"Can you believe [1:21]?!": Content and Time-Based Reference Patterns in Video Comments

Matin Yarmand
University of British Columbia
Vancouver, British Columbia, Canada
matin.yarmand@alumni.ubc.ca

Dongwook Yoon
University of British Columbia
Vancouver, British Columbia, Canada
yoon@cs.ubc.ca

Samuel Dodson
University of British Columbia
Vancouver, British Columbia, Canada
dodsons@mail.ubc.ca

Ido Roll
University of British Columbia
Vancouver, British Columbia, Canada
ido.roll@ubc.ca

Sidney S. Fels
University of British Columbia
Vancouver, British Columbia, Canada
ssfels@ece.ubc.ca

ABSTRACT

As videos become increasingly ubiquitous, so is video-based commenting. To contextualize comments, people often reference specific audio/visual content within video. However, the literature falls short of explaining the types of video content people refer to, how they establish references and identify referents, how video characteristics (e.g., genre) impact referencing behaviors, and how references impact social engagement. We present a taxonomy for classifying video references by referent type and temporal specificity. Using our taxonomy, we analyzed 2.5K references with quotations and timestamps collected from public YouTube comments. We found: 1) people reference intervals of video more frequently than time-points, 2) visual entities are referenced more often than sounds, and 3) comments with quotes are more likely to receive replies but not more "likes". We discuss the need for in-situ dereferencing user interfaces, illustrate design concepts for typed referencing features, and provide a dataset for future studies.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Hypertext / hypermedia*; *Empirical studies in collaborative and social computing*; • **Applied computing** → *Hypertext / hypermedia creation*;

KEYWORDS

Video; Comment; Reference; Timestamp; YouTube; Engagement

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300719

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05.
https://doi.org/10.1145/3290605.3300719

1 INTRODUCTION

Commenting on videos is becoming increasingly common in social media and educational platforms [6, 40]. In video-based commenting, referencing specific content can increase engagement with the comment thread [7] and help establish common ground [8]. However, when it comes to watching and discussing a video, the referencing practices people use are underexplored.

HCI and CSCW researchers have suggested interface solutions for referring to specific video content in textual comments: using spatio-temporal markers [11] or embedding subtext links in comments [7]. Several off-the-shelf video interfaces support referencing video content through timestamps (e.g., YouTube). Although existing solutions are driven by established theories and novel design concepts, the field is still lacking *empirical accounts* of referencing practices. Core questions to be answered include: (1) what types of video content do people reference, (2) how do people display their referential intent in textual comments, (3) how do attributes of the video (e.g., genre) impact referencing behaviors, and (4) how do different types of references influence people's reaction to comments, such as social engagement behaviors (e.g., replies and upvotes)?

To examine how people create and use references in videos, we conducted a series of qualitative, quantitative, and preliminary design studies. First, we generated a taxonomy of video

referents[1] based on a thematic analysis of diverse YouTube comments. Second, we applied this taxonomy to approximately 2.5K incidences of referential features collected from public YouTube comments. Based on this coded dataset, we conducted a quantitative analysis to answer our research questions, suggested above, and discussed design implications for video referencing user interfaces. Third, we built and tested a proof-of-concept interface grounded on the implications generated from our empirical findings.

Two attributes of video referencing emerged from our initial analysis: content types (e.g., visual, auditory) and temporal specificity (e.g., time-point, interval). The major findings from our study are: (1) people refer to *intervals* of video more frequently than time-points, (2) type and temporal specificity of referents are closely correlated, (3) genre influences referential behavior, and (4) comments with quotes are more likely to receive replies, but not more "likes". We discuss the need for *in-situ* dereferencing[2] user interfaces, illustrate design implications for expressive video-based commenting, and provide a dataset for future studies. Preliminary evaluation of an empirically-motivated proof-of-concept interface demonstrates the pragmatic utility of the taxonomy and the design implications.

This paper makes three main contributions: (1) A taxonomy for video references and referents, including their relationships, (2) Empirical findings: (2.1) How and what people reference; how "what" people reference affects "how" they reference it; (2.2) How referencing behaviors interact with video attributes, (3) Design implications for video referencing interfaces; typed references and *in-situ* dereferencing, and (4) A coded dataset for future exploration. Our study highlights the importance of referencing types for communication and engagement for video commenting and shines light on the growing need for interfaces that support rich forms of video referencing.

2 BACKGROUND & PREVIOUS WORK

Referencing has been regarded as an important concept when understanding, evaluating, and designing collaborative and communicative tools. Clark theorized the mechanism of referencing that spatial/visual referencing helps interlocutors minimize their joint communicative efforts [9] by establishing common ground anchored at the referred entity [8]. Especially when a collaborative systems involves shared visual information, referring to a specific piece of information using a pointer, pen stroke, or hand overlay has been known to enhance communicative efficiency [14, 15, 24].

In social media studies user engagement is known to be manifested as user-initiated participatory response, such as "likes", replies, shares, and view-counts. Researchers found that affordances of different platforms [21] or types of content (e.g., images of faces [1]) can impact people's engagement with social media. When it comes to YouTube, Kahn's recent study showed that users' different level of media use experience, gender, and motives for social interaction can influence how they engage with YouTube videos [22].

The education literature elaborates on the significance of engagement around shared resources. Scholars such as Lave and Wenger [25] and Engeström [12] suggest that learning is a socially-constructed phenomenon, through which knowledge emerges from participation. This socio-cultural process of meaning-making is highly contextual and specific to learners' goals and environment. However, video interfaces rarely support these social activities that constitute the context of learning [10, 33]. Drawing on this literature, we suggest that further attention to design is needed in order to explore ways in which people can participate and exchange ideas by deepening their engagement with and interpretation of shared resources.

