NoteLink: A Point-and-Shoot Linking Interface between Students' Handwritten Notebooks and Instructional Videos

Ranjitha Jaddigadde Srinivasa
ranjitha.jsrinivasa@gmail.com
Department of
Electrical & Computer Engineering
University of British Columbia

Samuel Dodson
smdodson@buffalo.edu
Department of Information Science
University at Buffalo

Kyoungwon Seo
kwseo@seoultech.ac.kr
Department of
Applied Artificial Intelligence
Seoul National University of Science
and Technology

Dongwook Yoon
yoon@cs.ubc.ca
Department of Computer Science
University of British Columbia

Sidney Fels
ssfels@ece.ubc.ca
Department of
Electrical & Computer Engineering
University of British Columbia

2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) | 978-1-6654-1770-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/JCDL52503.2021.00026

ABSTRACT

When learning from instructional videos, students frequently take handwritten notes to improve recall and comprehension. When reviewing their notes, it can be difficult to return to the corresponding part of the video. In this paper, we present NoteLink, a mobile application that allows students to take pictures of their notes to re-find and play relevant videos on their smartphone or tablet. Our study followed four phases. In Phase I, we identified the characteristics of students' notes by analyzing 10 engineering students' handwritten notes taken while watching instructional videos. We found: 1) students' notes are comprised of four content types: text, formula, drawing, and a hybrid of two or more types, 2) at least 75% of the notes, regardless of content type, manifest some degree of verbatim overlap with the corresponding video content, and 3) videos are referenced at three scales of temporal granularity: point, interval, and whole video. In Phase II, we designed a prototype mobile application, NoteLink, that retrieves instructional videos that are similar to students' notes. In Phase III, we ran a usability study with 12 engineering students to evaluate their preferences for the temporal granularity of retrieved videos and how search results are displayed. Students reported a preference for matches at the interval temporal granularity. Interviews with participants suggest that NoteLink-like tools for re-finding instructional videos are useful. In Phase IV, we evaluated the retrieval accuracy of NoteLink using the data collected in Phase I. The overall accuracy was 78%, and 98% for textual notes. We also provide design recommendations for optimizing NoteLink.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Users and interactive retrieval*; Search interfaces; • **Applied computing** → Hypertext / hypermedia creation.

KEYWORDS

Handwritten Recognition, Re-Finding, Personal Information Management, Video-Based Learning, Note-Taking

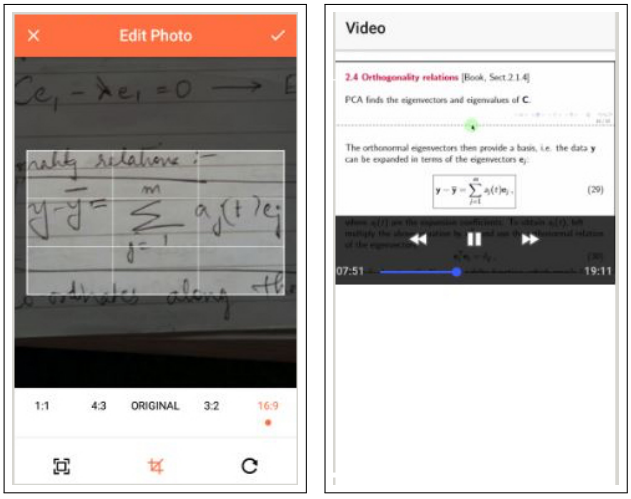

Figure 1: NoteLink, a mobile interface, enables a student to take a photo of their handwritten note (left) to find the corresponding video with topically similar content. The video is then played back within the app (right).

1 INTRODUCTION

Note-taking by hand can improve recall and comprehension [25, 47]. It also facilitates more flexible writing and drawing than using a keyboard and mouse [37, 43]. Students take and review notes in many learning situations, including during lectures and when preparing for exams. However, learning from notes can be difficult when the content is ambiguous [21, 52]. In these cases, students must navigate from their notes back to the primary learning material.

When taking notes on instructional videos, some students write down timecodes to create a link from their note to the video [9]. Unfortunately, it can still be challenging to match a timestamp to a specific video when students learn from many videos. Furthermore, these links are time-consuming to create and can distract from the learning activity. Previous work has found that students frequently record verbatim notes in courses involving mathematics

978-1-6654-1770-9/21/$31.00 ©2021 IEEE
DOI 10.1109/JCDL52503.2021.00026

and engineering topics [47]. Facts, definitions, and, graphical representations are often recorded with little or no paraphrasing [47]. We speculated that these types of notes could be used as cues to aid the content-based retrieval of related videos.

In this paper, we present a feasibility study of whether paper-based, handwritten notes can be linked to videos. We designed, built, and evaluated NoteLink (Figure 1), a novel mobile application that supports point and shoot linking between notes and videos. Students take a picture of their notes, and NoteLink identifies the content of the note and retries matching videos within a collection of instructional videos. The highlights of our work are presented below as answers to our three research questions:

(1) What are the distinctive attributes of notes taken as students watch instructional videos? We performed a needs assessment in Phase I, by collecting previously taken notes of watched videos through a lab study with 10 engineering students. We analyzed the types of notes students take while learning with video and identified characteristics of the note content that can be used to facilitate video retrieval.
(2) What is the preferred temporal granularity of the retrieved video? In Phase II, we designed and built NoteLink, a medium-fidelity prototype that recognizes the note content within pictures of students' notebooks then retrieves the matching videos. In Phase III, we conducted a lab-based usability study with 12 students to explore their preferences for and against different temporal granularities of retrieved video clips. We also investigate students' preferences for various video metadata in the video results.
(3) Can off-the-shelf handwritten text recognition APIs support these design requirements? In Phase IV, we present the results of our evaluation of the accuracy of NoteLink.

Overall, this paper makes three contributions.

(1) We present NoteLink, a novel video retrieval system that leverages handwritten notes as a query.
(2) We identify four types of handwritten note content (i.e., text, formula, drawing, and hybrid) that can be used as cues for finding relevant instructional videos.
(3) We showcase students' preference for and against three distinct temporal granularities (i.e., point, interval, and whole video) for displaying the videos.

2 RELATED WORK

2.1 Hypermedia Linking

While a growing number of video-based learning systems have been developed with features for textbook-style highlighting [12], note-taking [28, 35], and tagging [10, 14], previous work suggests that video-based learning occurs in coordination with other media types. Students often take notes with paper-based notebooks, even when provided with video annotation tools [20, 53]. Dodson *et al.* found that undergraduates in flipped classrooms interact with an heterogeneous information ecology, composed of learning materials that span text, video, and audio [9], accessed through various platforms, such as video players and learning management systems. A challenge for teachers and learners is the limited interoperability and linking within their information ecologies, resulting in what Jones calls "information archipelagos" [27].

