once the highlights or sticky notes are made, they can see the virtual content in AR by panning their smartphone. For tracking, we used Unity's image tracking library and its ARFoundation 4.1 to track the movement and rotation of both the user and the entire printed documents in six degrees of freedom (6DoF). On the iPhone 8 and later devices, the document tracking works at 60 frames per second (FPS).

3.2 Participants and Data Collection

We recruited eight iPhone users (4 male and 4 female, average age = 27, $\sigma = 4$) from a convenience sample. Each participant consented prior to the study and was compensated $15 for their participation. We collected video and screen recordings of the participants and took observation and interview notes.

3.3 Procedure

The study used a think-aloud protocol and lasted one hour over video call. The experimenter held individual video call sessions with each participant (i.e., one on one). Participants received a guided demonstration of the initial prototype and were given five minutes to practice using the system on their own. We told participants that their annotations would be shared automatically with others. Participants were given three printed articles on which to make annotations. Their specific tasks included: 1) highlighting interesting text (i.e., text worth revisiting and sharing), 2) marking unclear sections, and 3) making annotations as they usually would. After the task, we conducted a semi-structured interview with participants, aiming to understand their experiences and focusing on the benefits and challenges of using the initial prototype. The interview also covered how participants read, annotate, and share physical and digital documents in their everyday lives. Each interview was transcribed and analyzed using open coding [21], which allowed us to elicit topic categories that were not predetermined. The initial open codes were extracted from notes taken during interview sessions and recordings. Two authors then independently formed categories around potential benefits and challenges.

3.4 Results

3.4.1 Potential Benefits. **Augmenting Paper Reading.** Participants (6/8) believed a primary benefit of the AR prototype was being able to create and access digital content using a smartphone while primarily engaging with a printed document. Despite needing to hold a device to annotate, participants described the experience as "natural" (3/8) and easy to seek both AR and original text information (2/8). They felt the prototype enabled them to digitally see others' thoughts (3/8) while reading printed books, which was not possible for them before. Three participants indicated that they would feel more engaged and inclined to ask questions if they could see their peers' notes on a printed textbook as they were reading.

Sharing Annotations on Printed Documents. All participants habitually scan images, take photos, and share physical annotations on printed documents in their day-to-day lives. While their usual practices are sufficient for one-time sharing (4/8), many claimed they encountered problems when attempting to search for or revisit their archived items later on. They immediately identified the usefulness of the prototype for record management; P3 said: "*[AR] annotation is like a time capsule . . . I could surprise myself [when] I see . . . old annotations I left on a book years ago. Even if the book is lost, I can get a new one, and my words . . . won't be lost*". Some (3/8) participants used cloud drives to store their annotations (i.e., via scans), but noted that this method does not support new additions being made to annotations, unlike the AR prototype which can be continually amended (2/8). In addition, several (3/8) participants mentioned that their reading habits are multimodal. For example, half of the participants (4/8) interacted with and read the same article in both printed form and on their tablet, depending on factors such as mobility (2/8), device availability (3/8), or convenience (4/8); they preferred to have their AR annotations automatically sync to a digital copy to reduce task resumption overhead when transferring annotations from printed to digital documents.

Non-intrusive Annotations. Most participants (6/8) particularly desired the ability to annotate without damaging a book because the book may be precious (3/8), borrowed (3/8), public (6/8), or simply a form of preference (2/8). P7 said: "*I mostly annotate on my iPad, but this makes me also want to annotate on books now, because I can have my annotations . . . without writing on [my] books*". Unlike with writing directly on a printed document, many participants (5/8) said that they could easily edit annotations and "*squeeze annotations into small spaces*" without margin constrains.

Enabling Mobility. Half of the participants highlighted that the convenience and portability of a smartphone would enable them to annotation on the go. P7 said: "*"I don't want to carry books around, and if I have two [of the same] books in different cities, I can add annotations [in one place], and the next day [when] I travel to another city, I can still get my annotations from that book. All I need is my smartphone, which I always bring with me.*".

Reading Long Documents. Participants (5/8) said that digital reading (e.g., PDF reading) on a smartphone is easy, but that it is not an adequate substitute for paper reading, especially for lengthy documents. Participants ascribed its unsuitability to the device's small screen size (3/8), the increased strain on their eyes (4/8), and the negative effects on their posture (2/8). However, most participants (5/8) suggested that the initial prototype could act as an aide when reading long documents since the paper reading experience is preserved.

3.4.2 Challenges and Concerns. **Selecting Text is Difficult.** Most participants (7/8) complained that text was difficult to select because it is too small relative to the size of their thumb. This caused highlights to frequently misalign with the intended content. Some

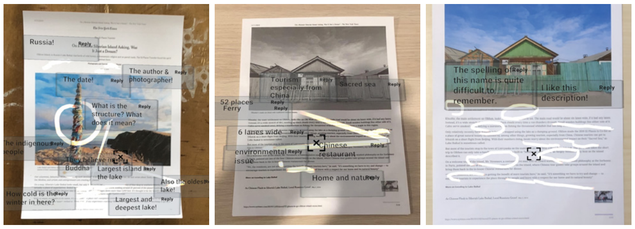

Russia!
The date!
The author & photographer!
What is the structure? What does it mean?
Buddha
Largest island in the lake
How cold is the winter in here?
Largest and deepest lake!
52 places Ferry
6 lanes wide
environmental issue
Chinese restaurant
Home and nature
The spelling of this name is quite difficult to remember.
I like this description!

Figure 2: During experience prototyping, participants struggled with annotations as they filled the view space. Although participants could move annotations out of the printed document to avoid occlusion, most still chose to place annotation near the text for better organization and context. The rightmost figure shows that how ray casting can make it challenge to align a highlight with the intended underlying text.

participants even gave up highlighting the desired text and only added comments. Others worked around this issue by moving their phones closer to the document to increase the size of the text, but this had the secondary drawback of limiting the amount of selectable text.

Cluttered View. As participants added more comments, the AR space became cluttered and hard to navigate (Figure 2). Many participants (7/8) moved their annotations next to the relevant text or figure to maintain priority (1/8), make room to highlight (5/8), and create a contextual link to the targeted text or figure (3/8). All participants expressed that pop-up annotations should not block their view of the AR document and suggested to "filter for what is most relevant" or "not displaying everything at once".

Annotation Drifting. Three participants noticed that pop-up comments began to drift when the smartphone was too close to the edge of the document, as the initial prototype relied on all four corners of the document to be visible as anchor points. This occurred in two cases when participants had difficulty selecting the text and wanted to move multiple comments to reduce clutter. Participants had to awkwardly "zoom out" to include the whole document again.

Arm Fatigue. Some participants experienced arm fatigue as a result of continuously holding the smartphone over the document. We observed that the most common reason participants held the phone for long periods was to highlight text; this process is often tedious due to "fat-finger" issues [33] and mapping on-screen sketches to the document in AR view. Three participants mentioned that they were able to reduce their fatigue by putting the phone down and only picking it up when they needed it to annotate. Surprisingly, participants indicated this did not interrupt their reading; P3 said: "*It is very natural for me to put down the phone when I have nothing to annotate and am focusing on the document.*".

4 DESIGN GOALS

Informed by the findings from the previous section, we aimed to design a system that improves speed and accuracy in smartphone AR selection, as well as elevates overall viewing experience. Because greater selection accuracy mitigates struggle during annotation, it may also reduce fatigue. Likewise, annotations must be organized and optimized for AR viewing. We want our smartphone AR system to liberate users in a similar manner to SurfaceFleet [15], using a server-client model to support cloud-based annotation and asynchronous interactions. The system should furthermore maintain ease of use and allow for as-you-go additions. Other useful features, suggested by participants, included synchronizing annotations across a digital environment and having the option to view the annotations of others.

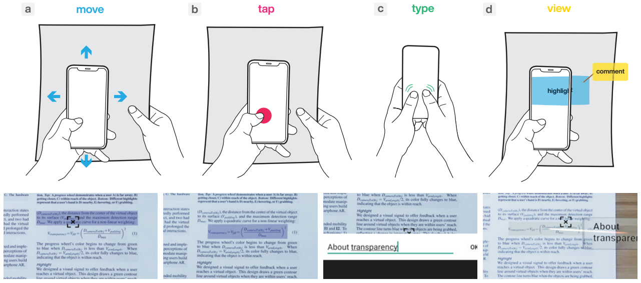


Figure 3: Dually Noted interface. *Upper row:* User interactions. *Lower row:* What a user sees on the screen. a) Hover to pre-select; b) tap to select; c) type the comment; d) the annotation floats to the nearest-side of the document with a line indicating its anchor.

