Sifter: A Hybrid Workflow for Theme-based Video Curation at Scale

Yan Chen
University of Michigan
Ann Arbor, Michigan
yanchenm@umich.edu

Andrés Monroy-Hernández
Snap Inc.
Seattle, WA, USA
amh@snap.com

Ian Wehrman
Snap Inc.
Santa Monica, CA, USA
iwehrman@snap.com

Steve Oney
University of Michigan
Ann Arbor, MI, USA
soney@umich.edu

Walter S. Lasecki
University of Michigan
Ann Arbor, MI, USA
wlasecki@umich.edu

Rajan Vaish
Snap Inc.
Santa Monica, CA, USA
rvaish@snap.com

ABSTRACT

User-generated content platforms curate their vast repositories into thematic compilations that facilitate the discovery of high-quality material. Platforms that seek tight editorial control employ people to do this curation, but this process involves time-consuming routine tasks, such as sifting through thousands of videos. We introduce Sifter, a system that improves the curation process by combining automated techniques with a human-powered pipeline that browses, selects, and reaches an agreement on what videos to include in a compilation. We evaluated Sifter by creating 12 compilations from over 34,000 user-generated videos. Sifter was more than three times faster than dedicated curators, and its output was of comparable quality. We reflect on the challenges and opportunities introduced by Sifter to inform the design of content curation systems that need subjective human judgments of videos at scale.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Computing methodologies**;

KEYWORDS

Crowdsourcing; video processing; social media; hybrid workflow; video content analysis

ACM Reference Format:
Yan Chen, Andrés Monroy-Hernández, Ian Wehrman, Steve Oney, Walter S. Lasecki, and Rajan Vaish. 2020. Sifter: A Hybrid Workflow for Theme-based Video Curation at Scale. In *ACM International Conference on Interactive Media Experiences (IMX '20), June 17–19, 2020, Cornella, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3391614.3393657

1 INTRODUCTION

Every day, millions of people around the world create, share, and consume short videos on platforms like Snapchat, TikTok, and Douyin. These platforms use a variety of curation approaches to help their users discover high-quality and recent ("fresh") content. These approaches leverage artificial intelligence (AI), user-sourcing, or dedicated curators [12]. AI techniques rely on algorithmic aggregation and the ranking of relevant content based on metadata, such as tags [28]. These approaches are scalable but limited in their capacity to identify content attributes that require subjective assessments and nuanced cultural understanding. User-sourcing approaches rely on end-users' votes or "likes" to identify high-quality popular content, such as on Reddit [30]. These approaches are also scalable, but have the potential to silence minority opinions or to be dominated by content manipulation strategies like "brigading" [10]. Lastly, curator-based approaches rely on staff curators to identify and organize compelling content, such as on Snapchat's Discover [35] or Twitter's Moments [38] (Fig. 1). These approaches give platforms editorial control and overcome machines' inability to make subjective assessments and prevent adversarial users from manipulating content selection, but are limited by scale [12]. Specifically, it is difficult to scale curators' ability to find appropriate content from a corpus of videos that is large and rapidly growing—on Youtube, for example, over 500 hours of video content is uploaded every minute.

In this paper, we introduce *Sifter* to scale the third type of curation strategy (dedicated curators). Sifter combines automated video processing techniques and crowdsourced human expertise to provide on-demand assistance to dedicated video curators in the process of selecting and collecting content (i.e., the "select and collect" phase in Fig. 2). In this phase, curators have to rapidly browse through large "fresh-content" corpora to collect just enough raw material that might fit a coherent narrative [2, 39], or theme (e.g., "magic tricks", or the movie Lion King). As the corpora often have more appropriate (e.g., interesting, relevant) materials than needed, curators do not have to exhaust all the items.

This setting makes our problem unique, but daunting for three main reasons. First, despite being short, videos often take more time to consume and interpret than other media, like images, as they contain multimodal signals (e.g., visual, audio, text caption). As a result, curators may have to watch the videos multiple times to grasp the essence, which extends the task time. Second, the corpora often contain many unqualified videos that are distracting, further slowing curators down. Although automated video analysis techniques have become promising, machines are still limited

IMX '20, June 17–19, 2020, Cornella, Barcelona, Spain

ACM ISBN 978-1-4503-7976-2/20/06…$15.00
https://doi.org/10.1145/3391614.3393657

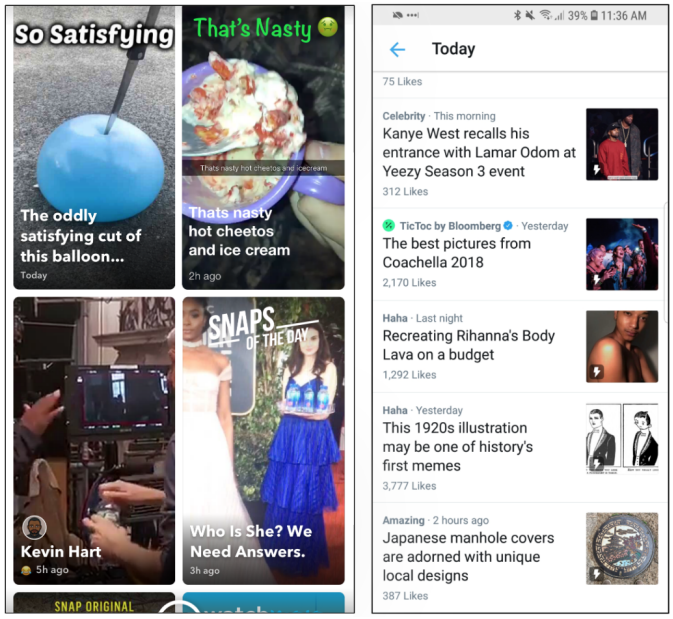

Figure 1: Screenshot from 2018 of some curated content from users' posts in Snapchat, e.g., "So Satisfying: The oddly statisfying cut of this balloon..." (left), and Twitter, e.g., "The best pictures from Coachella 2018" (right).