Hypertext often invokes digital environments, particularly the Web; however, hypertext can exist in non-digital environments [29]. For example, Marshall's study of university students' textbook annotations suggests that annotations are conceptually hyperlinks within and between content [29, 34]. When re-finding the source of hypertextual notes, students often make use of navigational cues [45, 55]. With physical information objects, such as textbooks, these cues include properties of the artifact (e.g., its color and size), paratext [17] (e.g., page numbers and headings), and relative positions of the information sought (e.g., "about half-way"). In e-books, some contextual cues are reduced or lost completely, so text-search, highlighting, scrolling, and annotations are commonly used to find information [32, 39]. These types of contextual cues can be less salient in videos.

In video, navigating can be difficult and time-consuming on account of fewer and less rich cues. Current video interfaces leverage visual, auditory, textual and temporal information streams to facilitate navigation and re-finding of video information through time-linked video-based annotation, transcripts, filmstrips, and table of contents [28, 35, 58]. However, re-finding material in video by other means, such as handwritten notes, is an underexplored area of research.

Previous work has explored linking between media; in particular, augmenting paper-based documents within digital information [18]. For example, Yeh *et al.* created an application to augment biologists' paper-based fieldnotes with links to digital photographs of butterflies [57]. Linking paper-based documents and video has also been examined. The ChronoViz system, for example, uses a special dot-marked paper and a camera to integrate paper notes into the composite time-coded data set of video files [13]. Embedded Media Markers can also be used as indicators, and are frequently employed in terms of glyph codes [50], barcodes [40], or transparent marks, signifying the availability of video associated with notes. Temporal referencing appears to be a widespread approach in video platforms for entertainment too. Yarmand *et al.* found YouTube comments reference a variety of temporal aspects of video: from single points, to intervals, to whole videos [56]. This paper builds on the previous work by exploring how novel tools can support interacting with video given students' paper-based note-taking practices.

JCDL has been a site of pioneering work on video-based interactive information retrieval since the early 2000s [e.g., 6, 12, 16, 19, 23, 36]. The TRECVID [2] community has also made significant contributions to this research area. Of particular importance to this paper is semantic indexing, which requires methods for detecting visual, auditory, or multi-modal concepts in videos that are assigned as semantic descriptors of the video. For example, in VCenter [23] Hsiao and Wang segment video into a series of frames from which only the most representative frames are used for indexing. iVIEW [31] is a system that supports full-content searching of multilingual text and audio extracted from the video. Likewise, our approach utilizes the visual, textual, and auditory data available from a video collection to calculate a similarity score between a handwritten note and a video.

2.2 Handwritten Word Recognition

While much progress has been made applying optical character recognition (OCR) to video recordings [1, 22, 44, 54], accurately recognizing handwritten notes continues to be a challenge [42, 48]. A promising area of research on the recognition of handwriting has leveraged stroke data, such as the direction and order letters are drawn [7, 46]. This information is readily captured by mobile phones and tablets, which support digital inking. However, many students continue to use pen and paper for taking notes [20], meaning stroke data is not always available.

Studies of handwriting word recognition have demonstrated substantial improvements in recognition accuracy by employing a combination of convolutional and recurrent neural networks [5, 41, 51]. Deep-learning approaches require a large amount of training data to discern the most important features for character recognition. Collecting and annotating a sufficiently large dataset of handwritten notes in different settings remains expensive and laborious. To our knowledge, no such dataset is publicly available. Therefore, in this work, we use off-the-shelf OCR engines leaving machine learning-based recognition optimization for future work.

3 PHASE I: CHARACTERISTICS OF HANDWRITTEN NOTES

In Phase I, we investigated the diversity of both note content and their corresponding video references. The procedure involved identifying the distinct characteristics of students' notes to inform implementing our hand-writing recognition techniques. We specifically targeted notes taken in technical courses, such as computer science and engineering. We analyzed previously taken notes of instructional videos to identify links between notes and videos. Each link is comprised of two entities: 1) the content of the notes and 2) the link destination within the video, allowing us to compare the students' notes to the correct point in the video. This resulted in a taxonomy of handwritten expressions and linking types. We also developed a dataset of mappings between each note to a video time-point. The dataset was used in our Phase IV study.

3.1 Procedure

We recruited 10 undergraduate students — four from an electromagnetics course, two from a machine learning course, three from a data analysis course, and one from a software engineering course. We analyzed participants' existing handwritten notes that were recorded while they learned from instructional videos as part of their coursework. Before the study, each participant confirmed they had at least eight pages of previously captured handwritten notes.

Participants were asked to locate pages in their notebooks that referenced content from an instructional video. Each instance of a video-related note was recorded to populate our dataset of notebook-video mappings. To record the content of participants' notebooks, we used a flatbed scanner to capture high-quality images and maintain a consistent resolution, luminance, and color quality. The notes collected were full-page copies of students' notebooks. We provided each participant with a piece of paper to conceal any content they were uncomfortable sharing. The notebooks were returned after completing the data collection, which took approximately 45 minutes per participant.

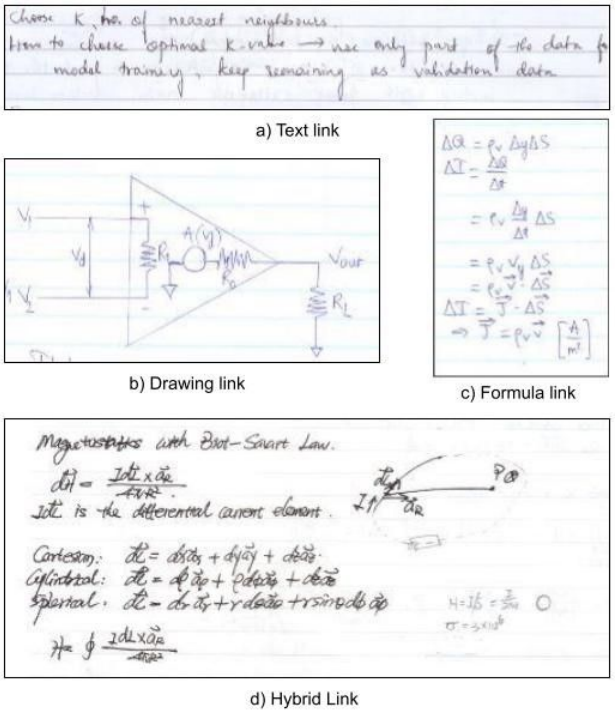

Figure 2: Handwritten links collected during Phase I. From top: a) text, b) drawing, c) formula, and d) hybrid note types