Taxonomy of Video Content & Comments

Several scholars have developed means for categorizing video content. Chorianopoulos [5], Santos-Espino et al. [35], and Sugar et al. [39], for example, classified the content of instructional videos. Our taxonomy was specifically created to understand referencing behaviors in video-based comments. Online commenting is a common feature of many websites. There have been efforts to classify how people comment on the web. Classification schemes can be used by researchers and practitioners to better understand the form and function of comments. Comments on YouTube [20, 26, 28, 32], and other social media services have been analyzed. Madden et al. [28] created a detailed classification scheme for YouTube comments, with ten broad categories and 58 subcategories. Madden et al., however, focused on general YouTube commenting and did not specifically investigate users' referential behaviours. The lack of work on video-based referencing and dereferencing highlights the utility of our taxonomy for understanding these practices and generating design implications.

Interfaces for Referencing Visual Content

Referencing has been a topic of interest in many HCI and CSCW works on collaborative systems, and types of interfaces to support such behaviors have been suggested. When commenting on textual materials, anchoring discussion threads to a specific point [4] or area [45] of text is used by many online discussion platforms. Going beyond visual anchoring, multimodal gestures, such as stylus hovering, ink

[1] We use the term *video referent* to indicate the video content that the user is referring to using their comment reference.
[2] *Dereferencing* means to access the referenced content.

markups, and touch pointing, afford more expressivity, and have begun to appear in experimental systems for text [43] and 3D content [41].

Video referencing interfaces have also been a continuing area of interest. Temporal [7, 27] or spatio-temporal [11] anchoring of a comment thread can be placed at specific times within a video and at specific locations within a frame, leading to a better understanding of what is being discussed in the thread. Stylus inking on video content affords fluid and expressive ways to explicate the content that the addresser is referring to [14, 31]. Fong et al. [13] leveraged time-aligned captions of video as a proxy for anchoring annotations to video contents. Mu [30] developed a tool that automatically generates hyperlinked timestamps between peoples' video comments and the associated time-intervals. Glassman et al. [16] suggested semi-translucent overlays of circles on educational video as students' collective indication of "muddy" or unclear parts of video.

With regard to the research on interaction techniques, our study offers conceptual and empirical frameworks for ideating, designing, and evaluating a new interface. Our empirical findings also provide rich implications for designing novel interfaces in the future, and a coded dataset to be used when testing such interfaces.

3 PHASE I: TAXONOMY OF VIDEO REFERENCES

This section describes our approach for establishing a systematic understanding of what types of video content people referenced in the YouTube comments, and how different types of references and referents were structured. At this stage, we focused on what kind of content people referred to in video-based comments, not on how they used referential functions supported by the existing platforms (e.g., the YouTube timestamp feature). Therefore, in Phase I, a *reference* is any word or phrase in a comment that is clearly pointing to the content in a video. Our mode of inquiry to answer such questions was primarily exploratory. Thus, we took a qualitative approach, generating a taxonomy of reference types. Since we intended to establish a taxonomy that *would be beneficial for inspiring new design concepts*, our interpretations of the incidences of references were guided by the utilities of categories. In other words, we focused on how the *viewer* who is watching the video and reading the comment *interprets* the given reference, instead of taking a purely conceptual or analytical approach that focuses on the linguistic features (e.g., existence of certain prepositions such as "at" or deixis such as "there"), which goes beyond the scope of this paper.

Methods

The qualitative phase of our study consisted of three steps. First, we crawled YouTube comments of a sufficiently large sample size (N = ~1.2K) such that the dataset would contain diverse selections of genre, video, and comment types. Second, we analyzed the comments to look for any patterns, interesting observations, and categories of referent types and temporal specificity. Third, we developed a taxonomy of the aforementioned attributes to be used in the second phase of our study reported in Section 4.

Data Collection. To diversify our dataset, we sampled from 7 main representative video genres. From the YouTube classification system, we first selected 4 out of 32 genres that are closely related to Yuan et al.'s [44] hierarchical ontology on video genres: *News & Politics, Music, Sports, and Film & Animation*. We added 3 other categories that are of particular interests to the HCI research community. We selected *Education* to capture the popularity of Massive Open Online Courses [6] and the interesting visual and speech elements typically found in educational videos. Moreover, we picked *How-To & Style*, due to the value of this type of video in HCI crowd-sourcing research [23] and *People & Blogs* to take into account the uniqueness of vlogs (video blogs) as visual media [29]. Our selection process is in accordance with previous studies on YouTube genres [e.g., 34, 42].

Using the YouTube Data API, we crawled the top 700 most-viewed videos published to YouTube between 2010[3] and 2018 (7 genres × 100 videos per genre); incorporating popularity to collect data simulates the users' behavior of viewing videos, and this measure is well explored in previous research [36, 38]. Out of 100 videos in each category, we selected 6 distinctively different videos. In order to maximize variation, purposive sampling was used based on video attributes (e.g., duration, date of publish, and YouTube channel) and video content (e.g., whether it contains speech or music, and the participating cast).

Once the collection of 42 videos (7 genres × 6 videos per genre) was selected, we crawled 30 top-level (non-reply) comments for each video,[4] for a total of 1,260 comments (42 videos × 30 comments per video). By only collecting top-level comments, we limited the dependency of our dataset on other comments in the same thread, thus ensuring that each reference is an independent data point. The top-level comment extraction was based on *relevance*, as determined by YouTube. Incorporating the YouTube relevance measure simulates the viewers' experience by seeing the comments in the same order as they appear on the YouTube website.

Analysis. Thematic coding of the set of 1,260 comments was performed by the lead investigator and discussed further with the other authors for conflict resolution and validation.

[3] We chose to sample videos uploaded starting from 2010, because the YouTube timestamp feature was in widespread use by then. The earliest mention we found is on 2008 [18].

[4] Each video contained many more top-level comments.

Our analysis took two steps: (1) identifying references and (2) categorizing them. When determining whether or not a keyword was a reference in step 1, our decision was governed by the existence of concrete referential words and phrases in comments, for the sake of robust and concrete sampling. We refrained from guessing an insinuated referential intent of the commenter (with no clear referential expression) based on surrounding context, because this can be highly ambiguous and subjective. For example, the following comment on a makeup tutorial video was tagged to contain no reference: "Gorgeous, with or without makeup". However, if the comment was '*You* are gorgeous, with or without makeup', the word "You" would be considered referential to a person in the video. During the categorization step (2), an identified reference was coded based on emerging thematic categories. When determining the type of attributes of a reference, the investigators read the comment and watched the video to reach an agreed interpretation about what is the referent content.