on assessing engaging or novel content recognition due to the social nature of content [1], and the difficulties in obtaining labelled training data [15, 16, 40]. Third, existing crowd workflows might help more accurately assess video content than machines can by leveraging human capacity, but they often require workers to reach a specified agreement level either in parallel, or through sequential refinement [7, 11, 13, 26, 27, 41]. Agreement can be difficult when the goal is to quickly extract a small set of videos from large video corpora where many are qualified, because, while they may not disagree with each others' selections, the sheer volume of content may result in workers selecting non-overlapping sets of responses.

Sifter addresses the above three challenges in the following ways. For the first challenge, we designed a custom interface for Sifter (Fig. 4) to make browsing videos more efficient. For the second challenge, Sifter automatically refines and reduces the dataset by filtering out the obviously unqualified videos (e.g., too dark, noisy) using video processing techniques. For the third challenge, we aimed to increase the overlapping sets among workers' selections while smoothing out individual workers' biases. So we developed a human-powered pipeline that further refines and reduces the output, and then draws a dozen or two qualified videos by agreement. Together, Sifter consists of a three-stage pipeline:

(1) Sifter leverages automation and video processing techniques to filter out low-quality videos from a large set.
(2) Sifter leverages human workers to rapidly select and collect thematically relevant and interesting videos.
(3) Sifter leverages separate groups of workers to make selections from the refined set of videos, and reach an agreement.

Figure 2: Common curation pipeline. We focused on the "Select and collect" step (gray).

We evaluated Sifter using publicly-available content from Snapchat's "Our Stories," which are "collections of Snaps submitted from different Snapchatters throughout the community" that are "collected and categorized to capture a place, event, or topic from different points-of-view" [36]. More specifically, we used the themes and keywords that were used to curate 12 published compilations by staff editors or by Team Snapchat [34], and we used Sifter to select and collect videos to attempt to recreate those compilations. Furthermore, we recruited three external dedicated content evaluators to assess the compilations, and we found that it took Sifter less time (μ = .71min, σ = .41min) to pick a video than the staff curators (μ = 2.35min, σ = 1.49min, $p < .0014$), and with no discernible differences in quality for eleven of the twelve compilations. Our findings aim to inform the design of systems that rely on subjective human judgments at scale.

In this research, we make the following contributions:

- The design of a human computation workflow that rapidly identifies high-quality and thematically-relevant videos from large sets of user-generated videos,
- Sifter A system that instantiates our approach and helps curators focus on creative tasks by handing off routine ones,
- An evaluation with over 34,000 videos showing that Sifter can help find videos of comparable quality to professional curators, but more quickly,
- Ethical guidelines for the adoption of Sifter.

2 RELATED WORK

Two key challenges with curation at scale are the large amount of video content, and the subjective aspects of content assessment. In this section we discuss prior relevant work and potential gaps that Sifter can fill.

2.1 Automation Techniques

One intuitive first approach for scaling up video curation is to use automated video processing techniques. Prior work in video analysis has explored methods to automatically detect activity [4], measure "interestingness" [19], identify complementary content [3], and even assess the level of creativity in a video [31]. However, these attempts to subjectively understand video content are still at an early stage and prone to algorithmic bias [5, 6].

User-generated content platforms often use simpler, but more reliable techniques to assist the video curation process, such as grouping videos by using user activity logs (e.g. clicks) and metadata (e.g., title entered by user who uploaded a video). These approaches have several limitations. First, using logs of user activity

relies on exposing behavioral analysts to video corpora in order to collect behavior data. This might not be feasible if a platform strives for tighter editorial control, and wants to shield its users from unvetted content. Second, user-generated metadata itself is not always accurate or detailed enough to understand the content of a video (e.g., videos with a caption like "Best Day").

As we describe in the next section, we chose to use some of these metadata automation techniques, but did not rely on them alone. This gave us scalability benefits without requiring us to compromise on quality.

2.2 Human-powered Video Analysis

Current video curation practices rely mainly on humans to set the criteria used for selecting high-quality content. These criteria are largely dependent on the available data and curators' tastes. By browsing the videos returned from a search query, curators constantly discover new contextual information and reshape the desired final video compilations in their mind. This complex selection model, which is confined to the curator's mind, can be difficult for even the curator to precisely articulate.