"Ground truthing" was a crucial step for validating our system in the later stages of our project. We collected participant-labeled ground truth observations along with the actual copies of video-related notebook content. Once copies of all relevant pages were taken, participants were asked to mark a rectangular box around all the video-related part of notes. All the participants used a laptop provided by us to mark the boxes and chose either Paint or Microsoft PowerPoint software for editing the annotations. Each annotation included 1) the *name* of the video, 2) the *timestamp* of the referenced video material where the note content can be linked, and 3) a *confidence interval*, measured using the Remember, Know, Guess paradigm [11], of their responses. Participants were asked if they *Remembered* exactly where the note content occurs in the video, *Knew* about the occurrence in the video but cannot remember the exact point of time, or *Guessed* a video based on its semantic relevance to the note content.

3.2 Collected Video Links

We recorded between five to 10 pages of notes per participant, which contained a minimum of 10 to a maximum of 35 video links. In total, 181 notes were collected and analyzed. We categorized the notes into four types: 1) notes with textual content were called *text*, 2) notes with mathematical symbols were called *formula*, 3) notes with drawings, such as circuits, were called *drawing*, and 4) notes comprising a combination of the three types were called *hybrid*. This was based on the examples described in the hierarchical annotation of online handwritten documents [8, 24]. Examples of the four note

types are presented in Figure 2. We determined the content types of the notes, and found that 39.22% of the references ($n = 71$) were written letters or words, 34.25% ($n = 62$) were formulas and 4.97% ($n = 9$) were drawings. We also found 21.54% ($n = 39$) hybrid notes.

Hybrid content type consists of both textual and non-textual components, which poses a challenge in selecting the appropriate recognition technique. However, when analyzing hybrid notes, we found that one content type often dominates the other(s). For example: the proportion of a content type may be far greater than the other type(s); the similarity score between a content type and the video may be significantly greater than the other type(s); or one content type may be emphasized (e.g., circled, highlighted, or starred). This led us to reclassify hybrid content as either text, formula, or drawing based on the dominant content type. In the end, 98 of all notes were tagged as text (54.14%), 65 as formula (35.91%), and 18 as drawing (9.94%).

Students marked most note links with a confidence level of 'Remember'. We identified two note links marked with 'Know' and four with 'Guess'. One participant tagged a part of the note page as video-related content, but did not remember the video and marked the confidence level as 'Unsure'. An important aspect to note is that each student took considerable time to scroll through the videos and the video timeline to record the relevant timestamp. However, the time taken to identify links did not indicate a general pattern. Also, we did not see any consistency in the number of links recorded per participant and their confidence level of recalling the related video context as 'Remember', 'Know', or 'Guess'.

Further, we analyzed *verbatim overlap* (VO), a measure of the word-for-word overlap between the lecture content and the content of the notes. VO is computed as the ratio of number of matching chunk of notes to the total number of chunks in notes. A chunk may represent each word of a text, each line of a formula, or a complete drawing. The boxplot in Figure 3 visualizes the overall distribution of the VO across text, formula, and drawing representations in the watched video note content. The boxplot suggests that the inter quartile range (IQRs) of all the three boxes are above 50% VO. This inference implies that at least half of the note links identified in each type, that is, text, formula, and figure, matched half or more of the content chunk for chunk as found in the videos. We outline some of the observations specific to each type below:

(1) In the case of texts and drawings, at least 25% of the note samples showed a VO of less than 50%. In the case of text, possible interpretations are that either students paraphrased the content in the notes in their own words or, the matching video timestamp as indicated in the ground truth did not exactly match the context of the notes. The IQR for the text type also exhibits a larger variance in the 50–100% window of overlap compared to the other two types. This indicates that with the text type, there is more paraphrasing when compared to formula and drawings.
(2) The median is the lowest for the text type pointing to 75% VO and is at 100% VO for the formula and drawings. Formulas, specifically, manifest 100% VO in the case of at least half of the formula-based notes. This implies that when learning from videos, formulas are transferred literally to notes for reviewing purposes.

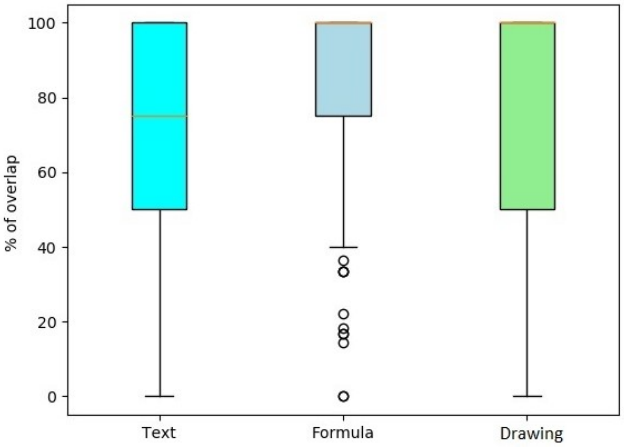

Figure 3: A comparison of content overlap in the text, formula, and drawing representations, respectively.

(3) Furthermore, for both text and drawing, there are cases when there is no overlap at all, with a minimum of no VO. But, it is extreme in the case of formula, indicating some amount of obvious overlap for most cases.

3.3 Conceptualizing Video Timestamps

Investigating how students marked video links to their notes was crucial in returning the accurate video timestamp as predicted. We looked into the timestamp — i.e., the link destination — recorded as the ground truth for each note link from the dataset, and found that the timestamps are 1) not always where the note content exactly occurs in the video and 2) not always the start of a section where the topic of interest is discussed, for example, a student wanted the returned video to play from the middle of a section where the software code implementation is covered. The timestamp of a retrieved video does not indicate whether a student is trying to refind a whole video, an interval of similar content in the video, or a specific point in the video. Thus, we conceptualize the timestamp to be retrieved based on the temporal granularity in the video content.

Previous work has emphasized the importance of temporal context within search results in video [4, 56]. For example, Yarmand *et al.* [56] identified three distinctive temporal scales: 1) a *point* that references a frame in a video, 2) an *interval* that references a span of video frames, and 3) *whole video*. We applied these scales of temporal granularity in our work.