Two core attributes of video referencing emerged: type of referred content (referent type) and time range where the referred content appears in the video (referent temporal specificity).

Video Referencing Taxonomy

After analyzing the coded comments, a taxonomy was developed for content type and the associated temporality of referents, which can be viewed in Tables 1 and 2, respectively.

Referent Type

We grouped the 14 identified referent types into 5 high-level categories (see Table 1); *Visual* denotes referents that can be seen in the video. *Auditory* referents include verbal and non-verbal elements that can be heard. Even though a Lyrics element is part of a song, we decided to separate it from Auditory because of the unique characteristics that each category exhibit, and how distinctively Auditory and Lyrics elements were referred to in the dataset. Some comments clearly referred to video content, but the referent was not a substantial entity, but rather a conceptual one (e.g., idea or concept). We categorized these referents as *Generic/Conceptual*, in which we grouped Concept (Textual) and Concept (Oral) depending on the mode through which the concept was conveyed. Metadata (attributes such as view count and the channel) and Meta-information (not part of video content but relevant to video, such as the singer's newest album in one of his or her older music videos) both refer to entities outside the scope of the video content. Lastly, *Inconclusive* refers to the use of a timestamp which has no further contextualized description provided.

Defining Inconclusive References for Usable, Reliable Taxonomy. Some comments have clear referential features, such as a timestamp, but lack an added descriptive text that specifies the referent (as shown in the example in Table 1). In these cases, even though the referential intent of the commenter was apparent, the lack of context could generate a high degree of uncertainty in the way comment viewers interpret the meaning of the reference. When analyzing such context-deficit references, we found that coders—the investigators—often disagreed with each other's interpretation, even after watching the timestamped sections of videos for further context. This suggests that commenters might be minimizing their level of effort for describing a reference *at the expense of the clarity of their messages*. Fortunately, whether or not a reference has descriptive text was a very distinctive criterion for identifying such cases. Thus, to generate a robust and useful taxonomy, we categorized these comments as a separate type called *Inconclusive* references, in which the way a viewer understands the referent can vary (be inconclusive) by one's experience and context. In our taxonomy, referent type and time specificity of Inconclusive reference are denoted as "Inconclusive (Type)", and "Inconclusive (Time)" respectively.

Temporal Specificity

As depicted in Table 2, the three temporal characteristics of referencing were identified as *Point*, *Interval*, and *Whole Video*. *Point* is when a referent is noted at a distinct moment in the video (e.g., "when she almost fell over at 3:40 tho"). *Interval* referents occur over a span of time, such as when an actor performs an activity or makes a speech. *Whole Video* is when the referent is apparent during the entire video; for example, talking about a song in a music video is considered *Whole Video*. Although *Inconclusive* references have a timestamp, they were categorized as a separate type, because it was unclear whether commenters intended to refer to the exact moment of the timestamp or the beginning of an interval. *Metadata* and *Meta-information* were not associated with time.

Prioritizing Specific Referent for Resolving Ambiguous Cases. We came across several ambiguous incidences that warranted two different interpretations, in which we favored the more specific referent. For example, Actor and Event tend to be accompanied together; for instance, the comment "*God bless whoever sneezed 0:06*" could be referencing the act of sneezing (*sneezed*) as well as the person who sneezed (*whoever*). In this example, the actor appears throughout the video, but his sneeze is relevant to the very moment of the event, so *sneezing* is the reference. Regarding temporal specificity, confusion arose from interpreting Point versus Interval. For example, an Actor or Object can appear in multiple spans

Table 1: Taxonomy of Referent Types with Examples
Bold words indicate the specific keywords that were considered to identify the referent.

Category	Referent Type	Definition	Example(s) from Dataset
Visual	Actor	Any distinctive agent with visual characteristics which is associated with performing an action	- 2:50 **this kid** is soooo funny xD
	Object	Any distinctive item with visual characteristics that is not capable of performing an action	- What if I dont have **a jacket** 3:08
	Event	An occurrence; could be performed by an actor, or just a happening in the video.	- You could see how happy that marine was to be given the honor of **shaking the presidents hand** :)
Auditory	Speech	The spoken, verbal content which is narrated by an actor (verbatim)	- **"When the game is on the line I'm on my prime"** - Ty 2016
	(non-verbal) Sound	Any sound/noise that is captured in the video but cannot be associated with words	- @Jogwheel what's name of the **background music** pleeeeeeeease???
	Song (whole song)	Representing a song as a whole; including lyrics, melody, musical instruments, and etc	- This is the best world cup **song**! So much energy!
	Lyrics	The verse of a song	- **"Sugar Yes please"** I wanna hear your voice [...]
Generic/Conceptual	Concept (Textual)	An idea or perception that textually appears	- i thought you would get a bit more room for **23000$**... (referring to the ticket price that is displayed in the video)
	Concept (Oral)	An idea or perception that is orally presented in the video (non-verbatim)	- No big deal just **3 stories** !!
	Scene	A non-specific subsection of the video	- **this** is extremely disturbing
	Time & Location	The settings; when and where the video takes place	- Always beautiful **India**
Inconclusive	Inconclusive (Type)	When a timestamp is present, with no further contextual description of the reference	- **10:34** ... In case you're wondering
Meta	Metadata	Descriptive data related to the video, but not the actual content in the video	- **57 m views** lol
	Meta-information	Usage of video content in a hypothetical or real-world context (doesn't appear in video)	- **Justin Bieber** is best singer

Table 2: Taxonomy of Temporal Specificity with Examples

Category	Temporal Specificity	Definition	Example(s) from Dataset
Temporal	Point	Occurring at a distinct moment in the video	- **6:01** Simons head shot up with a smile of his face.
	Interval	A continuous time span in the video	- **First audition** was just wow
	Whole Video	Presented in the majority of the video	- **This song**'s one of my favorites from Ed. So cute.
Other	Inconclusive (Time)	Ambiguous temporal specificity, used with Inconclusive (Type)	- **10:34** ... In case you're wondering
	N/A	Not associated with time (Metadata and Meta-information)	- 57 m **views**?

of time, but content of the comment can be relevant to the referent's presence at a distinct moment. In such cases, we favored Point over Interval.