Prior work has used crowdsourcing techniques for visual analysis, but has mostly focused on object or event recognition tasks. For instance, systems like Glance and Legion:AR [24, 25] leverage the crowd to identify events in a set of long videos in real-time. Similarly, Zensors [23] and CrowdAR [32] use the crowd to help alert end-users when certain events or objects occur in a live streaming video. Shamma el al. proposed a community-supervised technique that leverages online users and machine learning for image selections [33]. We build on this previous work and shift our focus to assessing the subjective attributes of videos, such as identifying whether or not a video is interesting.

Crowds are not only called upon for object recognition tasks: prior work also explored how to enable crowds to identify interesting content in a large corpus of video data. For example, Kim et al. analyzed videos of students interacting with MOOCs, to find which content sparked confusion or engagement [21].

Similarly, Carlier et al. worked on identifying regions of interest within a video by analyzing log data that showed users' zooming interactions [8]. These studies, however, rely on user interaction data, which may not always be available.

3 SIFTER

We created Sifter to address the challenges of scale and subjectivity by combining human and machine computation. Sifter uses video processing and human computation techniques to help scale the video curation process for a given theme (second stage in Fig. 2).

The system enables staff curators to delegate the time-consuming and monotonous tasks of sifting through thousands of videos that vary in quality and selecting a small set of high-quality and thematically relevant videos. By high-quality (HQ), we mean videos that might capture viewers' attention and engagement, and by thematically-relevant (TR), we mean those that are well-suited for a collection of a particular topic.

Sifter addresses these two challenges in the following ways:

(1) **Scale.** Sifter addresses the challenge of scale by first leveraging automated video processing techniques (Table 1) to identify HQ videos. The guidelines for identifying HQ videos are derived from prior experience with videos on the platform and the existing literature (inline citations in Table 1). Then we propose a human-powered pipeline added to the automated filter. This workflow was derived from analyzing human workers' performance in our pilot studies.
(2) **Subjectivity.** Sifter addresses the need for subjective interpretation by relying on human workers to execute instructions. We evaluate this method by comparing human workers' results with those of staff curators (e.g., using curators' prior search keywords).

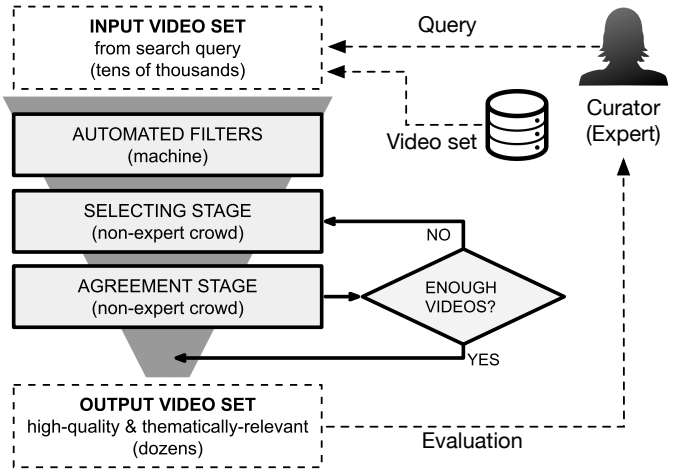

Figure 3: Sifter pipeline.

Sifter includes two parts: the sifting pipeline, and the user interface. We first explain Sifter's three-stage pipeline, and then describe our iterative user interface design process.

3.1 Sifting Pipeline

The pipeline of Sifter (Fig. 3) consists of three stages:

(1) R1: Automated filters aimed at removing trivial low-quality data for the purpose of time saving.
(2) R2: Selection stage of human-worker filters quickly refine a subset of videos.
(3) R3: Agreement stage of human-worker filters get multiple people's perspectives on the final decisions.

On a high level, the pipeline works as follows. When the search results are returned from the public Snap post database, which often contains thousands of videos, the pipeline will first automatically process these videos to filter out the low-quality videos. Then the output of R1 is a quality-refined set of videos which will then be sent to a human-powered pipeline (R2,R3) for further review in a randomized order. The human-powered pipeline can happen almost simultaneously for each stage in the long term. The method is used as follows: the videos selected from workers in R2 are pushed to R3 in real time. We assign the next group of workers to R3 when the number of videos in the R3 pool reaches a threshold determined by how long we want to keep the workers in the R3 pool. This

design helps us to streamline the final set of videos such that the staff curators can see more quickly what videos are selected.

Instead of using fixed values, Sifter's pipeline parameterizes the input and output of each step, as well as the number of needed workers. In both R2 and R3, workers use the aforementioned user interface and the number of videos and workers in the pool can vary. In our final evaluation, the values we used for the parameters were derived from our pilot studies, which were conditioned on the case that we want Sifter to generate 10-20 refined videos for each compilation (customized for the platform). This design process and the structure of the Sifter pipeline are generalizable as we evaluate it with a large number of videos. In this section, we discuss the pipeline design and implementation.

3.1.1 R1: Automated Filters. R1 includes four automated filters (Table 1) which take a set of a few thousand videos retrieved from a public Snap posts database as input, and outputs a refined set of videos by filtering out hundreds of unqualified videos. We implemented these filters by using video properties (e.g. duration) and the OpenCV library (https://opencv.org/). Table 1 presents the implementation details.