Although mechanisms such as locking [39] or live camera freezing improve accuracy, they reduce real-time engagement for users [6] and require users to find the AR scene after unfreezing the view [40]. We aimed to design an interaction technique that provides a continuous AR experience (e.g., akin to that of a headset environment) while improving annotation efficacy in the real-time, in-situ coupling experience.

Based on the findings of experience prototyping, we hypothesize that an effective AR annotation system should support:

- **Accurate Selection:** Smartphone AR presents challenges to highlight on text. The final experience should facilitate easy and accurate selection of text, figures, and other elements on the document.
- **Compact and Accessible Comments Viewing**: Pop-up annotations in AR should not block the user's view of the text, but should remain accessible when the user wishes to view them.
- **Reduce Drifting Near the Document's Boundary:** Annotations should be as stable as possible. Pop-up annotations drift when they are anchored outside the bounds of the printed document and when the user zooms in too close to the printed document.
- **Automatic Physical-Digital Synchronization:** Annotations should be automatically synchronized across digital copies for easy sharing among multiple devices and users. Changes to digital files should be reflected in AR. This lets the smartphone AR annotation tool bring digital reading, saving, and sharing experience to printed documents.

5 DUALLY NOTED SYSTEM

Based on our design goals and in order to build a system that works well with the smartphone's small screen, we explored the idea of using a document's layout structure data to facilitate both selection and viewing experience, see Figure 3. Here the *layout structural data is defined* as a document's structural data, which consists of individual words, sentences, paragraphs, figures, and tables. The main reason for using layout structure is increasing selection error tolerance without losing meaningful resolution for selecting text or figures, which were common needs identified in experience prototyping. Additionally, knowing the layout allows the system to automatically arrange annotations in a structured manner; for example, multiple annotations on a single paragraph can display in its vicinity and anchor to the same segment of text, avoiding ambiguity and providing direct visual guidance.

Here, we refine our experience prototyping apparatus and present a system pipeline for synchronization, as well as an interaction technique that uses document layout structure for AR authoring and viewing with printed documents.

5.1 System Pipeline

In reflection of the design goals, the system pipeline describes the components required to enable **automatic physical-digital synchronization** across devices and to reproduce this work. The pipeline leverages the cloud service to offload computation heavy tasks and file management from the smartphone AR and let the device deliver fluid experience. The system uses a server-client structure across three main components: *Cloud Service*, *AR Client*, and *Digital Client*. Collectively, the pipeline supports real-time document tracking, to obtain the document's layout structure for annotation interaction, to share the digital media, and to allow dynamic content to flow between AR and desktop users. (Figure 4).

5.1.1 *Cloud Service.*

The main role of the cloud service is processing document snapshots, image conversion and mapping, and communicating with application programming interface (API) servers for OCR and layout structure. The service also stores snapshots, annotations, and user replies that can be dynamically fetched by the AR and digital clients. To do this, a Python PDF library is used to load and save annotations from a PDF file and synchronize changes over the network.

5.1.2 *AR Client.*

The smartphone AR client lets users scan a document and send its snapshot to the cloud service to receive the document's structural data and fetch related annotations. Communication is performed over HTTP using JSON strings that contains the tags, 2D locations, sizes, and contents of the structural data (e.g., text-run, paragraph, figure, and table). The AR client allows users to select printed document text by words, phrases or paragraphs, and select figures and tables. In AR, virtual annotations and highlights appear to be superimposed over the document. Any changes made by the user are automatically saved to the digital copy of the document (i.e., PDF).

The document structure data can take up to 2MB per letter-sized page, resulting in delays in a real-time experience. We therefore separate the data receiving process into threads to reduce the network waiting overhead, obtaining an independent 20-40KB data segment containing information for the paragraphs being presently displayed to the user. As a result, users can start viewing or selecting annotations on paragraphs or figures while the system asynchronously loads the remaining data.

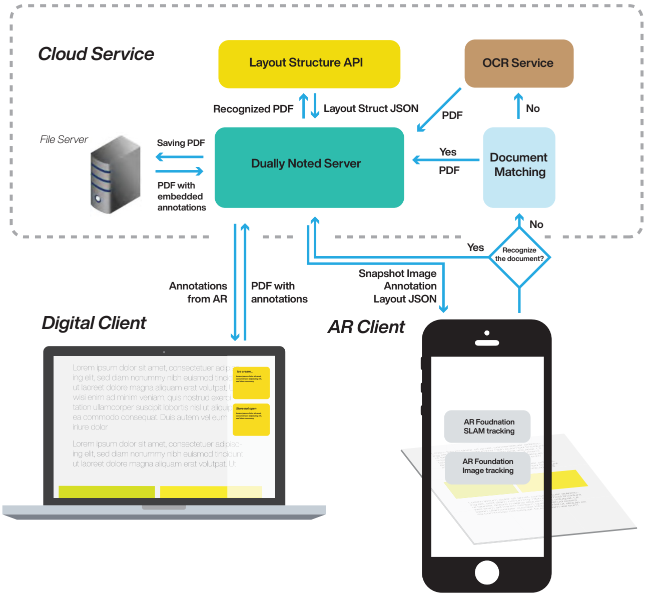


Figure 4: System Diagram. Dually Noted uses a cloud service to support document recognition and layout structure extraction. When using smartphone AR to annotate, the document's layout data is processed and sent to the smartphone to aid interaction. Digital client and smartphone AR client annotations and comments are distributed to other clients via the server.

5.1.3 *Digital Client.*

The digital client is mainly designed for desktop users to create, edit, and update annotations through the cloud service. It uses a PyMuPDF Python library 1.16.2 to read and save annotations from PDF files, and synchronize them with the cloud service. For the purposes of the present research, we did not implement security or access-level controls, though these controls could be added to a production version.

5.2 Recognizing Layout Structure from a Printed Document

While the system pipeline provides the infrastructure to support our design goals, layout structure is the key to enabling content-aware selection and optimizing the viewing experience. To obtain the layout structure, Dually Noted must generate or link to a digital copy of a printed document (e.g., a PDF). To generate a PDF from a printed document, we use a two-step process after the cloud service receives the snapshot of the document: 1) matching, and 2) extraction. In the matching step, we compare an image snapshot of the document to a database of image snapshots on the cloud service. Identifying a match allows us to skip the PDF generation process and reduce the overall processing time. We used the SIFT algorithm and k-nearest neighbors to compare snapshots [45]. Snapshots from smartphone cameras are robust for SIFT comparison even when the smartphone's camera has much lower resolution than is available today [7, 19]. This method provides good invariance with different image orientations and snapshot perspectives;

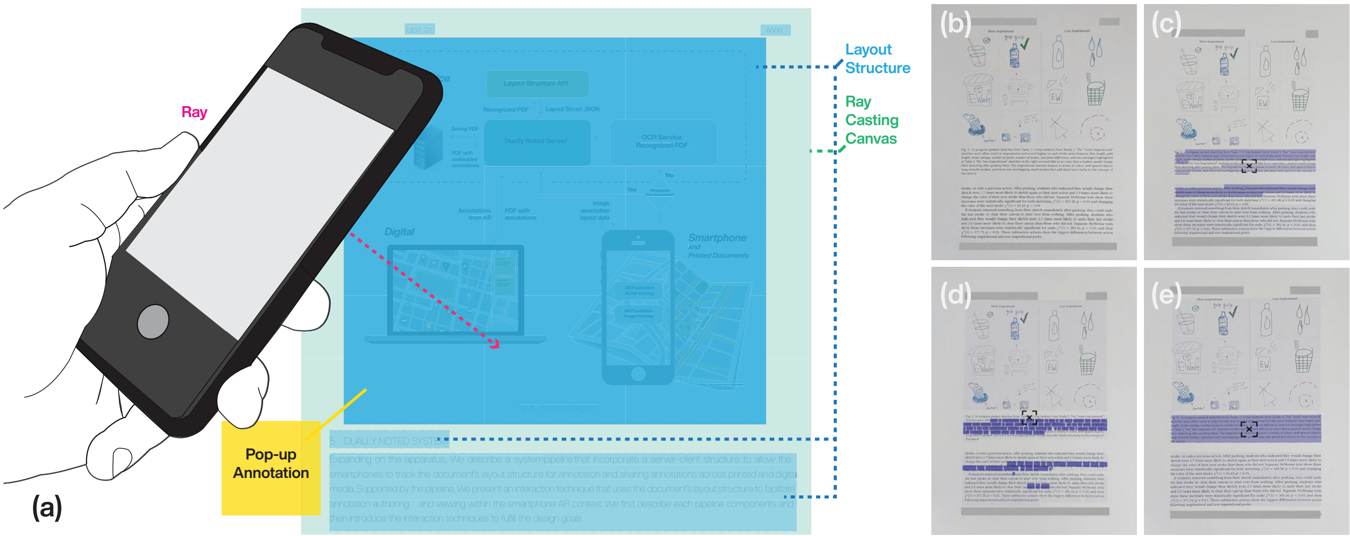


Figure 5: (a) shows how the virtual layers overlay the printed document to work together with ray casting. Layout structure (in blue) overlays the printed text's, figure's or table's boundary to capture users screen taps. Ray casting canvas (in green) is the same size to the document and can detect *continuous* ray casting; this is used to support screen-based strokes and detect when rays are not in any of the layout structure layers. Sub figures: (b) is a printed document without layout structure; (c) each region spans over a sentence; (d) each region spans over phrases or words; (e) each region spans over one paragraph.