Referential Features. We found that several features or expressions were used frequently to signify referential intent. The YouTube timestamp feature was the most popular one that comes with a syntactic highlight (blue markup) and dereferencing function (click-to-replay). Other non-functional, but popular referential indicators included quotes—both double and single—and colons, which are usually associated with Speech or Lyrics types.

4 PHASE II: HOW PEOPLE USE REFERENTIAL FEATURES IN YOUTUBE COMMENTS

In this phase, we aimed to generate empirical evidence of how people use the referential features to comment on video content. Our focus was not on generic referential expressions, but on the *referential features/indicators* identified in Phase I that served referential functions. These referential features can be detected using regular expressions, which allowed us to collect a large number of incidences for generating and testing empirical claims. In this step, it is worth noting that the unit of analysis changed from comments to references; In Phase I, references were identified from comments; In Phase II references were directly detected by textual expression. In

this phase, multiple referential features in a single comment were considered as multiple incidences.

Methods

Here we describe how we collected ~2.5K references and analyzed them based on the taxonomy from Phase I.

Data Collection. We collected a preliminary dataset consisting of 3,500 auto-detected *potential* references.[5] The data was collected through an iterative crawling strategy that scraped 100 most relevant comments of 50 most viewed videos per each of the 7 genres. While extracting the references in the scraped comments, this procedure continued until we had at least 500 potential references for each genre (7 genres × 500 potential references per genre). Total 64,373 comments have been tested. When computationally detecting references, we used a language detection algorithm to filter out non-English data. To obtain the final dataset of annotated references, we conducted an analysis as follows.

Coding and Post-Processing. For each reference in the dataset, two native English-speaking coders independently identified (1) whether it is a reference or a false positive, and (2) the referent type and temporal specificity, if it is a reference. We provided them with a code book similar to Table 1 and Table 2. The code book was generated from generic comments and were not often accompanied by referential features. However, the coders successfully applied the taxonomy to the new dataset without encountering breakdowns or requiring a restructuring of the taxonomy. Throughout the training period, and as the coders were familiarizing themselves with our taxonomy, they met regularly with the lead investigator to discuss the ambiguous cases and major conflicts of opinion.

A cross-validation score was calculated over part of data (N = 1,711, 57.15%, with the remaining references used for training). The score indicated a strong agreement between the coders. Referent type and temporal specificity had Cohen's kappa scores of .72 and .64, respectively. The kappa scores were calculated based on 1,711 references of the total dataset. After the Cohen's kappa scores were calculated, disagreements were resolved by a three-way discussion between the two coders and the lead investigator. The coding process took approximately 40 hours per coder.

During the post processing of the coded preliminary dataset, we removed several types of irrelevant data. First, 16% of the dataset contained false positives, with no reference, such as spam comments. Second, during annotation, we found that the majority of colons (151 of 196, 77%) were not associated with a reference; colons were used in various non-referential ways, such as listing items in a comment. Therefore, we decided to remove colons from the dataset and demote colons from our list of referential indicators. Single quotes and double quotes had lower false positive rates of 40% and 47% respectively, therefore we decided to keep those two indicators. As a result, our final dataset has a total 2,517 references, 2,086 contained timestamps and the rest quotation marks, of which 405 contained double quotes and 26 single quotes.

[5] At this stage, some auto detected referential expressions might not be real references, such as a comment about the time ("who else is watching this at 11:00 pm?") which does not reference a time in the video.

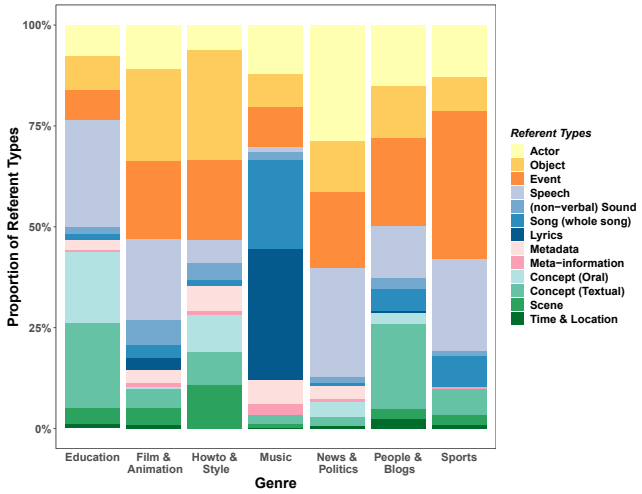

Figure 1: Referent type distribution across different Genres. For clarity and convenient comparison of high-level ideas, the referent types in the same high-level category were colored similarly.

5 RESULTS

To examine how referencing behaviors are related to attributes of video and user engagement, we analyzed the coded reference dataset. The dataset containing quotes and timestamps was addressed separately to prevent introducing biases arising from the type of indicator.

What Impacts Video Referencing Behaviours?

Genre Affects Referent Types. We were interested in whether people refer to different types of video content when commenting on different genres (see Figure 1). Music, for instance, attracted many Auditory referents (Song and Lyrics). Education contained mostly Speech—e.g., what the instructor says—and Concept (Textual)—e.g., what is written in the video—referents. Object was the most popular referent type in How to & Style, perhaps because of product reviews or Maker-contents. Sports had a large portion of Event references. Interestingly, News & Politics attracted the most Actor referents among all genres, even more than Film & Animation which was expected to include more references to Actor.