3.1.2 R2: Selection stage. In this stage, a group of workers review the output videos from R1 and identify a small set of them that are high-quality and thematically relevant. The number of needed workers for each compilation is a parameter that depends on the number of output videos from R1. Based on a set of pilot studies, we found that workers perform optimally when reviewing up to around 1,000 videos and are asked to select up to about 100. We derived these threshold values based on observation of quality and workers' efficiency (duration of selection process). Other researchers could derive suitable values for these parameters by exploring the trade-offs among quality, cost, and time based on their own needs.

3.1.3 R3: Agreement stage. In R3, a different group of workers is given the same UI but with the output videos from R2. The purpose of this stage is to smooth out the differences of how people interpreted the instructions and performed the task in R1.

Two workers were assigned per compilation in this final evaluation stage. The resulting output is a set of videos that both workers in R3 selected. In other words, the videos that have unanimous consent among all three workers from R2 and R3 are included in the output.

We designed R3 with two considerations. First, the quality of the input videos (from R2) is higher than those in R2 (from R1). With the assumption that a video with higher quality would require longer attention to review, we decided to have workers select fewer videos than in R2. Second, our pilot study results suggested the agreement rate is often about 40–50% between two workers when the number of selected videos is 30 and the total is 100. This relatively low level of agreement is due to the fact that there are more qualified videos than needed. Thus, we decided to match those numbers given that the final number of videos selected is 10. Again, other researchers may find different values for these parameters if the number of needed videos is different.

3.2 User Interface

The user interface (UI) we designed for the human-powered stages of Sifter is intended to give concrete instructions that are readily interpretable and enable the completion of open-ended tasks, i.e., find HQ and TR videos. Designing a user interface that facilitates efficient video selection is a non-trivial process. The large number of permutations of UI parameters (e.g., video playback speed, number of videos per page, autoplay on/off, sound on/off, etc.) makes it challenging to find and test a single UI for efficient video sifting. Pavel et al. developed a video review tool aimed to help with the video review process [29]. However, this tool focused on frame-by-frame editing of a single long video, rather than a large number of small videos. We addressed these concerns by first conducting a series of small studies to compare the outcome (in terms of speed) of different combinations of the major UI elements and of variants of their parameters. Inspired by prior work [22], we implemented a unique time-enforced interface that provided video auto-looping (max 10 seconds) of all eight videos on the task page at once, enabling workers to make rapid decisions.

We iteratively designed Sifter with two considerations in mind: the UI should be easy to use with a minimal learning curve (for speed), and workers should have enough context during the task to make a clear judgment (for quality). These features emerged from observing workers using the system during formative studies and a series of user studies where we tested each UI component one at a time. The final interface consisted of a landing page with instructions and a task page with five main components:

3.2.1 Landing page with example-based instructions. Before a worker starts executing the task, they see a landing page where they get instructions (e.g., Please QUICKLY select ## videos that satisfy all the requirements.) and example videos from a previously published compilation (randomly selected). To make sure workers understood the task details, such as what HQ and TR videos are, we first tested their comprehension by providing different levels of contextual information in the instructions (see below a, b, c) and measured the quality of the outcome (rated by researchers) and the completion time. We found that providing the goal of the task (a, b) and information about how the videos were found (c) increased the quality of selected videos without increasing the task time.

3.2.2 Task page. The component index is corresponding to the numbers in Figure 4. **1. Contextualized instructions.** These instructions reiterate what was presented on the landing page, but without the example videos. **2. Progress bar.** We used a progress bar to show workers how many videos are still needed, how many are left in the pool, and how many they have selected. **3. No scrolling.** To design a user interface that lets workers rapidly sift through videos, we first implemented a web interface with all videos on one page and asked workers to select interesting videos. From follow-up interviews with workers, we found that displaying all the videos on a single page is inefficient because workers would forget what videos they had reviewed already as they scrolled up and down the page. Thus we designed a layout to display as many videos as we could per page while avoiding having workers scroll. **4. Looping videos and audio on mouse over.** To enable fast visual scanning of the videos, each task page was populated with eight

Filter	Sifter Implementation	Motivation
Shorter than 3s	Use video duration property	Avg. minimum duration from published compilations
Small pixel differences between frames	Use systematic sampling to extract five frames and compute their pixel differences in the center 200px X 200px (most important part), if more than three frames have a difference of less than 1,000 we remove the video).	Prior work showed that videos with motion of objects (e.g., human) can be more engaging than static display (e.g., text) [17].
Low aesthetics score	Compute the average colorfulness scores [18] for the same five frames and remove those videos with a score that is below a threshold we derived from a training data set of published compilations.	Prior work showed that video aesthetics impact engagement [9] and that color in particular is highly correlated with aesthetics.
From same session	Use video metadata to keep only one posted video from a user (we picked the first in our experiments) and eliminate the rest that are posted within the next 120s. More advanced techniques can be applied to select the best one.	Using multiple videos from the same scene and the same person would reduce the diversity of the compilation which has been shown to lower the engagement.

Table 1: A list of automated filters, how we implemented them in Sifter, and the motivations of using them.