4 PHASE II: DESIGN & IMPLEMENTATION

After finding that students make links to video content in their notebooks using a variety of content forms (text, formulas, and drawings), we set out to design and build a prototype application that 1) recognizes these handwritten content types, 2) matches these to a collection of video, and 3) presents the matching video(s) to users.

4.1 NoteLink: A Point and Shoot Application

Today, smartphones have great potential for delivering just-in-time information [15, 26]. Considering students' use of paper notebooks for their coursework, we designed a mechanism for a mobile device to serve as an interface between handwritten notes and the videos that were watched when the notes were made. The user takes a picture of their notes and the recognition system uses that image as a query to find the corresponding video elements from whence the notes came. The goal is to employ an interface that takes minimal effort from a user to provide the desired input (i.e. point and shoot at their handwritten notes) and delivers video results that convey the information associated with the notes they took.

Figure 1 illustrates the use of NoteLink interface. When a user opens the application, they are asked to either take a picture of their notes or select an existing image from their device's photo library. For either option, the user can crop the image containing the text content to delimit the part of the video-related note content. The video(s) most similar to the image is then presented to the user. The mobile application was built on Android platform using the NativeScript plugins[1] API for using the camera, the background-http plugin for enabling HTTP calls, and the video player plugin that uses the native video players to play remote content. At the back end, the image input is sent to the Recognizer (Section 4.2) and Matcher (Section 4.3) blocks to find the corresponding video from a collection of videos.

4.2 Recognizer Block Implementation

We identified and compared available off-the-shelf OCR technologies to recognize each note type differently. A random note sample from the data collection in Phase I was used to select an OCR API that performs well with our requirements. We evaluated the OCR APIs from PixLab,[2] Google Cloud Vision,[3] and Microsoft.[4] The Microsoft API worked best with rotated text characters, but it did not reliably recognize the special characters, symbols and lines in figures that did not contain text. To extract special symbols, such as mathematical notations, from a given image, we used the Mathpix OCR API.[5] To read lines in pictorial representations, we used a two-step recognition process. First, we extract feature-based key points and descriptors from a scaled and slightly rotated image using the Scale Invariant Feature Transform (SIFT) algorithm [30]. Second, we compare the structural similarity, using Structural Similarity (SSIM) index, between the note image and the filtered video slides from the first step with SIFT. Finally, with the Recognizer block in place, we implemented the Matcher block in order to find similar videos from the collection.

4.3 Matcher Block Implementation

Figure 4a illustrates our approach for selecting videos that are similar to the handwritten text and formula note types. We remove stop words from the text identified by the OCR API, and look for matches in the video transcripts. The region of interests (ROIs) for each video are computed to differentiate mathematical symbols or formulas from text-heavy video frames. We then compare the OCR-ed ROIs for each frame with the handwritten input data to calculate the best matching ROI, which gives the best matching video with the timestamp where the ROI occurs.

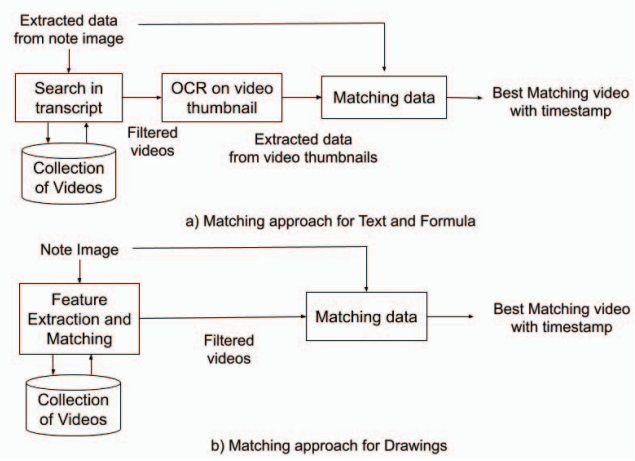

Figure 4: Matching approaches for a) text and formulas and b) drawings.

Drawings are processed differently than text and formulas. A flowchart is provided in Figure 4b. Again, we match feature descriptors from SIFT with input image descriptors. We compare the SSIM of the note images with the list of ROIs from filtered videos to produce the final matching video using the Fast Library for Approximate Nearest Neighbors (FLANN) algorithm [38].

5 PHASE III: EVALUATION

Students' opinion on how to employ a linking device like NoteLink as an educational tool is vital, if the right technology is to be designed, evaluated, and rolled out. We studied the acceptable temporal difference between the retrieved and the expected video along with the inclusion of other video-related objects, to conveniently establish the match. The objective of this phase involved three main steps: 1) to learn about students' preferred *temporal granularity* for retrieved videos, 2) to assess their preference for and against *video metadata* in the search results, and 3) to evaluate the usability and usefulness of NoteLink for video-based learning.

5.1 Procedure

We conducted a lab-based usability study with 12 engineering students who had previous experience learning with video in one or more university courses. Half-hour appointments were scheduled with each participant. We used NoteLink as a medium-fidelity prototype to evaluate the design decisions made in Phases I and II. Each session was audio recorded, which enabled us to focus on participants' verbal prompts throughout the study.

First, we asked participants about which video-based course(s) they have taken, and, in particular, whether or not they have developed any video-related learning and/or note-taking practices, in

[1] https://docs.nativescript.org/plugins/building-plugins
[2] https://pixlab.io/api
[3] https://cloud.google.com/vision/docs/reference/rest/
[4] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text
[5] https://docs.mathpix.com/

order to help us explore more formative design opportunities. Second, we asked participants to watch an instructional video and then take a comprehension quiz. The purpose of this task was to help participants recall note-taking when learning from videos and also generate sufficient real-time notes to guide the subsequent tasks. However, participants were requested to learn the video content so that they can perform well in the quiz to facilitate more ecological validity. Our approach follows Borlund's work on using simulated work tasks in lab studies [3] to obtain as realistic note-taking behavior as possible in a simulated setting. Participants chose one video from three pre-selected instructional videos on engineering topics. Each video was approximately five minutes. We encouraged the participants to take notes using a sheet of paper. Third, we asked participants to identify all notes that referred to content in the video, following a procedure similar to Phase I. Fourth, we gave participants a hands-on demonstration of NoteLink, explaining how NoteLink retrieves videos using handwritten notes. Fifth, participants reported their views for and against retrieving video at three different temporal granularities and presenting search results with different video metadata. They indicated their preferences verbally and through Likert scale questions. Sixth, we asked participants about the usability and usefulness of NoteLink. Again, participants indicated their preferences verbally and through Likert scale questions for effectiveness, efficiency and satisfaction.