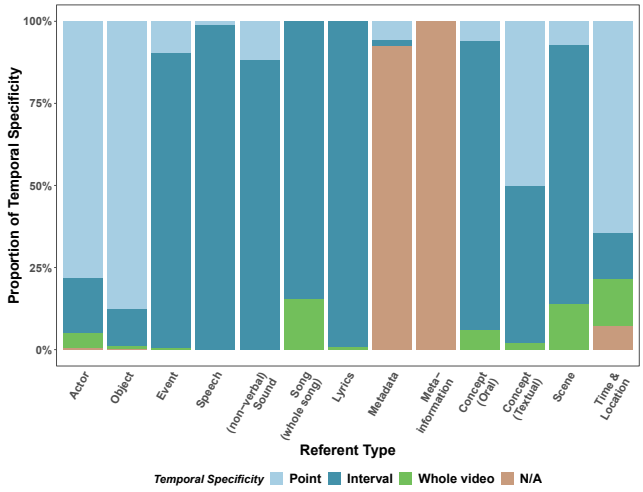

Figure 2: Proportion of referent temporal specificity across different referent types.

Referent Type is Dominated by Specific Temporal Specificity. To better understand the nature of referencing practices, we studied how referent type and referent temporal specificity are related. Chi-square testing showed that the correlation between temporal specificity and referent type is statistically significant, $\chi^2(56, N = 2{,}517) = 7{,}360.8$, $p < .001$. As shown in Figure 2, nearly each type of referent was dominated by a particular temporal specificity. Event and Scene referents were almost all interval-based, while Actor and Object were usually time-points. Concept (Textual) and Time & Location were the only two referent types that used more than one predominant type of temporal specificity. Concept (Textual) contained an equal mix of Point and Interval temporal specificity.

How Do People Use Timestamps?

How to & Style Genre Attracts Many References. We were interested in how the YouTube timestamp feature is used in each genre. A straightforward quantitative measure of timestamp use is the proportion of comments that contain a timestamped reference to all comments. We found that How to & Style videos attract the most referential comments (4.22%), while Sports the least (1.60%). This result is consistent with previous work by Shultes et. al. [37], where ~2% of comments were observed to have a timestamp. Shultes et. al. also suggested that Sports is an anomaly genre with a timestamp rate of 13%, which is much higher than average. However, our data is more extensive in terms of time ranges (2010–2018) and total number of comments considered (~2.9K potential references[6] from 64,374 comments).

[6]For the sake of compatibility to the previous work, results we applied *unfiltered preliminary dataset.* in this analysis.

Table 3: Distribution of Referent Types with Use of Timestamp and Quotes

Referent Types	Timestamp	Quotes
Visual	685 (32.84%)	35 (8.15%)
Auditory	281 (13.47%)	283 (65.97%)
Generic/Conceptual	251 (12.03%)	60 (13.99%)
Inconclusive (Type)	848 (40.65%)	0 (0%)
Meta	21 (1.01%)	51 (11.89%)
Total	2,086 (100%)	429 (100%)

Table 4: Distribution of Temporal Specificity with Use of Timestamp and Quotes

Temporal Specificity	Timestamp	Quotes
Point	445 (21.33%)	37 (8.62%)
Interval	757 (36.26%)	297 (69.23%)
Whole Video	14 (0.68%)	37 (8.62%)
Inconclusive (Time)	848 (40.65%)	0 (0%)
N/A	22 (1.05%)	58 (13.53%)
Total	2,086 (100%)	429 (100%)

Popularity of Visual, Interval, and Inconclusive References. We were interested in the overall distribution of referent types, as presented in Phase I of the study. Among Conclusive references, of which our coders successfully identified the referent, *Visual* (referent type, 32.84%) and *Interval* (referent temporal specificity, 36.26%) were the most popular. Also, 40.65% of our ~2.5K references were Inconclusive. Further qualitative analysis of Inconclusive comments revealed additional functions of the context-deficit reference: functions such as (1) indicating distinct moments in the video that could be easily understood once the video was watched at the timestamp, (2) conveying a humorous or satirical message; explaining the reference in details could jeopardize the delivery of the joke, and (3) minimizing the effort necessary for describing the comments, especially when the referential element is hard to describe textually [19].

How Do People Use Quotes?

Quotes for Speech and Lyrics Reference. During Phase I, we observed an abundance of double quotes that were used for *verbatim* referential intent (Speech and Lyrics). In Phase II we found that out of 405 quoted references, 142 were about Speech (35.06%), 117 were referencing Lyrics (28.89%), followed by Concept (Textual) and Metadata with 35 (8.64%) and 32 (7.90%) instances, respectively. The other 10 types of referents were used 79 times in total (19.51%).

Bi-Modal Distribution of Verbatim Transcription Accuracy. As quoting emerged as a major means for referencing verbal content, we measured the level of accuracy of verbatim, quoted text. We analyzed 65 quoted comments made to videos that contained manually created captions. The majority of the references (76.9%) fell within the first 10% of Levenshtein distance (0% representing perfect matching). A noticeable 12.3% of the data lies between 20% and 30% of Levenshtein distance. In the discussion section, we offer a potential reason why the distribution is bi-modal.

Conclusive Reference Can Increases Replies, but Inconclusive Cannot

The referential intent of Inconclusive references can be difficult to assess by reading the comments alone, due to the deficiency of within-comment context about the referent. Having to watch the referenced video by clicking the timestamp *jumps* the viewer's viewport to the top of the page and can divert the viewer's attention away from the comment, potentially being disruptive for following the comment thread.

To statistically test this hypothesis, we examined how comments with references, and specifically having different types of reference (Inconclusive and Conclusive) can engage viewers with the comment in the form of a *"like"* or *reply*. Following [1]'s analysis for social media, we used Negative binomial regression for our analysis as it is suitable for over-dispersed count variables (see Table 5). In our analysis, the number of views and comments were incorporated as two additional covariates, as these factors can introduce biases into the analysis of engagement.

The results showed that Conclusive references, either timestamped or quoted, received more replies. Interestingly, having Inconclusive references does not impact the number of replies or likes that a comment will get, as we hypothesized above. Similar patterns were shown for "likes" too, but what is different from replies is that the number of "likes" was not affected by Conclusive timestamped references. It is also worth noting that Quotes may draw attention to both types of social engagement behaviors.