however, it cannot cope well with folded or wrinkled documents; we consider these cases outside the scope of the present research. Additional challenges emerge with comparison and recognition across a large number of documents, including identifying variants of prior documents. Addressing these challenges is not the primary goal of this paper; however, using algorithms such as Locally Likely Arrangement Hashing (LLAH) [66] or leveraging existing software that can handle image matching across millions of documents (e.g., Vuforia Cloud Recognition [1]) could help in addressing those issues in future.

If no match is found, we digitize a new PDF file from the snapshot using Adobe's Acrobat OCR service (Figure 4) and add this snapshot to the image database. For this research prototype, the image database does not contain pre-existing images, but is dynamically expanded from snapshots created from the AR client. However, in reality such a database could be generated by publishing houses and various other contributors; a database could be generated from a house's catalogue of electronic books (i.e., a book's PDF), for instance. Otherwise, the database could be pre-converted from digital copies and uploaded by users, linked to commercially available image databases, or, in cases where the document has not yet had a digital copy generated, captured using the Dually Noted AR client. We envision the database growing over time as different users contribute to it; the methods for managing such a database and optimizing its performance are not entailed in the present work.

Most digital document formats, including PDF, do not actually contain structurally meaningful layout. To parse the layout, the server uses Adobe's Layout Structure Extraction model (PDF Extract API [2]) in the extraction step. We test 100 different snapshots to measure the average time to extract the layout. These snapshots were taken from different magazines, news articles, and book covers. On average the API returns the layout structure data in 6.1 (σ = 1.3) seconds for first-time extraction. The result is cached in the smartphone and in the server, so the latency for the subsequent visit to the same page is negligible. One case report indicates that the PDF Extract API maintained over 90% accuracy while digitizing over 50000 questions from paper documents [1]. This API provides similar functions to PubLayNet [78]; both detect the structural regions comprising the locations of words, paragraphs, figures, and tables. Then, word-group regions are formed into sentence regions based on conventional terminal punctuation (e.g., a period or question mark). Each structural region's location and size are indicated by the (x, y, w, h) coordinates of its 2D bounding box.

5.3 Mapping 2D Layout Structure to AR

Since the extracted layout structure is two dimensional, we need to transform its 2D location, orientation, and scale into three dimensions in AR. Given a digital document's width (W) and height (H), we let u and v be the normalized x, y positions in the digital file ($u = \frac{x}{W}, v = \frac{y}{H}$; Figure 6). Next, we find a virtual plane with horizontal and vertical axes matching those of the digital document in AR space. We use the AR camera's position as the zero-rotation origin and map the digital document's horizontal axis to the Right vector (**U**), vertical axis to negative the Up vector (**V**). Finally, we use the tracked document's translation matrix (**T**), rotation matrix (**R**), and physical width S_w and height S_h to transform **U** and **V**, and

[1] https://library.vuforia.com/articles/Training/Cloud-Recognition-Guide.html

[2] https://www.adobe.io/apis/documentcloud/dcsdk/pdf-extract.html

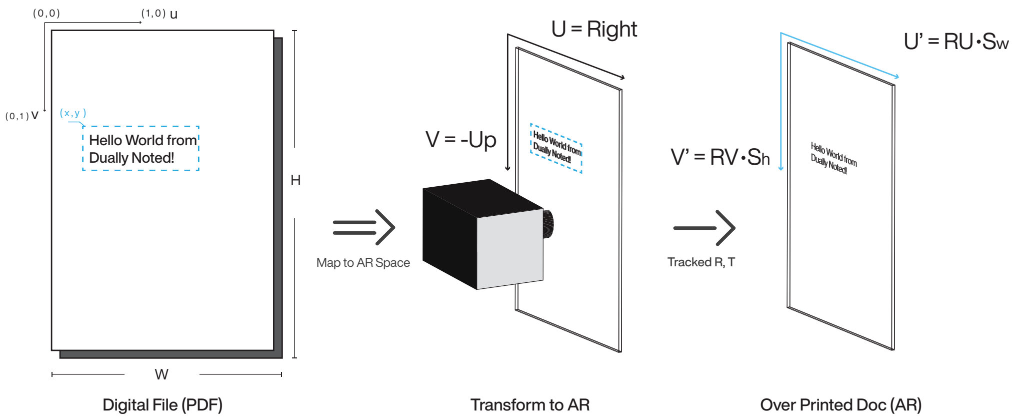


Figure 6: Structural regions are transformed from 2D digital files into 3D coordinates.

so obtain the final transformation matrix to map the 2D bounding boxes:

$$P_{AR} = \begin{bmatrix}\mathbf{R} & | & \mathbf{T}\end{bmatrix} \begin{bmatrix} S_w \cdot \mathbf{U} & S_h \cdot \mathbf{V} & 1\end{bmatrix} \begin{bmatrix}u & v & 1\end{bmatrix}^T .$$

The resulting transformed 2D coordinate system (P_{AR}) represent an area superimposed on their corresponding text, figures, or tables on the printed document when viewing through the AR device.

5.4 Configuring Layout Structure for Interactions

We used three different configurations to implement selection tolerance for interaction: at *word and phrase*, *sentence*, and *paragraph* configurations. In those configurations, we call the bounding boxes that wrap around text or figures structural regions. In the *word and phrase* configuration, each structural region spans one word and users can select any number of words at a time. In the *sentence* configuration, the structural regions of individual words are connected, recalculated to form continues regions that cover the entire sentence. Finally, in the *paragraph* configuration, structural regions cover a paragraph, a figure, or a table (Figure 5). As a result, users can tap "roughly" on the region to successfully highlight.

5.5 Selection Text or Figures with Layout Structure

With structural regions defined and anchored to the printed document, four types of interactions are implemented to fulfill the design goal of **accurate selection**. For a better and less distracting viewing experience, these regions are transparent by default.

- **Pre-selection or Hovering:** The user *moves* the phone (cursor) to aim at text, a figure, or a table. The screen's center position continually casts a ray from the AR camera's forward direction into the scene. A 3D structural region is *highlighted* when the ray hits it and clears that highlight when the ray exits it.
- **Selection:** The user *taps* once on the screen while preselecting. Different configurations will trigger corresponding structural regions that may be selected simultaneously. For example, an entire paragraph will be highlighted when the user taps on any sub-elements (e.g., text or subfigures) in the *paragraph* configuration. Once tapped, the invisible structural region becomes visible, and users are given an opportunity to type in comments.
- **AR Dragging Start:** The user *taps-and-holds* on the screen and moves the smartphone.
- **AR Dragging Stop:** The user *releases* the tap.

5.6 Annotation Viewing in AR

Three-dimensional structural regions also help to achieve **compact and accessible comments**. Anchoring annotations in the printed document space while supporting temporal coherence improves users' ability to locate annotations [47]. Additionally, dynamically arranging the annotations improves user viewing experiences [69], as does ensuring that document content and other annotations are not occluded [8]. Inspired by prior work and feedback from the experience prototyping (Figure 3), Dually Noted automatically anchors annotations to the closest empty space in the document and arranges them to minimize occlusion. Further, to save viewing space, annotations display a short preview by default and only display their full content when users interact with them (Figure 7, a and b). Annotations also automatically shrink when they are out of the AR camera's range.