5.1.1 Preference for Temporal Granularity. The first objective of this phase was to examine whether or not students preferred one of the temporal granularities for presenting the retrieved videos. We showed participants three counterbalanced designs of displaying a timepoint at the point, interval, and whole video scale. The designs were demonstrated relative to the notes participants' took in the video-learning task. For example, participants noted a video timestamp for each of their notes as *mm:ss*, specifying the minute and second in the video the note refers to. In the case of the point temporal granularity, the video was retrieved at three timepoints, *mm:ss − 0:02*, the timestamp, and *mm:ss + 0:02* in the same frame that pointed to the noted timestamp. In the case of interval, the timestamp was shown as *mm:ss − 0:30*, *mm:ss + 0:30*, and for the whole video, timestamp at the beginning or middle of the video. For each temporal granularity, the participants shared their thoughts with us aloud on what is an acceptable divergence from the timepoint. Additionally, participants indicated their preference on three temporal scale based on the three questions using a Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*): 1) It is easy to find the information I need from this timepoint; 2) The information is useful in helping me complete the tasks and scenarios for learning; and 3) I'm satisfied with the retrieved video timepoint.

5.1.2 Preference for Video Metadata. The second objective of Phase III was to better understand participants' design expectations for displaying the search results with video metadata. To display the search results in a list, we used video metadata with textual and non-textual attributes, following previous work [33]. We showed participants three designs, with four types of video metadata: 1) Title + Thumbnail, 2) Title + Thumbnail + Keywords, 3) Title + Thumbnail + Keywords + Summary (see Figure 5). Participants expressed their preference for the three designs aloud and also

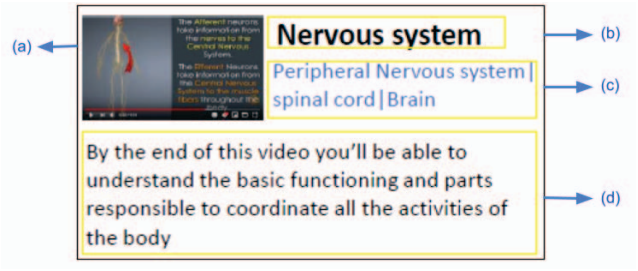

Figure 5: An example search result, comprised of a Thumbnail (a), Title (b), Keywords (c), and Summary (d).

rated their agreement on same three statements listed for temporal granularity.

5.1.3 Usability and Usefulness. Each participant also reported on the usability of NoteLink in terms of efficiency, effectiveness and satisfaction [4]. The three elements' ratings were captured through a set of Likert scale questions from 1 (*strongly disagree*) to 5 (*strongly agree*). Each usability element's items were phrased with two positive and one negative question to avoid bias.

5.2 Findings

5.2.1 Temporal Granularity. We compared participants' preferences among the temporal granularities (i.e., point, interval, or whole video) based on the data from the three Likert scale questions. Since the data collected was ordinal and failed to meet the assumptions for parametric tests, we used the Friedman test. An exact p-value was used to account for the small sample size ($n = 12$). A *post-hoc* analysis was conducted using the Wilcoxon signed-rank tests. A Bonferroni correction was applied, resulting in a significance scale set at $\alpha < 0.017$.

Participants' perceived ease of finding information from the retrieved time point ($\chi^2(2) = 7.4, p = 0.021$) and perceived effectiveness in completing tasks for learning were different across the three temporal granularities ($\chi^2(2) = 6.897, p = 0.029$). However, the *post-hoc* test did not locate significant differences between the three scales. Participants' satisfaction with the retrieved timepoint was significantly different ($\chi^2(2) = 9.190, p = 0.008$). Additionally, the *post-hoc* test manifested significant difference in the interval and whole video temporal granularities ($z = -2.240, p = 0.011$). Participants felt higher satisfaction when the retrieved video timepoint was in the interval when compared to a timepoint in the whole video.

To learn more about participants' preferences for and against specific temporal granularities, we looked for explanatory quotes in the verbal prompts. We coded participants' comments at five different levels of preference: *Very good*, *Good*, *Neutral*, *Bad*, or *Very bad*, (see Figure 6). For example, P13 provided reviews on each of the scales: whole video: "*That can be really* inconvenient, *because some videos are like two hours*" was coded as *Very bad*; interval: "*It is somewhat helpful. I'll probably be pretty* satisfied" was coded as *Good*; point: "*I'll just be pretty* happy. *Yeah, it gets me to where I want to go*" was coded as *Very good*. Comments were coded as *Neutral* if a participant's preference changed in context.

Figure 6: Coded mapping of preference levels (i.e., very good, good, neutral, bad, and very bad) to each scale of temporal granularity (i.e., point, interval, and whole video)

For example, P1 reported that the length of the video influences his timepoint preference: "*If it's a 15-minute video, yeah. I would be* okay *with it, because it's just 15 minutes. But, if it's an hour or two hours, then it becomes a slight bit of an inconvenience.*" Following the coded mapping Figure 6, the point scale had nine *Very good*, one *Good*, one *Neural*, and one *Bad*; the interval scale had four *Very good* and eight *Good codes*; and the whole video scale had two *Very good*, three *Good*, one *Neutral*, five *Bad*, and one *Very bad*.

5.2.2 Video Metadata. To assess students' preferences for and against different search result designs, we used the Freidman and Wilcoxon signed-rank tests. We compared the three presentation styles for metadata: 1) Title + Thumbnail, 2) Title + Thumbnail + Keywords, and 3) Title + Thumbnail + Keywords + Summary. Participants' perceived ease in finding information from the retrieved timepoint ($\chi^2(2) = 2.889, p = 0.240$) and design expectation showed no statistically significant differences ($\chi^2(2) = 2.8, p = 0.270$). However, perceived information clarity varied ($\chi^2(2) = 8.914, p = 0.008$). A *post-hoc* analysis found a statistically significant reduction in the clarity of information organization between the Summary and Keywords presentation styles ($z = -2.762, p = 0.004$).

We found that participants had no clear preference for any of the metadata combinations, as there was no statistically significant difference in responses in two of the three Likert scale questions. However, verbal excerpts highlighted some differences between the presentation styles. P3 preferred keywords: "*I would prefer the keywords most, because that one is short. For the abstract one, if it's one or two sentences, I think that will be better.*" P10 wanted a summary: "*I think more the information more it will be easy for me. So, okay, from here it is like more information so I can just read through and recall if that's the video I'm looking for.*" Therefore, we suggest that future systems provide users with a choice of how to display search results.