Table 5: Impact of Reference Types on Social Engagement (Results of Negative Binomial Regression test)

Type	Variable	Estimate	Standard Error	*p*-Value
Replies	Inconclusive (Timestamped)	.07	.105	.50
	Conclusive (Timestamped)	.16	.037	<.001***
	Conclusive (Quoted)	.24	.108	.028*
Likes	Inconclusive (Timestamped)	-.11	.092	.23
	Conclusive (Timestamped)	.03	.033	.37
	Conclusive (Quoted)	.49	.095	<.001***

6 DISCUSSION AND DESIGN IMPLICATIONS

Here we discuss the results from Phase I and II and generate implications for designing improved referencing interfaces.

Reference is not the Message: Needs for Dereferencing Interfaces that Engage Comment Viewers

If we view video comments as a communication channel, where benefits of the two ends of communication—message producer and consumer—need to be balanced [17], the high frequency of Inconclusive comments suggests a significant imbalance caused by the particular design of interface. It seems that some commenters create Inconclusive comments for minimizing their effort or maximizing intended communicative effect. However, our test on user engagement indicated that the viewers can be disengaged from the comment due to the specific *dereferencing feature* of YouTube that moves the viewers' viewpoint to the top-video location and away from the comment. Quotes can be easily understood without additional context, and it turns out that quoted comments attract both more replies and "likes". This result suggests that there is a need for viewing users to access referenced video contents without leaving the comment section:*in-situ* dereferencing.

Conclusive references attract replies, not "likes". Khan [22] explicates different motivations for the two indicators of user participation in YouTube. "Likes" tend to manifest motivation for relaxing interaction, while replies are for social interactions with a higher level of engagement. This implies that referencing using a timestamp can be an effective way to draw users' attention when they are motivated to participate in the social interaction with the commenter. However, if viewers want to casually enjoy a video, timestamps that require clicking may not be the best means of engagement.

Needs for Better Video Referencing Interfaces & Their Requirements

By analyzing the general distribution statistics regarding users' referential behaviours (see Tables 3 and 4), we observed limitations regarding the use of timestamps. Table 4 shows that Intervals are the most prominent form of temporal specificity in our dataset. The timestamp feature is incapable of addressing such reference; using a timestamp indicates a point and not using it indicates the whole video. Therefore, commenters have to creatively work around the limitation of popular commenting interfaces; for instance, some users were observed using dashes (e.g., *"4:40-4:43 is kinda dirty especially if a boy does it!!"*) and time-specifying keywords (e.g., *"*clicks on video and watches it **until** **1:53**"* or *"Good info **up to** 5:30 w/ one correction [...]"*). Due to the

popularity of referencing video elements that appear in a span of time, unique interval selection techniques can facilitate a better referencing experience; for example, the user may select an interval directly from the seek bar and include it into the comment section.

Table 3 reveals another limitation of timestamps; without any further description, a timestamp is referential to the entire scene that is specified at a particular point in the video. In our taxonomy, Scene is a subcategory of Generic/Conceptual which is used far less than Visual. Thus, it is important to develop advanced techniques to incorporate various referent types. For example, an interface may accommodate referencing an Object/Actor element by allowing the user to directly select an associated region of video for inclusion in the comment.

Correlated Type & Temporal Specificity: Implication for Typed Video (De)Referencing

Figure 2 depicts the strong dependency of most referent types on a particular temporal specificity which suggests interesting design implications for anchoring the type of referent to some temporal feature. For example, Actor and Object tend to be anchored to Point. Thus, showing a snapshot of that time point is enough to understand what the referent is. Event and Auditory Referents are mostly intervals, which means that they need to be played (as audio snippet or video clip) to be understood at the point of dereferencing. The referential method for the mixed types—Concept (Textual) or Time & Location—needs to be determined by the commenter at the point of referencing.

Genre as Predictor of Referent Type

We presented how genre affects the referent type in Figure 1. Such a relationship shows an opportunity for referential systems to account for the type of the video content in order to adapt the type of referential link to users' needs. For instance, video interfaces that are primarily focused on playing Sports can consider the abundance of references to Events to better facilitate their users' referential needs, such as incorporating a drop-down menu of the most common events.

Automation of Verbatim References

Nearly two-thirds of double quoted references were dedicated to verbatim referential behaviours (Speech and Lyrics for 63.95%). While there were cases of double quotes referencing other types of referents, it seems reasonable to associate double quotes with spoken content. This suggests a design implication for using double quotes as a keyword for activating a tool for referencing a Speech or Lyrics element, while giving the user the option to dismiss the tool.

Table 6: Examples of Imperfect Verbatim References

Cause	Original Caption	User's Reference	Levenshtein Distance
Expressivity	oh no	oh noooo	37.5%
Typo	benjamin button	bejnamin button	11.1%
Mishearing	finish her off	finish it off	21.4%

Our findings regarding the Levenshtein distance suggest that when making verbatim references, most people accurately type the intended piece of the spoken content in their comment. In order to get a better sense of the imperfect verbatim references, we qualitatively analyzed the dataset and found the most common explanations (see Table 6 with an example).

By combining the previous two claims, we suggest that once a double quote is typed in the comment section, autocompletion can be used to help the user create an error-free verbatim reference. However, in case the intention is not to reference Speech or Lyrics, pressing the Escape key exits the auto-complete mode and the user is back to the normal typing interface.

Proof-of-Concept System

To demonstrate the feasibility of typed referencing and *in-situ* dereferencing, we designed and built a proof-of-concept interface. The design concepts of the interface were firmly grounded on the implications from the previous sections. For typed referencing (3 types, as displayed in Figure 3), users can drag the seek bar for Interval referencing, video screen for Visual referencing, and in-video caption for Speech referencing. Each type of reference is time-mapped and dereferenced along with the temporal specificity, as suggested by Figure 2. For *in-situ* access to the referent, we came up with the hover-and-replay concept where viewers can access the different types of referent through a tool-tip without leaving the comment section. To accommodate Speech referencing, an auto-complete menu is summoned when a user types a double-quote.