5.6.1 Layered Visualization for Replies. When dealing with multiple replies on one annotation, instead of expanding the vertical or horizontal AR space, Dually Noted uses the 3D space along the z-axis to virtually stack replies, similar to the interaction design proposed by Wilson and Benko [76]. This minimizes the space needed to display replies. By moving the smartphone closer to and further from the document—similar to using a magnifying glass to see contents on a paper—a user can navigate through all annotations without those annotations occluding each other (Figure 7). This design leverages the smartphone AR's intrinsic move-to-view interaction to avoid displaying content unnecessarily.

5.6.2 Reduce Annotation Drift. One challenge from the design goals (Section 4) indicated that AR annotations drifted if placed beyond the bounds of the printed document. Reducing the drift

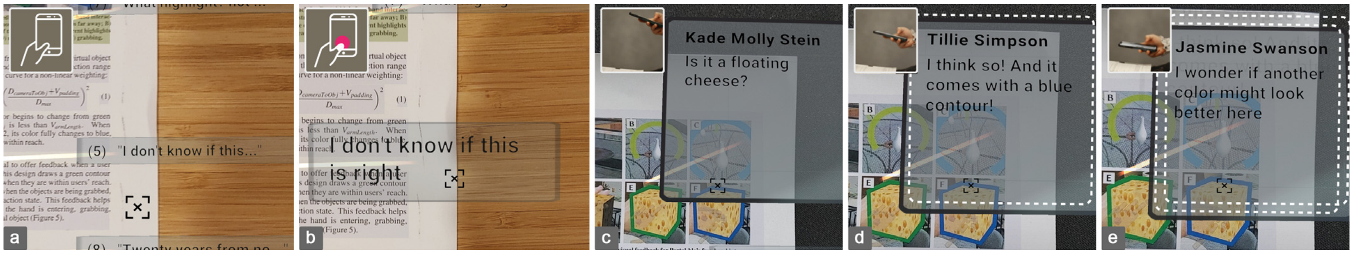

a
(5) "I don't know if this..."
b
I don't know if this is right
c
Kade Molly Stein
Is it a floating cheese?
d
Tillie Simpson
I think so! And it comes with a blue contour!
e
Jasmine Swanson
I wonder if another color might look better here

Figure 7: Our viewing technique. a) shows the pop-up comment in a tab format; b) shows the full view; c)–e) show that when the user moves closer to the page, different layers of annotation reply shows up.

of these annotations expands the interaction region and grants users access to a larger usable space. Similar to Vuforia’s extended tracking[3], our system aims to support annotation tracking beyond the document’s boundary. While spaces beyond the bounds of the document may not be trackable via the image tracking library, they can be tracked with the SLAM tracking used for AR localization. Additionally, these two types of tracking have the same degree of freedom: AR image tracking determines the document in 6DoF (denoted as transformation T_D), while AR SLAM tracking determines smartphone’s localization in the physical space with 6DoF (denoted as transformation T_S). As such, automatically switching between document tracking and SLAM tracking (Figure 8) can reduce drift errors. The final transformation (T_A) applied to the AR annotations is determined by:

$$T_A = \begin{cases} T_D, & \text{if } Tracking\ State = True \\ T_S, & \text{if } Tracking\ State = False \end{cases} \tag{1}$$

The resulting AR annotations fall back to relative positions on the smartphone when the document loses tracking. Additionally, after piloting for selecting different captions and text with layout structure, we noticed that layout structure-based selection reduces the user’s need to magnify the text to highlight it, mitigating the tracking issues caused by document’s being only partially detectable to the image tracking library. However, in cases of very small text, such as is sometimes seen in footnotes, better document tracking is required. Finally, we have applied a Kalman filter for annotation x, y, z positions to reduce random jitter.

6 EVALUATION

The goal of our evaluation was to understand the efficacy of Dually Noted’s interaction and viewing techniques, as well as how those techniques evoke other everyday applications for participants. Specifically, we asked two research questions: RQ1) How would Dually Noted affect users’ timed-performance, accuracy, and cognitive-load while annotating printed documents in AR? and RQ2) How feasible would our proposed interactions for everyday AR annotation? The evaluation consisted of two sessions: a controlled experiment (**Task I**) for objective performance measurement and an open-ended exploration (**Task II**) for assessing real-world usability.

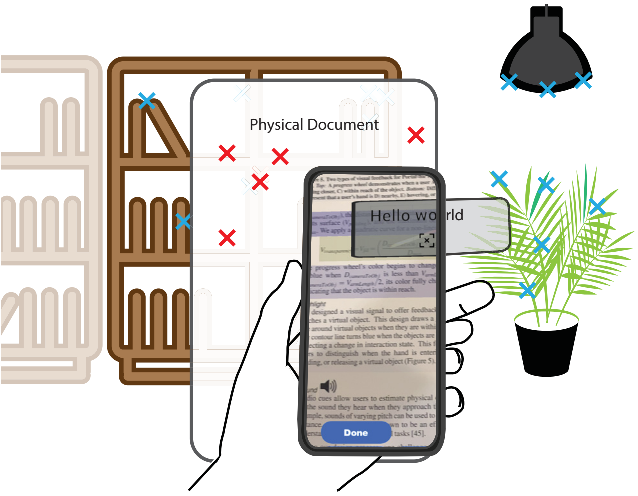


Figure 8: When the document is fully in view, it will be tracked by the image tracking algorithm (marked with a red cross). When users view annotations on the edge of the document, the system switches to SLAM tracking (marked with a blue cross)

6.1 Task I: Controlled Experiment

Participants annotate a letter-sized, journal style paper using a smartphone. They highlight an indicated subtask and are free to use the smartphone keyboard to add comments. Participants are instructed to perform the highlight as fast and accurate as possible. For fair comparisons among all conditions, we exclude the time of text entry whose performance is orthogonal to our system goals while correlated to the comment length.

Experimental Design. Task I uses a $[2 \times 5]$ within-subject design: two conditions (Dually Noted and a baseline) and five subtasks (annotating words, phrases, sentences, paragraphs, and figures). Each subtask is performed three times. These tasks reflect typical annotation activity on physical documents [50]. The **baseline** is the apparatus in Section 3. Both Dually Noted and the baseline use the same document tracking algorithm provided by ARFoundation’s ImageTracking. To reduce learning and order effects, we use a pre-determined ordering to randomized subtasks and balance task performance by alternating conditions.

[3] https://library.vuforia.com/features/environments/device-tracker-overview.html

Data Collection. We collect the following data: 1) efficiency as measured by task performance time, 2) accuracy determined by noticeable user-made errors, and 3) cognitive workload as assessed through a NASA-TLX survey. Qualitative results also describe the observation of participants. We used a script to automatically record the completion time. The logging starts and stops when a participant taps a button on the screen. This button stores two timestamps in the log; one when participants starts to highlight and one when they stop. Accuracy was interpreted based on whether the resulting highlights are sufficient to convey the user's intention to human judgment. We opted to skip automatic ways to score accuracy because users naturally annotate in different ways (e.g., drawing a circle, bracketing a paragraph, etc). As a result, two authors separately grade the accuracy with a $\{0, 0.5, 1\}$ rating scale: 0 for not reflecting the task goal; 0.5 for comprehensible with obvious mistakes; 1 for comprehensible without obvious mistakes (Figure 9b). A final score is generated after any scoring discrepancies were resolved via discussion. A NASA's official TLX application is used to measure the raw scores on a 0 to 100 with 21 gradations in six subscales. In total, we recorded 360 trials (30 trials per participant × 12 participants). Two trials were discarded due to the system overheating.

6.2 Task II: Open-ended Exploration

Experimental Design. Task II is an open-ended session that collects qualitative and holistic views from users within real-life settings. Participants interact with a letter-size printed document using Dually Noted for at least three minutes. The document contains 15 pre-existing digital annotations threaded with at least three replies. Participants are given a chance to view the annotations and reply to them or add their own. We ask participants to think-aloud while performing the task. At the end of Task II, a semi-structured interview collects their experience for viewing and annotating, as well as their thoughts on usability and potential daily applications.

Data Collection. We assess the viewing experience via a questionnaire that asked: 1) if participants could see all annotations clearly; and 2) participants' self-rated ease-of-use and satisfaction on a 7-point Likert scale. We use a semi-structured interview to asks questions about their overall experience, the feasibility of Dually Noted for everyday use, their thoughts on potential applications, and anything else they wished to contribute.

6.3 Participants

We recruited 12 participants (7 male, 5 female, average age = 30, σ = 4) with convenience sampling. Each participant received a $30 gift card as compensation after completing the study. Eight participants reported that they annotate both physical and digital documents in their daily lives, while four participants indicated they annotated digital files only. No participant had prior experience with AR annotations, nor were they aware of the research or research hypothesis. We submitted the study protocol to the Institutional Review Boards (IRB), and it was determined that the study was a program evaluation and did not constitute a human subjects study. But we still applied the principles of informed consent and followed relevant protocols.