The inclusion of video transcripts also seemed to be influenced by the video content, compact video viewing in mobile applications. P1 said, "*I would want them for a few videos where the profs are really fast. But if they are like slides, I don't know. May be having an option is good, like we have in youTube but not every time.*" P6 said, "*You know, just like I searched here, there's little ability in a smartphone to search like that.*" An additional observation was that all the participants except one (P3) preferred seeing a thumbnail that showed the video content that matches the note content. For example, P9 explained, "*If that slide was in the thumbnail, that would be the most optimal scenario.*" However, P3 said that displaying key frames may not always be useful: "*I think for a video they always have a thumbnail selected for that video in the system, and I remember that thumbnail. If you change the thumbnail during the search, I'll probably get confused.*"

We found that 11 of the 12 participants mentioned that they would like to see a short ranked-list of matches. For example, P1 preferred having at most three search results: "*Options are definitely good, but how many options are there? I don't want something like Google, where it has pages of matches. Maybe three? Not more than that.*"

5.2.3 Usability and Usefulness. Final survey data on the usability was analyzed. Participants thought that NoteLink was easy to use ($M = 4.42, SD = 0.49$) and were able to become productive quickly using NoteLink ($M = 4.33, SD = 0.62$). Overall, participants were satisfied with our NoteLink ($M = 4.33, SD = 0.47$). When participants were asked about using NoteLink as an app on their mobile phones, all the 12 participants reported that NoteLink would be useful for learning with videos if it was a real application and could be customized. P1 said, "*If it had things I want, yes. Very useful. Saves so much time, and maybe I'll watch more videos then.*" P2 said, "*Definitely, for videos that have to be bookmarked. This is very useful.*"

The interviews suggest that NoteLink would not change students' current note-taking practices. P13 said, "*My note-taking wouldn't change but would be far more helpful.*" The possibility of extending the application of NoteLink to digital notes also played a role in indicating no influence. For example, P4 said, "*Because I write electronic notes I can imagine that could also be something that I can use this snipping tool to take pictures and do that.*" Overall the ability of the system to retrieve video without requiring explicit links, such as timecodes, means students can continue to use their current note-taking style. P12 said, "*I don't think my note-taking process would change as much. No, I don't. I would still write my notes like this. Because it works even now, if I scan this word.*"

6 PHASE IV: ACCURACY

Based on the user feedback collected in Phase III, we redesigned NoteLink to present the three highest-ranking results on each scan. Then, we assessed the ability of the NoteLink to retrieve videos related to the notebook data collected in Phase I. We used NoteLink to test a total of 181 watched video notes and mapped the retrieved timepoint to the point, interval or whole video scales. The evaluation metric used was *accuracy* determined by the ratio of matched notes to the total number of notes.

If any video in the list of three search results was the correct video, it was marked as *Match*. Otherwise, it was marked as *No match*. The timestamp of the retrieved videos were compared with the ground truth timestamp. The video content in both the retrieved timestamp and the expected timestamp were checked and marked if they belonged to the same point or interval, or whole video. If the Recognizer block encountered an error or returned no information, it was marked as *No data*. We excluded 38 video links marked as *No data* from the accuracy calculation. In these cases, the recognizer

Table 1: The number of notes, number of matches, and accuracy for each content type (i.e., text, formula, and drawing) at the three scales of temporal specificity (i.e., point, interval, and whole video).

Note type	Temporal Granularity					
	Point	Interval	Whole video	Overall		
	Count (n)	Count (n)	Count (n)	Count (n)	Notes (n)	Accuracy (%)
Text	36	27	15	78	80	97.5
Formula	5	10	10	25	44	56.8
Drawing	1	2	4	7	18	38.9
Total	42	39	29	110	142	77.5

failed to detect the content because the notes were illegible, due to poor penmanship, or the notes provided insufficient information to calculate similarity scores. Of the remaining 143 links, one link was removed because it did not include a timestamp and video. Therefore, the data from the Matcher block that was available for comparison was 142 links.

About 77.5% of the total links ($n = 142$) returned video(s) that matched the expected videos that participants reported as ground truth. Table 1 presents the number of matches for the three content types on the temporal granularity of point, interval and video. The system matched textual content with an accuracy of 98%, formulas with 57%, and drawings with 39%.

7 DISCUSSION

Our work's contribution extends beyond developing a novel system that retrieves videos related to handwritten notes. In Phase I, we articulated the various content types in handwritten notes: text, formula, drawing, and hybrid. Textual content accounted for about half of all notes and video links collected, indicating its broader use over the other types of content. In addition to the identified types of notes, our investigation of VO also provides insights into each content type. Figure 3 suggests that students tend to copy formulas and figures as they appear in videos, but paraphrase textual information in their notes. Overall, at least 75% of the notes in each type showed some amount of VO, suggesting notes may be effective queries for retrieving videos.

In Phase III, our results indicated that students prefer videos retrieved with a timepoint at the interval scale. Displaying the retrieved video with a timepoint anywhere in the related interval provides a considerably large window of time difference between the expected and the retrieved video timepoint. This is an important finding as it aids future video retrieval systems to conceptualize the timepoint difference in determining the right video context associated with notes. A deeper investigation into the effects of video metadata on video retrieval systems is necessary. Phase III also revealed that students find the proposed linking approach usable and useful for learning from videos.

We calculated the accuracy of NoteLink for different types of note content and scales of temporal granularity in Phase IV. NoteLink matched about 78% of the data to one of the temporal scales. Most matches were at the point and interval temporal granularities, which participants preferred to whole video. The participants' positive attitude towards NoteLink suggests our approach has value as a design specification for other systems for linking handwritten notes and videos.

7.1 Limitations

In this paper, we only collected data from computer science and engineering students. Future work could study other disciplines. This may involve recognizing handwritten notes with content that was not examined in this work (e.g., musical notation). In addition to investigating how disciplinary differences affect note-taking styles, lecture characteristics, such as the modality and the lecture structure, is another critical factor that could be considered.

NoteLink performed well at matching textual note content to videos as there has been extensive work in plain text matching and high-quality APIs in word-spotting exists [49]. However, NoteLink matched less than half of the formulas and drawings links. One reason for this low accuracy could be the difficulty identifying non-textual ROIs in the Matcher block. Similarly, advancements in feature matching are needed to find video correspondences for graphical representations, as there can be more false-positive matches in the case of free-form drawings.