To evaluate the feasibility of our design concepts and preliminary user attitude toward them, we conducted an informal comparative study of our proof-of-concept interfaces against the YouTube timestamp feature. Eight undergraduates were recruited using convenience sampling. After the experimenter's demonstration, participants used each interface to create references to three types of video content (Interval, Visual Object, and Speech) and to dereference different types of content referred in the given comments. We collected System Usability Scale (SUS) [3] ratings and a written qualitative response after each session. The order of sessions and materials was balanced.

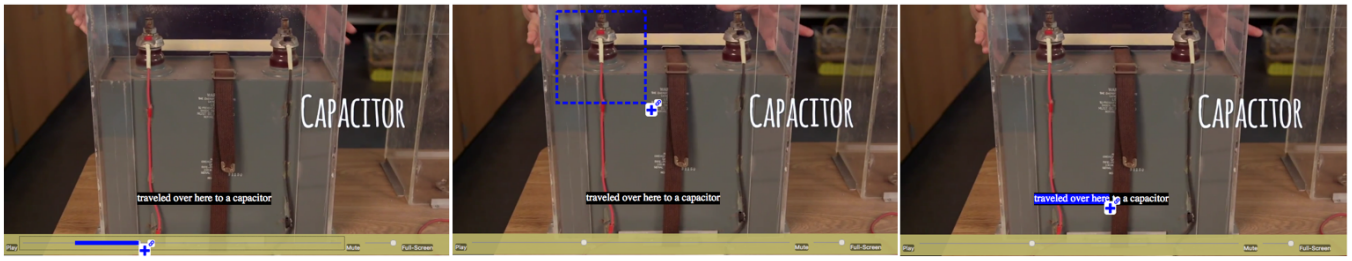

Figure 3: Interfaces for typed referencing; selection of Interval (left), Visual Object (middle), and Speech (right).

Participants rated the overall usability of our referencing interfaces *acceptable by a large margin* with the mean SUS score of 81.3 ($SD = 10.2$), while a score > 70 deemed acceptable in [2]. This score is slightly higher than the SUS score for the YouTube timestamp feature ($M = 77.5, SD = 11.6$), but these results did not show a significant statistical difference. The preliminary themes of the qualitative written responses indicate (1) participants enjoyed our *in-situ* dereferencing design concept as it allowed them "to hear/see/watch the referenced items without having to jump back to the video", (2) dragging to select a referent using our interfaces is more convenient than "continuously replay(ing) frames to find the exact time to reference" in YouTube, (3) participants found the typed referencing interface "intuitive", "easy to use", and feasible ("this is the next step for YouTube to improve their platform" and they could "see many people using" such features).

7 CONCLUSION & FUTURE WORK

We described our three-pronged approach to better understand and support video referencing and dereferencing. First, we analyzed YouTube comments to understand what people refer to (type) and how they do so (temporal specificity). Second, using this taxonomy, we classified 2.5K YouTube comments. We found that most comments describe visual elements in the video, and that the use of quotes is very common. More surprising is that most temporal references described time intervals and not points. This is especially interesting given that the YouTube interface is much more amenable to talking about time points. Moreover, certain referent types are more likely to be referenced in specific ways, and certain video attributes (such as genre) encourage specific ways of talking about the video. The way in which people refer to video is associated with the amount of engagement that the comment generates. We have also provided public access to the dataset of annotated references for continued research.

These findings inspired the design of novel referencing mechanisms. Our new designs include affordances that support the commenting patterns identified above. Namely, the proof-of-concept interface supports referencing video intervals and screen regions, enables quick quoting, and above all, supports dereferencing in the comment itself, so that the users can engage with comments without leaving the comment. A preliminary evaluation of the interface found that it supports intuitive ways of video referencing.

The study has several limitations. With regard to the taxonomy, we focused only on referent types and temporal specificity. Other aspects of the referencing behaviors that may be important fall outside the scope of the current study. While we diversified the content of the videos in our quantitative analysis, all of the comments were harvested from YouTube. It may be that the YouTube interface affects the nature and distribution of comments. Additionally, while a larger span of time for video selection (i.e, 2010 to 2018) allowed us to consider a diverse set of videos, this period might have introduced biases into our analysis due to the changes in YouTube user interface over the last nine years. Furthermore, with regard to the proof-of-concept interface, a more thorough evaluation may teach us more about the nature of the suggested mechanisms. Future work will need to address these limitations, as well as explore additional opportunities that are afforded by the annotated data.

Overall, this study took us from defining a taxonomy of video references, to evaluating frequencies of reference attributes and their relationships, to designing a proof-of-concept video interface informed by the taxonomy and empirical findings. This approach helps us get closer to a contemporary video interface supporting users describing, conversing, and commenting about what they see and hear in a way that fosters meaningful engagement and communication.

8 ACKNOWLEDGEMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC, CRDPJ 508852-17 and RGPIN-2018-04591), the University of British Columbia Teaching and Learning Enhancement Fund, and Microsoft Corporation.

REFERENCES

[1] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on Instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 965–974.

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] A. J. Bernheim Brush, David Bargeron, Jonathan Grudin, Alan Borning, and Anoop Gupta. 2002. Supporting Interaction Outside of Class: Anchored Discussions vs. Discussion Boards. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community (CSCL '02)*. International Society of the Learning Sciences, 425–434.

[5] Konstantinos Chorianopoulos. 2018. A taxonomy of asynchronous instructional video styles. *The International Review of Research in Open and Distributed Learning* 19, 1 (2018), 294–311.

[6] Gayle Christensen, Andrew Steinmetz, Brandon Alcorn, Amy Bennett, Deirdre Woods, and Ezekiel Emanuel. 2013. The MOOC phenomenon: Who takes massive open online courses and why? https://dx.doi.org/10.2139/ssrn.2350964.

[7] Soon Hau Chua, Toni-Jan Keith Palma Monserrat, Dongwook Yoon, Juho Kim, and Shengdong Zhao. 2017. Korero: Facilitating complex referencing of visual materials in asynchronous discussion interface. *Proceedings of the ACM on Human-Computer Interaction* 1 (2017), 34:1–34:19.

[8] Herbert Clark. 1996. *Using language*. Cambridge University Press, Cambridge, United Kingdom.