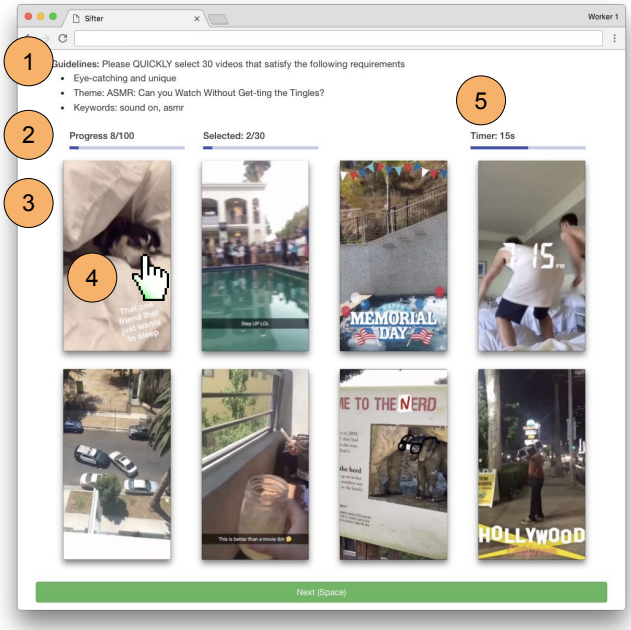

Figure 4: Sifter UI task page. Workers can see the instructions, their progress (number of videos they have seen and selected), and a timer on every page. All videos autoplay silently when workers arrive on the page; they can mouse over videos to turn on audio, and click to select.

looping videos. These videos were muted, however, workers could move their mouse over any video to trigger its audio. This approach helped workers to rapidly go through a large corpus of videos. Furthermore, we experimented with using keyboard shortcuts to play, pause, and select videos, but we found that workers were faster with the mouse-based approach. Also, using the mouse resulted in more videos being mouse hovered (reviewed) and selected. We also experimented with different video preview speeds but found no difference in execution time compared to normal video speed.

5. Timer. The sifting task is such that it is not necessary to select every good video, but only a small set of them. Additionally, as one of our goals is to speed up the process, we set a 30-second limit on each page to prevent workers from getting stuck watching videos in great detail.

4 EVALUATION

Sifter was meant to scale subjective human judgment. Also, we wanted Sifter to perform fast and reliably. To this end, we compared the quality of the videos that Sifter generated to those generated by dedicated curators. We report the details for each compilation we evaluated in Table 2.

4.1 Dataset

To effectively validate the video quality, we selected 12 compilations that were previously created and published by staff curators by the partner platform. We put together these compilations with the following guidelines:

(1) All videos were in English to make it easier for the researchers to run the evaluation.
(2) Each compilation had more than 1,000 videos to curate.
(3) The videos represented a wide range of themes from international events to evergreen topics. For example, "ASMR," which contained videos with soothing sound, or "That's Nasty" with videos of showing a ruined ice cream.
(4) The topics of the videos were globally recognizable, i.e., we avoided local news.

For each compilation we gathered (a) the number of staff curators that worked on putting together those compilations, (b) the name or theme of the compilation, (c) the keywords used to find all the videos that were considered for the compilations, and (d) the time spent searching for and collecting the videos. Table 2 lists (b) and (c); Table 3 shows the aggregate of (a) and (d).

4.2 Workers and Evaluators

We recruited both the workers and evaluators from an online freelancing platform that enables workers to set their own rates. Although the platform allows requesters to bargain, we took the rate proposed by the workers at face value.

Compilation	**Theme**	**Keywords**	**Input videos**	**Videos after automated filters**	**Workers selection stage**	**Workers agreement stage**
C1	ASMR: Can You Watch Without Getting the Tingles?	sound on, asmr	1,922	329	1	2
C2	Black Panther: Does It Live Up to the Hype?	black panther	1,984	586	1	2
C3	Conspiracy Theories: 12 Videos That'll Make You Believe	ufo, ghost, conspiracy, alien	3,262	1,262	2	2
C4	Flashback Feels: Gone But Never Forgotten	flashback, throwback, 80s, 90s, 2000s	2,118	980	1	2
C5	Magic Wins: These Weird Tricks Will Fool You	magic, tricks	3375	1,508	2	2
C6	Fun Moms: They've Gone Wild... But We're Here For It	mom, mother, ma	2,741	1,185	2	2
C7	That's Nasty: 10 Videos That'll Make Your Skin Crawl	gross, disgusting, ew	3,204	974	1	2
C8	So Over It: Cue the Eye Roll 🙄	so over it, i'm done, not amused, ugh, bummer	3798	1,670	2	2
C9	Happy St. Patrick's: Are You Ready to Shamrock & Roll?	patricks's, patrick	3,121	1,933	2	2
C10	School's Out For Snow: The Weather's Got Us Wild	snow, school, campus, canceled	847	426	1	2
C11	Streaks: 🔥🔥🔥🔥🔥	streaks	5,490	1,672	2	2
C12	Weddings: These Brides & Grooms Are the Real MVPs	wedding, weddings, brides, grooms	2,206	1,390	2	2

Table 2: List of compilations used for evaluating Sifter. All of these compilations had been previously published. Keywords were the words used by the platform's curators to find the videos to create the compilations. The number of videos retrieved using those keywords for the evaluation of Sifter is in the fourth column. The fifth column presents the number of videos left after the automated filter step. The last two columns present the number of workers involved in each stage of filtering.