6.4 Apparatus

An iPhone 8 or later version is required to run Dually Noted with steady document tracking at 60 FPS. We sent participants digital files to print before the study. To remotely deploy the system on the participants' devices, we used the TestFairy [4] platform with installation instructions.

6.5 Overall Procedure

The experiment runs in a remote setting over Zoom due to pandemic restrictions. Each participant first receives and completes a formal consent and follow up by the experimenter's instructions. Participants enable screen sharing on both their computers and their smartphone for our remote observation. They then practice in both Dually Noted and the baseline condition for a maximum of five minutes until they are comfortable continuing. During the practice session, participants interact with system annotation functions, (e.g., moving the phone toward and away from their printed document to view annotative comments).

In Task I, the experimenter assigns the participant a condition and a subtask. After the experimenter indicates which line is being targeted (e.g., first line in the second paragraph), the participant locates that segment of the text. Then, they tap on the screen to start timing and create the highlight, before tapping the stop button when the task is completed. After, the system prompts them with a text input option. Once they complete all five subtasks (15 trials total), they are asked to rate their cognitive ratings via an official NASA-TLX application. Afterwards, participants complete another set of five subtasks in the alternate condition, followed by another NASA-TLX rating.

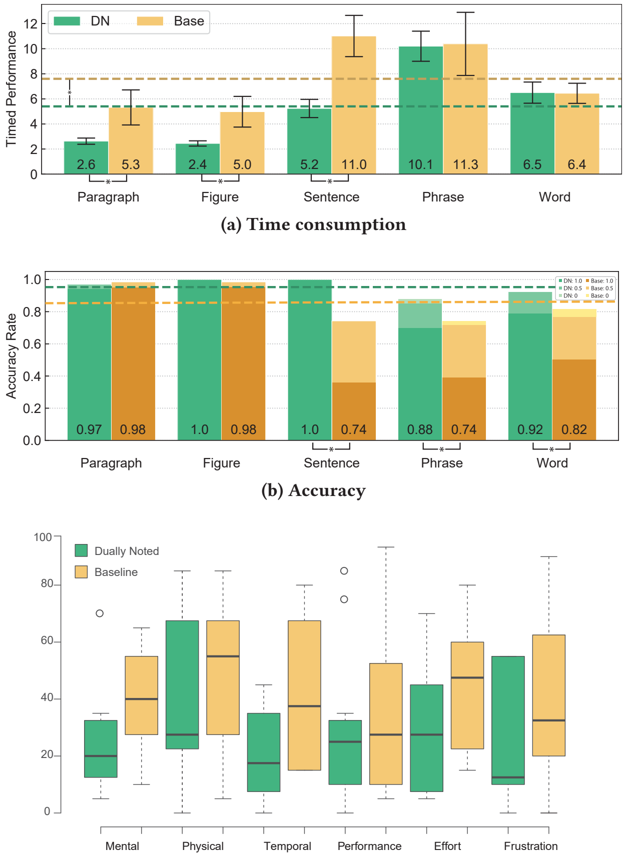
In Task II, participants are given a printed document with pre-existing AR annotations and are told that their annotations would be saved to later appear for others. We instruct the participants to read pre-existing annotations and annotate or reply with the AR system as they might normally. During that process, their think-aloud observations are recorded and the experimenter took notes on their interactions. At the end of the study, we collect participants' qualitative feedback from the interview and analyze screen recordings. The entire experiment took about one hour (σ = 20).

7 RESULTS

7.1 Quantitative Evaluation

Completion Time. We first log-transformed the completion time and tested the significance with a repeated two-way ANOVA. We checked the sphericity assumptions and adjusted p-value with Bonferroni correction for 5 subtasks. The results revealed participants' significantly faster (42%) task performance in Dually Noted (DN) compared to the baseline ($F(1, 11) = 84.84, p < 0.01$). We also found a significant interaction effect between the condition and subtasks (F (4,128) = $16.76, p < 0.01$) with a large effect size ($\eta^2 = 0.34$). A paired-samples t-test revealed that the DN condition is significantly faster than the baseline in selecting paragraphs ($t(32)= -8.85, p < 0.01$), figures ($t(32)= -9.32, p < 0.01$), and sentences ($t(32) = -5.05, p < 0.01$). Figure 9a shows the breakdown details for timed performance.

[4] https://www.testfairy.com/



(a) Time consumption

(b) Accuracy

(c) TLX reports

Figure 9: Summary of the results of evaluation. (a)/(b): comparisons time/accuracy metrics. The dashed lines indicate average values and the error bars indicate standard deviation. (c): Dually Noted had lower mental load than the baseline across five of the six NASA-TLX measures (lower is better).

Accuracy. We found a significant improvement in accuracy for DN over the baseline condition ($U = 16414, z = 4.733, p < 0.01, r = 0.26$). The average selection accuracy was 95% for DN and 85% for the baseline. Phrase selection has the lowest accuracy of 88% for DN and 75% for the baseline. DN has highest accuracy scores on paragraph (97%), figure(100%) and sentence (100%) selections; the baseline has highest accuracy scores on paragraph (98%) and figure (98%) selection. See figure 9b for details. Mann-Whitney U tests indicated that the DN condition was significantly more accurate than the baseline in sentences ($U = 825, z = 4.75, p < 0.01, r = 0.58$), phrases ($U = 689, z = 2.21, p < 0.05, r = 0.27$), and word ($U = 666, z = 1.98, p < 0.05, r = 0.24$).

Cognitive Load. Following Hart [28], we calculated both the total and subscales of the raw NASA-TLX scores. A Mann-Whitney U test showed that the overall cognitive load is significantly lower for DN compared to the baseline ($U = 1773.5, z = -3.3, p < 0.01, r = 0.3$). The interaction between the conditions and TLX subscales revealed DN significantly reduced temporal ($U = 36, z = -2.1, p < 0.05, r = 0.4$) and mental demand ($U = 36.5, z = -2.1, p < 0.05, r = 0.4$).

Overall Preference and Viewing. The exit interview and survey revealed that 83% participants preferred DN over the baseline for its being intuitive and satisfactory. All participants (12/12) reported that they could read annotations clearly and did not experience cluttered view. Most of them (10/12) were satisfied with the viewing experience and found it easy to navigate.

7.2 Qualitative Evaluation

Impressions of Interactive Layout Structure. Half of the participants made a sound of surprise or remarked "cool" when first seeing paragraphs, sentences, or words automatically pre-highlighted. Each participant immediately understood the design concept of both the baseline and DN in the training session. Although the conditions are counter-balanced in an alternating order, participants (P1, P5, P7, P10) still noted during their think-aloud process that the layout structure simplified the selection process and allowed them to select more quickly. Three participants (P6, P7, P9) reported that the baseline selection was harder to use than DN. P9 reacted saying, "that is much easier than the previous one," and P6 commented on the baseline,"I think this needs to be more stable." In general, participants (P1, P3, P7, P8, P9) felt it was fun and "cool" to select the text and images in AR with the smartphone.

Behavioral Insights. We observed difference behavioral patterns when participants used AR drag (described in Section 5.5). For example, P2 and P3 quickly dragged the AR pointer to the target without paying attention to text in between, whereas P9 dragged the AR pointer to highlight every single word, despite achieving the same outcome as P2 and P3. With the baseline, constant adjustment of the smartphone's perspective was the most frequently noted behavior. Most of these adjustments were performed to find the right angle from which to annotate (P3 and P9) or to see the printed text more clearly.

Difficulty of Reading for Extended Periods on the Smartphone. When moving the smartphone closer to the document to view layered annotations, all participants reported that reading short comments was easy and undemanding. However, reading for an extended period in a fixed position was reported as tedious by some. P10 and P12 mentioned that reading shorter comments was convenient and accessible but that achieving the same with longer comments would require a different format. We observed that those who held their arms in mid-air to read reported fatigue more often than those who rested their elbow on the table or held the smartphone with two hands. This fatigue especially presented itself in Task I where participants were annotating continually; annotations would typically be made in a more intermittent fashion in real-world scenarios. As for the feasibility of highlighting or annotating, four participants (P3, P5, P6, P7) stated that neither DN nor the baseline were suitable for long-term, continuous annotation sessions due to fatigue. They also, however, indicated that multiple short-term annotation sessions were realistic with DN but not with the baseline.