[9] Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.

[10] Samuel Dodson, Ido Roll, Matthew Fong, Dongwook Yoon, Negar M Harandi, and Sidney Fels. 2018. Active Viewing: A Framework for Understanding Student Engagement With Educational Videos. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, New York, NY, 24:1–24:4.

[11] 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.

[12] Yrjö Engeström. 1987. *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Cambridge University Press, Cambridge, United Kingdom.

[13] 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.

[14] Susan R Fussell, Leslie D Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam DI Kramer. 2004. Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction* 19, 3 (2004), 273–309.

[15] Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2004. Language Efficiency and Visual Technology: Minimizing Collaborative Effort with Visual Information. *Journal of Language and Social Psychology* 23, 4 (2004), 491–517.

[16] Elena L Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1555–1564.

[17] Jonathan Grudin. 1988. Why CSCW Applications Fail: Problems in the Design and Evaluationof Organizational Interfaces. In *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work (CSCW '88)*. ACM, New York, NY, USA, 85–93. https://doi.org/10.1145/62266.62273

[18] Christian Heilmann. 2008. YouTube now offers deep links to timestamps (via URI hash). Retrieved Jan 05, 2019 from https://christianheilmann.com/2008/10/26/youtube-now-offers-deep-links-to-timestamps-via-uri-hash/

[19] Michel Hupet, Xavier Seron, and Yves Chantraine. 1991. The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology* 82, 4 (1991), 449–462.

[20] Graham M Jones and Bambi B Schieffelin. 2009. Talking text and talking back:"My BFF Jill" from boob tube to YouTube. *Journal of Computer-Mediated Communication* 14, 4 (2009), 1050–1079.

[21] Anastasia Kavada. 2012. Engagement, bonding, and identity across multiple platforms: Avaaz on Facebook, YouTube, and MySpace. *MedieKultur: Journal of media and communication research* 28, 52 (2012), 21.

[22] M. Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior* 66 (2017), 236–247.

[23] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 4017–4026.

[24] David Kirk, Tom Rodden, and Danaë Stanton Fraser. 2007. Turn It This Way: Grounding Collaborative Action with Remote Gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1039–1048.

[25] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, United Kingdom.

[26] Yun-Jung Lee, Jung-Min Shim, Hwan-Gue Cho, and Gyun Woo. 2010. Detecting and visualizing the dispute structure of the replying comments in the internet forum sites. In *Cyber-Enabled Distributed Computing and Knowledge Discovery*. IEEE, Piscataway, NJ, 456–463.

[27] Scott LeeTiernan and Jonathan Grudin. 2001. Fostering Engagement in Asynchronous Learning through Collaborative Multimedia Annotation.. In *INTERACT*. Citeseer, 472–479.

[28] Amy Madden, Ian Ruthven, and David McMenemy. 2013. A classification scheme for content analyses of YouTube video comments. *Journal of Documentation* 69, 5 (2013), 693–714.

[29] Heather Molyneaux, Susan O'Donnell, Kerri Gibson, and Janice Singer. 2008. Exploring the gender divide on YouTube: An analysis of the creation and reception of vlogs. *American Communication Journal* 10, 2 (2008), 1–14.

[30] Xiangming Mu. 2010. Towards effective video annotation: An approach to automatically link notes with video content. *Computers & Education* 55, 4 (2010), 1752–1763.

[31] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: Video-based asynchronous video review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, New York, NY, 517–528.

[32] Martin Potthast and Steffen Becker. 2010. Opinion summarization of web comments. In *European Conference on Information Retrieval*. Springer, Berlin, Germany, 668–669.

[33] Ido Roll, Daniel M Russell, and Dragan Gašević. 2018. Learning at Scale. *International Journal of Artificial Intelligence in Education* (2018),

1–7.

[34] Dana Rotman and Jennifer Preece. 2010. The'WeTube'in YouTube–creating an online community through video sharing. *International Journal of Web Based Communities* 6, 3 (2010), 317–333.

[35] José Miguel Santos-Espino, María Dolores Afonso-Suárez, and Cayetano Guerra-Artal. 2016. Speakers and boards: A survey of instructional video styles in MOOCs. *Technical Communication* 63, 2 (2016), 101–115.

[36] Peter Schultes, Verena Dorner, and Franz Lehner. 2013. Leave a comment! An in-depth analysis of user comments on YouTube. *Wirtschaftsinformatik* 42 (2013), 659–673.

[37] Peter Schultes, Verena Dorner, and Franz Lehner. 2013. Leave a Comment! An In-Depth Analysis of User Comments on YouTube. *Wirtschaftsinformatik* 42 (2013), 659–673.

[38] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, 891–900.

[39] William Sugar, Abbie Brown, and Kenneth Luterbach. 2010. Examining the anatomy of a screencast: Uncovering common elements and instructional strategies. *The International Review of Research in Open and Distributed Learning* 11, 3 (2010), 1–20.

[40] Clive Thompson. 2011. How Khan Academy is changing the rules of education. *Wired Magazine* 126 (2011), 1–5.

[41] Michael Tsang, George W. Fitzmaurice, Gordon Kurtenbach, Azam Khan, and Bill Buxton. 2002. Boom Chameleon: Simultaneous Capture of 3D Viewpoint, Voice and Gesture Annotations on a Spatially-aware Display. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology (UIST '02)*. ACM, New York, NY, USA, 111–120. https://doi.org/10.1145/571985.572001

[42] Dustin J Welbourne and Will J Grant. 2016. Science communication on YouTube: Factors that affect channel and video popularity. *Public Understanding of Science* 25, 6 (2016), 706–718.

[43] Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: Blending ink, speech, and gesture to support collaborative document review. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 481–490.

[44] Xun Yuan, Wei Lai, Tao Mei, Xian-Sheng Hua, Xiu-Qing Wu, and Shipeng Li. 2006. Automatic video genre categorization using hierarchical SVM. In *International Conference on Image Processing*. IEEE, Piscataway, NJ, 2905–2908.

[45] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful Classroom Deployment of a Social Document Annotation System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1883–1892.