- Workers. We recruited nine human workers (two female, seven male) whose self-reported expertise was "data entry." Workers came from Europe, Asia, and North America. Any worker who applied to work on the task was accepted on a "first come, first serve" basis. The last two columns in Table 2 report the number of workers involved per stage.
- Evaluators. We recruited three professional video producers (one female, two male) to evaluate the quality of the final videos. They came from North America and Asia. These evaluators were chosen because they all had prior experience with the process for Snap video curation, but none of them were familiar with the compilations we selected.

4.3 Procedure

Once the workers received the task, they first read the aforementioned task instructions (e.g., name, example videos). Based on how many videos have been selected in the human worker pipeline, the workers were assigned to either the "selection" or the "agreement" stage in the pipeline. After completing their tasks, workers filled out a survey about their familiarity with the topic of the compilation, the challenges they faced, and their selection strategy.

5 RESULTS AND DISCUSSION

Overall, Sifter is faster than the staff curators at generating a refined set of videos, and the quality of these videos is comparable to those identified by the curators. In this section, we discuss details of our study results.

5.1 Sifter is three times faster than curators.

We computed the average time spent per compilation for Sifter by adding up the completion times of the two workers who took longer to finish R2 and R3 stages (Eq.1). This is because the workers from the same stage can perform tasks in parallel. The fourth column in Table 3 reports the average time spent per compilation between two methods. With that, we computed the average human-time spent per video selection between the two methods (last column in Table 3). We found that Sifter ($\mu = 0.71$min, $\sigma = 0.41$min) can pick a video three times faster than curators ($\mu = 2.35$min, $\sigma = 1.49$min) ($p < .0014$).

$$\begin{aligned} \text{Sifter}_T &= \max(t_{\text{worker1,selection}}, t_{\text{worker2,selection}}) \\ &\quad + \max(t_{\text{worker1,agreement}}, t_{\text{worker2,agreement}}) \end{aligned} \tag{1}$$

To determine the time that the curators spent sifting each compilation, we used the conventional timeout cutoff technique to determine their query session and then added up all the query sessions per compilation [20]. We used 30 minutes as the timeout threshold, meaning that if the next search request happened more than 30 minutes after the current one we counted it as a new query session. The sum of all the session time is the final time spent per compilation. Additionally, because Sifter works asynchronously, curators could perform multiple compilation sifting tasks simultaneously.

	Avg. workers (s.d.)	Avg. generated videos (s.d.)	Avg. time spent (s.d.) per compilation	Avg. time spent (s.d.) per video
Sifter	3.58 (0.51)	21.08 (8.36)	13.55 min (5.50)	0.71 min (0.41)
Curator	2.58 (1.56)	120.67 (91.02)	259.79 min (252.82)	2.35 min (1.49)

Table 3: Comparing the production of workers using Sifter and professional curators. The first column reports the average number of people involved per compilation. The second column presents the average number of videos generated per compilation. The third column reports the time spent per compilation.

5.2 No quality difference for 11 of the 12 compilations.

We evaluated Sifter's quality by comparing its output against the output of the staff curators when performing the same "selection and collection" stage that Sifter aimed to replace, (step 2 in the pipeline from Fig. 2). We did not inform evaluators where the videos came from, so they did not know whether it came from Sifter or the staff curators. We were interested in measuring how relevant the videos output by Sifter were to the topic of the compilation. For example, for compilation C5, we wanted to know how relevant the videos selected by Sifter were to the topic of "Magic Wins: these weird tricks will fool you."

For each compilation, we calculated a rating for a sample of videos output by Sifter, and another for the ones output by the staff curators. We hired raters to evaluate each video in the sample on a 5-point Likert scale for "how relevant is this video for the topic 'insert topic'?" from 1 (not relevant at all), to 5 (very relevant).

To reduce the biases resulting from the raters' different background knowledge (according to their feedback), we also added a baseline condition that consists of a sample of randomly selected videos. These videos were retrieved from the corpus using the keywords described in Table 3, e.g., 10 random videos out of the 1,984 videos that were collected for compilation C2. Then we measured Sifter's and the staff curators' ratings relative to the baseline rating.

The rating for each video in the Sifter sample was calculated by subtracting the average of the baseline ratings from the rating given by the rater to that video. We then calculated the average of all of the individual ratings in Sifter and used that as the rating for Sifter for that compilation, e.g., 0.93 for compilation C5 in Figure 5. In this way, we took into account the individual raters' differences in perception and were able to make the difference comparison to determine the effectiveness of Sifter. We conducted 12 comparisons using two-tailed, paired-samples t-test, and with Bonferroni correction, we considered the comparison result significant if the p–value was below .05/12 = .0042. We found that the ratings for eleven out of the twelve compilations generated by Sifter showed no significant differences ($p > .0042$); the other one compilation, C3, were rated significantly lower ($p < .0042$). We analyzed the reasons in detail in a later section.

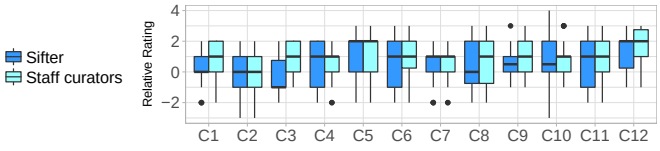

Figure 5: A box plot comparing Sifter with staff curators using a relative rating measurement. For C3, the curator's ratings were significantly different from Sifter's ($p < .0042$).