Additionally, participants had diverse opinions about the hand movement speed that should be mapped to viewing layered annotations. For instance, P1, P3 and P6 preferred that the amount of annotation visible should increase when their hand moved, but P9 thought otherwise.

Smartphone AR Annotations in Everyday Scenarios. All participants but two found making annotations based on the document layout structure feasible for their everyday use with printed documents; they cited the system's accuracy, speed of use and the smartphone stand-alone setup as reasons. Of the two who did not find it feasible for their everyday use, one participant did not typically read printed documents, and the other explained that his hands were not steady enough. Participants liked Dually Noted's portability (P4, P5, P6, P7), accuracy for annotating (P6, P7, P9, P10, P12), easy-to-view annotations (all participants), and ability to support online-offline discussion (P3, P4, P9, P10, P11, P12).

8 DISCUSSION

8.1 Operation Efficacy

For *RQ*1, significantly improved performance time, accuracy, and overall cognitive load indicates that Dually Noted overcome the **accurate selection** challenge identified in Section 3. Faster performance reduces arm fatigue (P5, P6, P7), improves usability, and leads participants to believe it would be easier than the baseline to use in real life. Figure 9a shows that significantly faster paragraph, figure, and sentence selections are key contributors in a 42% faster performance time , but selection speed does not improve for phrases or words (even though selection accuracy improves). One explanation, taken from observation, is that variations in participants' habits and preferences lead to challenges with fine movements, such as AR dragging and aiming (Section 5.5). The current design maps the aiming interaction movement in a linear ratio to the AR camera's movement. Future work could explore non-linear and individually-tailored mapping to improve performance related to fine movements.

The baseline condition reaches the same near-perfect tracking accuracy for paragraphs and figures as Dually Noted, though it requires significantly more time. One explanation is that participants use habitual methods for annotate (assuming there is no ambiguity in intention per our scoring); we observe participants drawing a circle, a rectangle, or a star to select large areas (e.g., a paragraph or figure). These approaches do not require exact alignment with the underlying content and, as a result, do not apply to highlighting shorter words or phrases where there is less tolerance for ambiguity. In these cases, Dually Noted's greater accuracy can help users to pinpoint selections without extra performance time.

8.2 Viewing 3D Annotations

Participants report that all annotations could be read clearly, which indicates Dually Noted successfully reduce clustering problems related to viewing. The layout structure automatically places annotations beyond the bounds of the page, while maintaining their semantic connections. None of the participants experienced content drifting while viewing annotations, indicating the SLAM and image tracking achieved the intended benefit of stabilization Section 5.6.2.

Participants easily understand and navigate layered replies. However, we observe that some participants use both hands for this interaction to gain better movement control and stability. There are probably no universal settings for how hand motion can be mapped to layered viewing, as the participants (P1, P3, P6, P9) have different inclinations and methods for maneuvering their phones. This leaves open the possibility of future work in studying hand motions for viewing layered content in AR.

8.3 Extended Reading Using Dually Noted

Reading multi-page documents while making smartphone AR annotations is challenging over a long period of time. Unlike using AR annotation with head-mounted displays, users hold their smartphones with one hand for in situ information retrieval. While the current implementation makes it easy to view AR annotations on single-page documents, annotations on the smartphone are not yet responsive enough to load in real-time while flipping through pages due to the latency needed to parse the layout structure and the fatigue from holding the phone while waiting. However, people usually flip through books and documents in sequential page order, so it may be possible to extend the current system to pre-fetch the upcoming digital content by predicting the next pages the user will read. This would allow users to quickly spot the AR annotations with negligible processing time for recognition.

Although we opted for a real-time AR annotation experience tailored for short reading sessions with this system prototype, there may be trade-offs between a live AR experience and traditional 2D digital information that can achieve a compromise for extended sessions. For example, retrieving the annotation in 2D format [43] or temporarily freezing the live view [6, 12] during interaction may provide a more familiar user experience. These methods allow users to read the annotations more comfortably for long periods but spoil the AR experience or ability to display spatially-aware AR content (e.g., annotations with AR animations).

When reading a printed document, flipping through pages is well supported by Dually Noted. However, Dually Noted does not provide the serendipity in spotting annotations as the pen-written annotations do. A user must hold the phone to discover the digital annotations while waiting for digital information updates.

8.4 Fatigue in Authoring AR Annotations

Although some participants report arm fatigue during Task I, they suggest that the fatigue is unlikely to occur in real-world as they would typically annotate more intermittently (P5, P6, P7, P9, P11). Further, Dually Noted users spent about 5.3 seconds per annotation and could reach annotation speeds as fast as 2.5 seconds per annotation, selecting an entire paragraph/figure with over 97% accuracy. P5 commented that "I think it is totally doable to take out my phone, aim at a paragraph, and add a comment [with DN]."

8.5 Layout Structure for Other AR Devices

Although Dually Noted is deployed and tested on smartphones, layout structure selection could also be used on a tablet; both devices support on-screen ray casting. The tablet may facilitate more extensive engagement with annotations due to its larger screen size. We envision that Dually Noted's interactions could be deployed on different devices that support ray casting or pointing, including HMDs, as most HMDs support handheld controllers or eye gaze input, which resembles a ray cast selection. HMD users could leverage Dually Noted to facilitate text and figure interaction with documents without a device needing to be held, preventing arm fatigue.

8.6 Practicality and System Implications

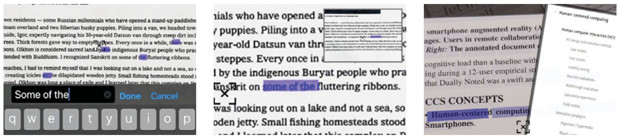
For *RQ*2, most participants (90%) report that Dually Noted is suitable for short, everyday annotation tasks. The implications of their feedback and applications for Dually Noted are discussed below.

Multimodal Annotation for Broader Use Cases. Although this study compares selection aided by layout structure with selection using common ray casting, in practice, a combination of these methods provides useful functionality. Ray casting yields greater freedom (P1, P10, P11, P12) for sketching on screen, while layout structure selection supports an organized and accurate link between annotations and their related content. A balanced, combinatory approach that considers the trade-offs between the two can be designed. For example, future design could allow a user to engage ray casting to create personalized sketches, signs, and notes on a textbook, while relying on the layout structure to annotate the text accurately.

Multiuser Experience. Dually Noted allows multiple users to create and view annotations *asynchronously*. Users can likewise view and respond to annotations from others in-situ (Task II). While Dually Noted provides a portable way for printed document readers to access updated digital information, future work should explore real-time multiuser highlighting for collaborative learning and mixed-reality co-annotating. This includes supporting interaction state synchronizations (knowing what other people are looking at), annotation stylization, and filtering out unwanted annotations.

Usage Scenarios. Participants proposed usage scenarios that can leverage the benefits of layout structure. For example, they suggested the possibility of digitally searching the printed document or looking up content via an external hyperlink. They mentioned that these tasks could be effectively supported with a smartphone's mobility and would allow them the benefit of digital functions when working with printed documents. Additionally, Dually Noted's portability let them leave notes in situ and initiate conversations about physical objects that have printed labels. For example, P3 mentioned that they would like to see how others comment on items such as "menus at a diner or reviews on a product."

Active Editing Documents. Currently, Dually Noted is designed for sharing annotations on static documents, such as digital copies of flyers, books, news articles in PDF form, and printed documents. Its AR annotation uses the meta-information (e.g., whether the region contains a figure, a table, or a text element) and text within a single 2D bounding box region to identify itself. As a result, moving text on the document does not affect the annotation but changing the text content removes the AR annotation. However, users can manually reset the AR annotations on an edited document and treat it as a new document for shared annotation. This allows for applications such as initiating new discussions cycles on a particular magazine design iteration. Accommodating structural changes from iterations or active editing remains work for the future.



(a) Type search phrase (b) Search and highlight (c) Hyperlink in AR

Figure 10: Example applications show how Dually Noted's layout structure enables additional digital functions on printed documents. (a) and (b) : AR search on the printed document allows readers to locate a target phrase quickly. This can be specifically useful for long or wordy documents where searching by eyes is tedious to do. Dually Noted allow smartphone users to type and search on a printed document without additional devices. (c) : Hyperlinking adds additional digital content into the physical reading experience. Leveraging the document's layout structure, this application retrieves external multimedia information in real-time and enabling new interactions such as tapping on a link (or words with links) to open up a pop-up window in AR. In the subfigure c), the user taps on the word "Human-computer" to open its hyperlink displayed as a pop-up canvas on the right.