We excluded 38 video notes marked as No-data in Phase IV, from the accuracy calculation. The ability of the scanned query in fetching suitable videos may be improved, taking into account the types of errors likely to occur when taking pictures of handwritten notes. Additionally, the same video(s) can often be searched at multiple points in a page(s) of notes that contain a mathematical derivation, for example. The mobile app can implement these and many other extensions of "paper/digital integration" without requiring changes to the current writing processes giving such conventional paper documents a whole new layer of digital functionality.

7.2 Design Implications

We demonstrated how a point and shoot mobile application can be used to retrieve instructional videos that are similar to handwritten notes. Future work could introduce a number of extensions to improve retrieval accuracy. For example, a NoteLink-like system could allow users to emphasize important parts of their notes to support filtering information and better ranking of search result. Systems could also allow users to append comments, notes, and free-hand drawings to the images of their notes in order to improve

the performance of the Recognizer and Matcher blocks. Future work could also consider matching notes to videos and other learning materials, allowing users to quickly search for information in place. We believe that NoteLink-like systems can eventually create an interactive platform for linking to and from multiple information sources in a usable and useful mobile application.

8 CONCLUSION

This paper presents the requirements and evaluation of NoteLink, a mobile application that uses handwritten note content as queries to retrieve relevant videos. The content of notes were characterized as either text, formula, drawing, or hybrid type. The timepoints videos could be retrieved at had three granularities: point, interval or whole video. Participants had an overall preference for returning videos at the interval scale. NoteLink matched 77.5% of the links to one of the temporal granularity, with 97.5% of text links matched to a video at one of the three scales of temporal granularity. Overall, the findings suggest that NoteLink-like tools are now possible using off-the-shelf OCR technology and could introduce a new approach for bidirectional linking between handwritten notes and instructional videos.

ACKNOWLEDGMENTS

This work was supported by the University of British Columbia Teaching Learning Enhancement Fund (TLEF), the Natural Sciences and Engineering Research Council of Canada (NSERC), and Microsoft. We also thank the ViDeX team, especially Gokberk Ozsoy for his work on handwritten note recognition.[6]

REFERENCES

[1] John Adcock, Matthew Cooper, Laurent Denoue, Hamed Pirsiavash, and Lawrence A Rowe. 2010. TalkMiner: A Lecture Webcast Search Engine. In *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, New York, NY, 241–250.

[2] George Awad, Cees GM Snoek, Alan F Smeaton, and Georges Quénot. 2016. Trecvid semantic indexing of video: A 6-year retrospective. *ITE Transactions on Media Technology and Applications* 4, 3 (2016), 187–208.

[3] Pia Borlund. 2003. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research* 8, 3 (2003). http://www.informationr.net/ir/8-3/paper152.html

[4] Daragh Byrne, Peter Wilkins, Gareth J F Jones, Alan F Smeaton, and Noel E O'Connor. 2008. Measuring the Impact of Temporal Context on Video Retrieval. In *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval*. ACM, New York, NY, 299–308.

[5] Dayvid Castro, Byron L D Bezerra, and Mêuser Valença. 2018. Boosting the Deep Multidimensional Long-Short-Term Memory Network for Handwritten Recognition Systems. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, Piscataway, NJ, 127–132.

[6] Sally Jo Cunningham and David M Nichols. 2008. How People Find Videos. In *Proceedings of the 8th ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, 201–210.

[7] Adrien Delaye and Cheng-Lin Liu. 2012. Text/non-Text Classification in Online Handwritten Documents With Conditional Random Fields. In *Chinese Conference on Pattern Recognition*. Springer, Berlin, Germany, 514–521.

[8] Adrien Delaye and Cheng-Lin Liu. 2014. Contextual Text/non-Text Stroke Classification in Online Handwritten Notes With Conditional Random Fields. *Pattern Recognition* 47, 3 (2014), 959–968.

[9] Samuel Dodson, Ido Roll, Negar M Harandi, Sidney Fels, and Dongwook Yoon. 2019. Weaving Together Media, Technologies, and People: Students' Information Practices in Flipped Classrooms. *Information and Learning Sciences* 120, 7/8 (2019), 519–540.

[10] Brian Dorn, Larissa B Schroeder, and Adam Stankiewicz. 2015. Piloting TrACE: Exploring Spatiotemporal Anchored Collaboration in Asynchronous Learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, New York, NY, 393–403.

[11] John C Dunn. 2004. Remember-Know: A Matter of Confidence. *Psychological Review* 111, 2 (2004), 524–542.

[12] Matthew Fong, Samuel Dodson, Xueqin Zhang, Ido Roll, and Sidney Fels. 2018. ViDeX: A Platform for Personalizing Educational Videos. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, 331–332.

[13] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D Hollan. 2011. ChronoViz: A System for Supporting Navigation of Time-Coded Data. In *Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 299–304.

[14] Marco Furini, Silvia Mirri, and Manuela Montangero. 2017. TagLecture: The Gamification of Video Lecture Indexing Through Quality-Based Tags. In *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, Piscataway, NJ, 122–127.

[15] Ramya Gangaiamaran and Madhumathi Pasupathi. 2017. Review on Use of Mobile Apps for Language Learning. *International Journal of Applied Engineering Research* 12, 21 (2017), 11242–11251.

[16] Gary Geisler and Gary Marchionini. 2000. The Open Video Project: Research-Oriented Digital Video Repository. In *Proceedings of the Fifth ACM Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, 258–259.

[17] Gerard Genette. 1997. *Paratexts: Thresholds of Interpretation*. Cambridge University Press, Cambridge, United Kingdom.

[18] François Guimbretière. 2003. Paper Augmented Digital Documents. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 51–60.

[19] Martin Halvey, David Vallet, David Hannah, and Joemon M. Jose. 2009. ViGOR: A Grouping Oriented Interface for Search and Retrieval in Video Libraries. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, 87–96.

[20] Negar M Harandi, Farshid Agharebparast, Luis Linares, Samuel Dodson, Ido Roll, Matthew Fong, Dongwook Yoon, and Sidney Fels. 2018. Student Video-Usage in Introductory Engineering Courses. *Proceedings of the Canadian Engineering Education Association (CEEA)* (2018), 1–8.

[21] Beverly L Harrison and Ronald M Baecker. 1992. Designing Video Annotation and Analysis Systems. In *Graphics Interface*, Vol. 92. Morgan Kaufmann Publishers, San Francisco, CA, 157–166.

[22] Alexander G. Hauptmann, Rong Jin, and Tobun Dorbin Ng. 2002. Multi-Modal Information Retrieval from Broadcast Video Using OCR and Speech Recognition. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, 160–161.