5.3 Workers use different strategies.

In order to improve the automation process in the future, we wanted to understand what strategies workers used when selecting videos. We then gave workers a questionnaire when they finished the selection task of each compilation. The questionnaire contained free-text questions about their strategy such as "what was your strategy for completing this task?" We then coded their answers with one of a set of five categories that we came up with through an inductive approach inspired by the text instructions.

Figure 6 shows how often each strategy was used. We found that for more than half of the compilations (7/12) workers reported using at least two strategies, and two of them had workers report using only one (C2, C4).

A relevance-centric strategy means the worker reported focusing primarily on identifying videos that were relevant to the theme of the compilation. For instance, a worker mentioned "I was looking for videos that match given theme and move on" (P33). A quality-centric strategy means that the worker reported having focused on identifying "eye-catching" or "interesting" videos. For example, a worker reported "I just see which video is most interesting, look good, have a magic or some funny or eye catching things in it" (P16). We also found that some workers reported applying both strategies equally, e.g. "Watching eye-catching videos and select them if they match given theme" (P40), and others reported applying one strategy before the other, e.g. "I was looking for interesting videos as fast as I could and then I was making selection if they are right match" (P14).

6 LIMITATIONS AND FUTURE WORK

6.1 Parameter values in the pipeline

One of our contributions is the design of Sifter's pipeline. However, the parameter values we derived for the final evaluation were based

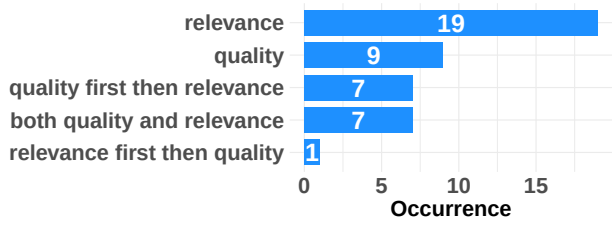

Figure 6: Histogram of the popularity of different strategies among human workers.

on our need to form compilations with 10 to 20 videos. Future users of Sifter's pipeline would need to find their own optimal values parameter for their data and scenarios. Future work should also explore the dynamics of the parameter values through additional controlled studies.

6.2 Evaluating Automated Filter

Because we designed our automated filters (R1) using previously published and evaluated heuristics, we did not evaluate their performance as part of this paper. However, future work would benefit from a re-evaluation of these processes for each new context.

6.3 Worker Biases

Although we strove to understand what is "universally" relevant by "averaging out" workers' biases, Sifter still serves as only a proxy and might not always perform as expected. For example, there was one compilation, C3, in which no worker reported unfamiliarity with the theme, but Sifter still generated videos that were given lower quality ratings than the curators' compilation. We analyzed the outcome and found that the majority of videos in the set were selfies with an alien animation, which was briefly a popular camera effect. Although the content is relevant to one of the keywords in the query 'alien', it is not considered to be a high-quality video based on our definition, and also made the majority of the selection look similar in style. A worker in T8 commented that:

> *[The] main challenge was to not pick all the same videos, cause [sic] there are many similar videos.*

We further analyzed the comments given by workers for this compilation. One worker from the agreement stage reported that *"There isn't that many eye-catching videos ..."*. As we asked workers in the agreement stage to select at least 30 videos, they might have selected some with low quality because of this requirement. One way to address this is to let workers self-report low quality video batches, especially in the agreement stage. We could do this by building on prior research done on quality control [14].

6.4 From Curation to Moderation

Our primary focus in this work was on video curation of interesting content; however, our approach is also promising for moderating inappropriate videos. For example, workers could potentially rapidly identify and remove videos of kids bullying or being engaged in violent activities. In the future, Sifter has the potential to delight the users of online social media platforms with safe content.

6.5 Opaque Strategies

We have made initial attempts at trying to figure out what strategies curators and workers used for successful sifting. However, a more formal field study with curators to explore how their strategies change for different compilation themes could offer valuable insights. For example, examining the portions of the video that people watched and the interaction patterns people have with videos might provide useful information. In addition, we focused on the sifting task in this work as the first step. Future work can expand our approach to other steps, such as leveraging human workers for complex queries, to further scale the curation process.

7 ETHICAL RECOMMENDATIONS

We believe the Sifter approach has significant potential for being widely adopted, so it is important for us to ensure that designers who build upon this work use it ethically. During the curation process, we expect curators to have their own control in the workflow—e.g., taking breaks when needed, driving their own work forward, and selecting videos that are potentially interesting to them. To ensure they are treated ethically and responsively, we include the following recommendations on how to appropriately deploy Sifter:

- **Compensation.** Workers should be paid a fair rate [37].
- **Sessions.** Session lengths should be capped, and breaks should be compensated to care for workers' mental and physical health.
- **Choice.** Workers' sensibilities and personal preferences should determine which topics they curate. We envision a scenario where workers get to see a list of titles and descriptions of stories they can curate, giving them the power to select which ones they work on. It is also important to understand their familiarity with the topics, as it would make them comfortable pursuing a task and improve the quality of work.
- **Transparency.** End-users who consume or produce content curated by a hybrid process like Sifter should be informed of the process by which the stories are curated.