9 EXAMPLE APPLICATIONS

We implement *searching* and *hyperlinking* applications with the goal of demonstrating and exploring the range of Dually Noted's capabilities. Prior work, such as that of Holodoc [41], explored searching and hyperlinking on printed documents using a headset and digital pen. We aim for an instrument-free experience that does not require external devices, extracting the entire document's layout in the initial AR process and allowing real-time interactions.

9.1 Digital Searching and Linking on Printed Documents

Searching a printed document for a specific snippet of text with just our eyes can be challenging, and our success is affected by document attributes such as font size and spacing [4]. We thus expanded Dually Noted to enable text searching on printed documents via its layout structure. Figure 10b demonstrates how a user inputs the text they wish to search for. Digitally searching the facsimile of the printed document, Dually Noted highlights all occurrences of a phrase or term in AR, allowing rapid target localization. Similarly, Dually Noted enables digital linking (Figure 10c, right). A user can tap on a hyperlink in AR to view external content via their screen. The content displays as a 3D image next to the link, allowing users to rotate their wrist to glance between the 3D image and AR annotations.

10 CONCLUSION

We have presented a smartphone AR system that synchronizes digital annotations with printed documents. It uses the printed document's layout to facilitate in situ annotation selection and viewing via a mobile device. Engaging an experience prototyping protocol, eight participants used an AR prototype with a naive content-agnostic selection technique, generating information on the benefits and challenges of the prototype. The results informed the design and implementation of Dually Noted: a smartphone-based AR annotation system that recognizes the layout of printed documents for effective real-time authoring and viewing of annotations. In a controlled experiment that compared Dually Noted to an AR prototype, our system enabled users to spend 42% less time performing common annotation tasks. It significantly reduced errors in selecting individual words or phrases and had a lower cognitive rating. Overall, 83% of participants found Dually Noted's viewing experience intuitive and satisfactory, and 90% of participants would want to use the system for short-term interactions with augmented annotations on printed documents. We consider layout-structure-aided interaction a new and important step toward AR annotation in everyday settings, and we envision a future where smartphone users can create, edit, and share their annotations anywhere with a lightweight and portable AR system.

ACKNOWLEDGMENTS

The authors thank reviewers for their constructive feedback that helps us improve this paper. We extend our gratitude to all participants who provide valuable lessons to make this work possible, and to Fumeng Yang for discussions and editing. This work is supported by the National Science Foundation under Grant No. IIS-1552663 and IIS-2107409.

REFERENCES

[1] [n.d.]. https://business.adobe.com/customer-success-stories/cambridge-assessment-case-study.html

[2] Annette Adler, Anuj Gujar, Beverly L Harrison, Kenton O'hara, and Abigail Sellen. 1998. A diary study of work-related reading: design implications for digital reading devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 241–248.

[3] Hend S Al-Khalifa and Jessica Rubart. 2008. Automatic document-level semantic metadata annotation using folksonomies and domain ontologies. *ACM SIGWEB Newsletter* 2008, Autumn (2008), 1–3.

[4] Nilsu Atilgan, Ying-Zi Xiong, and Gordon E Legge. 2020. Reconciling print-size and display-size constraints on reading. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30276–30284.

[5] Ronald Azuma and Chris Furmanski. 2003. Evaluating label placement for augmented reality view management. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* IEEE, 66–75.

[6] Huidong Bai, Gun A Lee, and Mark Billinghurst. 2012. Freeze view touch and finger gesture based interaction methods for handheld augmented reality interfaces. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*. 126–131.

[7] Dominikus Baur, Sebastian Boring, and Steven Feiner. 2012. Virtual projection: exploring optical projection as a metaphor for multi-device interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1693–1702.

[8] Blaine Bell, Steven Feiner, and Tobias Höllerer. 2001. View management for virtual and augmented reality. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. 101–110.

[9] Kavita Bhardwaj, Santanu Chaudhury, and Sumantra Dutta Roy. 2013. Augmented paper system: A framework for user's personalized workspace. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*. IEEE, 1–4.

[10] Andrea Bianchi, So-Ryang Ban, and Ian Oakley. 2015. Designing a physical aid to support active reading on tablets. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 699–708.

[11] Mark Billinghurst, Hirokazu Kato, and Ivan Poupyrev. 2001. The magicbook-moving seamlessly between reality and virtuality. *IEEE Computer Graphics and applications* 21, 3 (2001), 6–8.

[12] Sebastian Boring, Dominikus Baur, Andreas Butz, Sean Gustafson, and Patrick Baudisch. 2010. Touch projector: mobile interaction through video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2287–2296.

[13] Nathalie Bressa, Henrik Korsgaard, Aurélien Tabard, Steven Houben, and Jo Vermeulen. 2021. What's the Situation with Situated Visualization? A Survey and Perspectives on Situatedness. *IEEE Transactions on Visualization and Computer Graphics* (2021).

[14] Leonard D Brown and Hong Hua. 2006. Magic lenses for augmented virtual environments. *IEEE Computer Graphics and Applications* 26, 4 (2006), 64–73.

[15] Frederik Brudy, David Ledo, Michel Pahud, Nathalie Henry Riche, Christian Holz, Anand Waghmare, Hemant Bhaskar Surale, Marcus Peinado, Xiaokuan Zhang, Shannon Joyner, et al. 2020. SurfaceFleet: Exploring Distributed Interactions Unbounded from Device, Application, User, and Time. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 7–21.

[16] Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. 424–433.

[17] Jonathan J Cadiz, Anop Gupta, and Jonathan Grudin. 2000. Using Web annotations for asynchronous collaboration around documents. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 309–318.

[18] Bay-Wei Chang, Jock D Mackinlay, Polle T Zellweger, and Takeo Igarashi. 1998. A negotiation architecture for fluid documents. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*. 123–132.

[19] Tsung-Hsiang Chang and Yang Li. 2011. Deep shot: a framework for migrating tasks across devices using mobile phone cameras. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2163–2172.

[20] Yun Suk Chang, Benjamin Nuernberger, Bo Luan, and Tobias Höllerer. 2017. Evaluating gesture-based augmented reality annotation. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 182–185.

[21] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[22] Kun-Hung Cheng. 2017. Reading an augmented reality book: An exploration of learners' cognitive load, motivation, and attitudes. *Australasian Journal of Educational Technology* 33, 4 (2017), 53–69.

[23] Andreas Dünser, Lawrence Walker, Heather Horner, and Daniel Bentall. 2012. Creating interactive physics education books with augmented reality. In *Proceedings of the 24th Australian computer-human interaction conference*. 107–114.

[24] George W Fitzmaurice. 1993. Situated information spaces and spatially aware palmtop computers. *Commun. ACM* 36, 7 (1993), 39–49.

[25] Jan Grbac, Matjaž Kljun, Klen Čopič Pucihar, and Leo Gombač. 2016. Collaborative Annotation Sharing in Physical and Digital Worlds. In *IFIP International Conference on Human Choice and Computers*. Springer, 303–313.

[26] Jonathan Grudin. 2001. Integrating paper and digital information on EnhancedDesk: a method for realtime finger tracking on an augmented desk system. *ACM Transactions on Computer-Human Interaction (TOCHI)* 8, 4 (2001), 307–322.

[27] François Guimbretière. 2003. Paper augmented digital documents. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*. 51–60.

[28] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[29] Wahyu Nur Hidayat, Muhammad Irsyadul Ibad, Mahera Nur Sofiana, Muhammad Iqbal Aulia, Tri Atmadji Sutikno, and Rokhimatul Wakhidah. 2020. Magic Book with Augmented Reality Technology for Introducing Rare Animal. In *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*. IEEE, 355–360.

[30] Juan David Hincapié-Ramos, Sophie Roscher, Wolfgang Büschel, Ulrike Kister, Raimund Dachselt, and Pourang Irani. 2014. cAR: Contact augmented reality with transparent-display mobile devices. In *Proceedings of The International Symposium on Pervasive Displays*. 80–85.

[31] Juan David Hincapié-Ramos, Sophie Roscher, Wolfgang Büschel, Ulrike Kister, Raimund Dachselt, and Pourang Irani. 2014. tPad: designing transparent-display mobile interactions. In *Proceedings of the 2014 conference on Designing interactive systems*. 161–170.

[32] David Holman, Roel Vertegaal, Mark Altosaar, Nikolaus Troje, and Derek Johns. 2005. Paper windows: interaction techniques for digital paper. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 591–599.