[23] Jen-Hao Hsiao and Yu-Zheng Wang. 2007. VCenter: A Digital Video Management System with Mobile Search Service. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, 508.

[24] Emanuel Indermühle, Horst Bunke, Faisal Shafait, and Thomas Breuel. 2010. Text Versus Non-Text Distinction in Online Handwritten Documents. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, New York, NY, 3–7.

[25] Renée S Jansen, Daniel Lakens, and Wijnand A IJsselsteijn. 2017. An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review* 22 (2017), 223–233.

[26] Sanna Järvelä, Piia Näykki, Jari Laru, and Tiina Luokkanen. 2007. Structuring and Regulating Collaborative Learning in Higher Education With Wireless Networks and Mobile Tools. *Journal of Educational Technology & Society* 10, 4 (2007), 71–79.

[27] William Jones. 2007. Personal Information Management. *Annual Review of Information Science and Technology* 41, 1 (2007), 453–504.

[28] Matthew Kam, Jingtao Wang, Alastair Iles, Eric Tse, Jane Chiu, Daniel Glaser, Orna Tarshish, and John Canny. 2005. Livenotes: A System for Cooperative and Augmented Note-Taking in Lectures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 531–540.

[29] George Landow. 2006. *Hypertext 3.0: Critical Theory and New Media in an Era of Globalization*. Johns Hopkins University Press, Baltimore, MD.

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

[31] Michael R. Lyu, Edward Yau, and Sam Sze. 2002. A Multilingual, Multimodal Digital Video Library System. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, 145–153.

[32] Anne Mangen. 2008. Hypertext Fiction Reading: Haptics and Immersion. *Journal of Research in Reading* 31, 4 (2008), 404–419.

[33] Gary Marchionini, Yaxiao Song, and Robert Farrell. 2009. Multimedia Surrogates for Video Gisting: Toward Combining Spoken Words and Imagery. *Information Processing & Management* 45, 6 (2009), 615–630.

[34] Catherine Marshall. 1998. Toward an Ecology of Hypertext Annotation. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, 40–49.

[35] Xiangming Mu. 2010. Towards Effective Video Annotation: An Approach to Automatically Link Notes With Video Content. *Computers & Education* 55, 4 (2010), 1752–1763.

[6] https://matheqrecognition.blogspot.com/

[36] Xiangming Mu, Gary Marchionini, and Amy Pattee. 2003. The Interactive Shared Educational Environment: User Interface, System Architecture and Field Study. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE Computer Society, Washington, DC, 291–300.

[37] Pam A Mueller and Daniel M Oppenheimer. 2014. The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking. *Psychological Science* 25, 6 (2014), 1159–1168.

[38] Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)* 2, 331-340 (2009), 2.

[39] Jennifer Pearson, George Buchanan, and Harold Thimbleby. 2011. The Reading Desk: Applying Physical Interactions to Digital Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3199–3202.

[40] Jennifer Pearson, Simon Robinson, and Matt Jones. 2015. PaperChains: Dynamic sketch+ voice annotations. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 383–392.

[41] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout Improves Recurrent Neural Networks for Handwriting Recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*. IEEE, Piscataway, NJ, 285–290.

[42] Réjean Plamondon and Sargur N Srihari. 2000. Online and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and MacHine Intelligence* 22, 1 (2000), 63–84.

[43] Abigail Sellen and Richard H R Harper. 2001. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA.

[44] Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Frédo Durand. 2015. Visual Transcripts: Lecture Notes From Blackboard-Style Lecture Videos. *ACM Transactions on Graphics* 34, 6 (2015), 1–10.

[45] Yu-Chen Song, Gwo-Dong Chen, and Liang-Yi Li. 2013. Improving E-Book Reading With Information Cues: An User Investigation and Suggestion. In *2013 IEEE 13th International Conference on Advanced Learning Technologies*. IEEE, Piscataway, NJ, 261–263.

[46] Thomas F Stahovich and Hanlung Lin. 2016. Enabling Data Mining of Handwritten Coursework. *Computers & Graphics* 57 (2016), 31–45.

[47] Peggy Van Meter, Linda Yokoi, and Michael Pressley. 1994. College Students' Theory of Note-Taking Derived From Their Perceptions of Note-Taking. *Journal of Educational Psychology* 86, 3 (1994), 323–338.

[48] Prem Chand Vashist, Anmol Pandey, and Ashish Tripathi. 2020. A Comparative Study of Handwriting Recognition Techniques. In *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. IEEE, Piscataway, NJ, 456–461.

[49] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-To-End Scene Text Recognition. In *2011 International Conference on Computer Vision*. IEEE, Piscataway, NJ, 1457–1464.

[50] Lynn D Wilcox, Patrick Chiu, Makoto Sasaoka, Jun Miyazaki, David L Hecht, and L Noah Flores. 2010. System and method for video access from notes or summaries. US Patent 7,647,555.

[51] Kyle Williams and Hussein Suleman. 2011. Using a Hidden Markov Model to Transcribe Handwritten Bushman Texts. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, 445–446.

[52] Michael A Wirth. 2003. E-Notes: Using Electronic Lecture Notes to Support Active Learning in Computer Science. *SIGCSE Bulletin* 35, 2 (2003), 57–60.

[53] Amber E Witherby and Sarah K Tauber. 2019. The Current Status of Students' Note-Taking: Why and How Do Students Take Notes? *Journal of Applied Research in Memory and Cognition* 8, 2 (2019), 139–153.

[54] Haojin Yang, Maria Siebert, Patrick Luhne, Harald Sack, and Christoph Meinel. 2011. Lecture Video Indexing and Analysis Using Video OCR Technology. In *2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems*. IEEE, Piscataway, NJ, 54–61.

[55] Sheng-Jie Yang, Gwo-Dong Chen, and Liang-Yi Li. 2011. How Students Use Contextual Cues in Finding Information in Paper and Electronic Textbooks. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*. IEEE, Piscataway, NJ, 302–304.

[56] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S Fels. 2019. "Can you believe [1:21]?!": Content and time-based reference patterns in video comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 489:1–489:12.

[57] Ron Yeh, Chunyuan Liao, Scott Klemmer, François Guimbretière, Brian Lee, Boyko Kakaradov, Jeannie Stamberger, and Andreas Paepcke. 2006. ButterflyNet: A Mobile Capture and Access System for Field Biology Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 571–580.

[58] Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. 2017. A Novel System for Visual Navigation of Educational Videos Using Multimodal Cues. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, New York, NY, 1680–1688.