8 CONCLUSION

In this paper, we introduced Sifter, a system that utilizes automation and workers to enable curators to delegate the task of selecting eye-catching and thematically-relevant videos, allowing them to focus on more creative tasks. Sifter first leverages video processing techniques to remove unqualified videos, and then uses a human-powered pipeline that allows workers to rapidly browse, select, and reach an agreement on videos.

We evaluated Sifter by creating 12 different video compilations, and found that the quality of the majority of those compilations was indistinguishable from the ones created by staff curators. We believe that our findings can inform the design of systems in the future that rely on subjective human judgments at scale.

9 ACKNOWLEDGEMENTS

Special thanks to our colleagues Maarten Bos, Kelly Mack, and Aletta Hiemstra for their feedback.

REFERENCES

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Georgios Askalidis and Greg Stoddard. 2013. A theoretical analysis of crowdsourced content curation. In *The 3rd Workshop on Social Computing and User Generated Content*. ACM.

[3] Werner Bailer, Martin Winter, and Stefanie Wechtitsch. 2017. Learning selection of user generated event videos. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. 1–7.

[4] Sunil Bandla and Kristen Grauman. 2013. Active learning of an action detector from untrimmed videos. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 1833–1840.

[5] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. *Special Interest Group for Computing, Information and Society (SIGCIS)* (2017).

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. In *Conference on Neural Information Processing Systems, Long Beach, CA*.

[7] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

[8] Axel Carlier, Vincent Charvillat, Wei Tsang Ooi, Romulus Grigoras, and Geraldine Morin. 2010. Crowdsourced automatic zoom and scroll for video retargeting. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 201–210.

[9] Peter McFaul Chapman. 1997. *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. Ph.D. Dissertation. University of Waterloo.

[10] Wallace Chipidza. 2016. Negative Behaviors in Online Communities. (2016).

[11] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2382–2393.

[12] Edward Curry, Andre Freitas, and Sean O'Riáin. 2010. *The Role of Community-Driven Data Curation for Enterprises*. Springer US, Boston, MA, 25–47. https://doi.org/10.1007/978-1-4419-7665-9_2

[13] Peng Dai, Daniel Sabby Weld, et al. 2011. Artificial intelligence for artificial artificial intelligence. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[14] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.

[15] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. 2015. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE transactions on pattern analysis and machine intelligence* 38, 3 (2015), 563–577.

[16] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. 2018. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine* 35, 1 (2018), 112–125.

[17] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: an empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 41–50.

[18] David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. International Society for Optics and Photonics, 87–96.

[19] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. 2013. Understanding and Predicting Interestingness of Videos.. In *AAAI*.

[20] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 699–708.

[21] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 31–40.

[22] Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 3167–3179.

[23] Gierad Laput, Walter S Lasecki, Jason Wiese, Robert Xiao, Jeffrey P Bigham, and Chris Harrison. 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1935–1944.

[24] Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 551–562.

[25] Walter S Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P Bigham. 2013. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1203–1212.

[26] David Merritt, Jasmine Jones, Mark S Ackerman, and Walter S Lasecki. 2017. Kurator: Using The Crowd to Help Families With Personal Curation Tasks.. In *CSCW*. 1835–1849.

[27] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2019. Photo sleuth: combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 547–557.

[28] Alessandro Ortis, Giovanni Maria Farinella, Valeria D'amico, Luca Addesso, Giovanni Torrisi, and Sebastiano Battiato. 2015. RECfusion: Automatic Video Curation Driven by Visual Content Popularity. In *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane, Australia) *(MM '15)*. ACM, New York, NY, USA, 1179–1182. https://doi.org/10.1145/2733373.2806311

[29] 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, 517–528.

[30] Reddit. 2019. Reddit Video. https://www.reddit.com/r/videos/.

[31] Miriam Redi, Neil OHare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 4272–4279.

[32] Elliot Salisbury, Sebastian Stein, and Sarvapali Ramchurn. 2015. Crowdar: augmenting live video with a real-time crowd. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[33] David A Shamma, Lyndon Kennedy, Jia Li, Bart Thomee, Haojian Jin, and Jeff Yuan. 2016. Finding weather photos: Community-supervised methods for editorial curation of online sources. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 86–96.

[34] Snap. 2018. Brand Safety. https://support.snapchat.com/en-US/a/brand-safety.

[35] Snap. 2018. Discover. https://support.snapchat.com/en-US/a/discover.

[36] Snap. 2018. Our Story. https://support.snapchat.com/en-US/a/live-story.

[37] Stanford. 2019. Fair Work. https://fairwork.stanford.edu/.

[38] Twitter. 2011. Twitter Moments guidelines and principles. https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles.

[39] Annika Wolff and Paul Mulholland. 2013. Curation, curation, curation. In *Proceedings of the 3rd Narrative and Hypertext Workshop*. ACM, 1.

[40] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2018. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision* 126, 2-4 (2018), 375–389.

[41] Amy X Zhang, Jilin Chen, Wei Chai, Jinjun Xu, Lichan Hong, and Ed Chi. 2018. Evaluation and refinement of clustered search results with the crowd. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 14.