[33] Christian Holz and Patrick Baudisch. 2010. The generalized perceived input point model and how to double touch accuracy by extracting fingerprints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 581–590.

[34] Jochen Huber, Jürgen Steimle, Chunyuan Liao, Qiong Liu, and Max Mühlhäuser. 2012. LightBeam: interacting with augmented real-world objects in pico projections. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*. 1–10.

[35] Daisuke Iwai and Kosuke Sato. 2011. Document search support by making physical documents transparent in projection-based mixed reality. *Virtual reality* 15, 2-3 (2011), 147–160.

[36] Nancy Kaplan and Yoram Chisik. 2005. Reading alone together: creating sociable digital library books. In *Proceedings of the 2005 conference on Interaction design and children*. 88–94.

[37] Tereza Gonçalves Kirner, Fernanda Maria Villela Reis, and Claudio Kirner. 2012. Development of an interactive book with augmented reality for teaching and learning geometric shapes. In *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. IEEE, 1–6.

[38] Konstantin Klamka and Raimund Dachselt. 2017. IllumiPaper: Illuminated interactive paper. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5605–5618.

[39] David Ledo, Saul Greenberg, Nicolai Marquardt, and Sebastian Boring. 2015. Proxemic-aware controls: Designing remote controls for ubiquitous computing ecologies. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 187–198.

[40] Gun A Lee, Ungyeon Yang, Yongwan Kim, Dongsik Jo, Ki-Hong Kim, Jae Ha Kim, and Jin Sung Choi. 2009. Freeze-Set-Go interaction method for handheld mobile augmented reality environments. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*. 143–146.

[41] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. 2019. HoloDoc: Enabling mixed reality workspaces that harness physical and digital content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[42] Chunyuan Liao, François Guimbretière, Ken Hinckley, and Jim Hollan. 2008. Papiercraft: A gesture-based command system for interactive paper. *ACM Transactions on Computer-Human Interaction (TOCHI)* 14, 4 (2008), 1–27.

[43] Chunyuan Liao, Qiong Liu, Bee Liew, and Lynn Wilcox. 2010. Pacer: fine-grained interactive paper via camera-touch hybrid gestures on a cell phone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2441–2450.

[44] Robb Lindgren, Michael Tscholl, Shuai Wang, and Emily Johnson. 2016. Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education* 95 (2016), 174–187.

[45] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[46] Chris Lytridis, Avgoustos Tsinakos, and Ioannis Kazanidis. 2018. ARTutor—an augmented reality platform for interactive distance learning. *Education Sciences* 8, 1 (2018), 6.

[47] Jacob Boesen Madsen, Markus Tatzqern, Claus B Madsen, Dieter Schmalstieg, and Denis Kalkofen. 2016. Temporal coherence strategies for augmented reality labeling. *IEEE transactions on visualization and computer graphics* 22, 4 (2016), 1415–1423.

[48] Koji Makita, Masayuki Kanbara, and Naokazu Yokoya. 2009. View management of annotations for wearable augmented reality. In *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 982–985.

[49] Catherine C Marshall. 1997. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries*. 131–140.

[50] Catherine C Marshall and AJ Bernheim Brush. 2004. Exploring the relationship between personal and public annotations. In *Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries*. 349–357.

[51] Motoki Miura, Susumu Kunifuji, Buntarou Shizuki, and Jiro Tanaka. 2005. AirTransNote: augmented classrooms with digital pen devices and RFID tags. In *IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'05)*. IEEE, 56–58.

[52] Alaeddin Nassani, Hyungon Kim, Gun Lee, Mark Billinghurst, Tobias Langlotz, and Robert W. Lindeman. 2016. Augmented Reality Annotation for Social Video Sharing. In *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications* (Macau) *(SA '16)*. Association for Computing Machinery, New York, NY, USA, Article 9, 5 pages. https://doi.org/10.1145/2999508.2999529

[53] Kenton O'hara and Abigail Sellen. 1997. A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 335–342.

[54] Simon Olberding, Michael Wessely, and Jürgen Steimle. 2014. PrintScreen: fabricating highly customizable thin-film touch-displays. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 281–290.

[55] Saurabh Panjwani, Abhinav Uppal, and Edward Cutrell. 2011. Script-agnostic reflow of text in document images. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. 299–302.

[56] Klen Čopič Pucihar and Paul Coulton. 2014. [Poster] Utilizing contact-view as an augmented reality authoring method for printed document annotation. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 299–300.

[57] Gerhard Reitmayr, Ethan Eade, and Tom W Drummond. 2007. Semi-automatic annotations in unknown environments. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE, 67–70.

[58] Jun Rekimoto and Yuji Ayatsuka. 2000. CyberCode: designing augmented reality environments with visual tags. In *Proceedings of DARE 2000 on Designing augmented reality environments*. 1–10.

[59] Evan F Risko, Tom Foulsham, Shane Dawson, and Alan Kingstone. 2012. The collaborative lecture annotation system (CLAS): A new TOOL for distributed learning. *IEEE Transactions on Learning Technologies* 6, 1 (2012), 4–13.

[60] Hugo Romat, Emmanuel Pietriga, Nathalie Henry-Riche, Ken Hinckley, and Caroline Appert. 2019. SpaceInk: Making Space for In-Context Annotations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 871–882.

[61] Emily Louise Forrester Schneider. 2016. *Designing Digital Tools for Critical Reading: Multiple Perspectives on a Platform for Social Annotation.* Ph.D. Dissertation. Stanford University.

[62] Abigail J Sellen and Richard HR Harper. 2003. *The myth of the paperless office.* MIT press, Cambridge, MA, USA.

[63] Brett E Shelton and Nicholas R Hedley. 2004. Exploring a cognitive basis for learning spatial relationships with augmented reality. *Technology, Instruction, Cognition and Learning* 1, 4 (2004), 323.

[64] Hyunyoung Song, Tovi Grossman, George Fitzmaurice, François Guimbretiere, Azam Khan, Ramtin Attar, and Gordon Kurtenbach. 2009. PenLight: combining a mobile projector and a digital pen for dynamic visual overlay. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 143–152.

[65] Jurgen Steimle, Oliver Brdiczka, and Max Muhlhauser. 2009. CoScribe: integrating paper and digital documents for collaborative knowledge work. *IEEE Transactions on Learning Technologies* 2, 3 (2009), 174–188.

[66] Kazutaka Takeda, Koichi Kise, and Masakazu Iwamura. 2011. Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah. In *2011 International Conference on Document Analysis and Recognition*. IEEE, 1054–1058.

[67] Kazuma Tanaka, Motoi Iwata, Kai Kunze, Masakazu Iwamura, and Koishi Kise. 2013. Share me-a digital annotation sharing service for paper documents with multiple clients support. In *2013 2nd IAPR Asian Conference on Pattern Recognition*. IEEE, 779–782.

[68] Katsuma Tanaka, Kai Kunze, Motoi Iwata, and Koichi Kise. 2014. Memory specs: an annotation system on Google Glass using document image retrieval. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 267–270.

[69] Markus Tatzgern, Denis Kalkofen, Raphael Grasset, and Dieter Schmalstieg. 2014. Hedgehog labeling: View management techniques for external labels in 3D space. In *2014 IEEE Virtual Reality (VR)*. IEEE, 27–32.

[70] Azfar Bin Tomi and Dayang Rohaya Awang Rambli. 2013. An interactive mobile augmented reality magical playbook: Learning number with the thirsty crow. *Procedia computer science* 25 (2013), 123–130.

[71] Takahiro Tsujii, Naoya Koizumi, and Takeshi Naemura. 2014. Inkantatory paper: dynamically color-changing prints with multiple functional inks. In *Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology*. 39–40.

[72] Suppawong Tuarob, Line C Pouchard, and C Lee Giles. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. 239–248.

[73] Suppawong Tuarob, Line C Pouchard, Prasenjit Mitra, and C Lee Giles. 2015. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries* 16, 2 (2015), 111–128.

[74] Pierre Wellner. 1993. Interacting with paper on the DigitalDesk. *Commun. ACM* 36, 7 (1993), 87–96.

[75] Andrew D Wilson. 2005. PlayAnywhere: a compact interactive tabletop projection-vision system. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*. 83–92.

[76] Andrew D Wilson and Hrvoje Benko. 2010. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 273–282.

[77] YanXiang Zhang, Li Tao, Yaping Lu, and Ying Li. 2019. Design of Paper Book Oriented Augmented Reality Collaborative Annotation System for Science Education. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 417–421.

[78] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1015–1022.