Generating Audio-Visual Slideshows from Text Articles Using Word Concreteness

Mackenzie Leake[1], **Hijung Valentina Shin**[2], **Joy O. Kim**[2], **Maneesh Agrawala**[1]
[1]Stanford University, [2]Adobe Research
{mleake, maneesh}@cs.stanford.edu, {vshin, joykim}@adobe.com

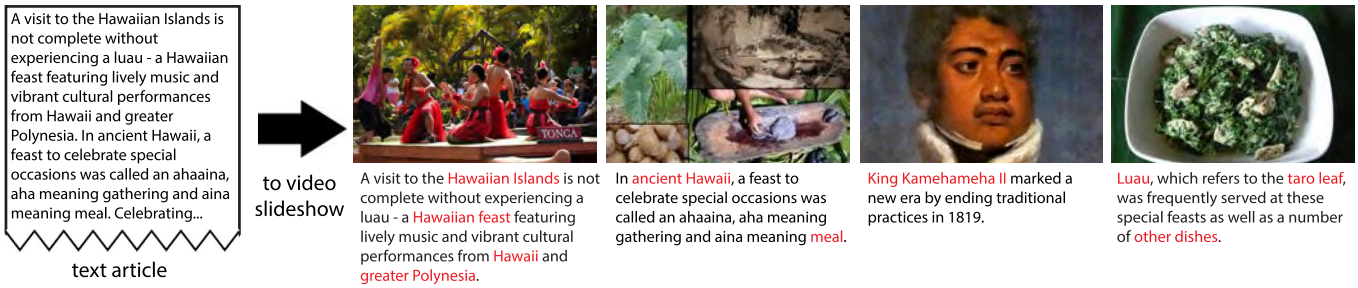

Figure 1. We present a method for automatically generating audio-visual slideshows from a text article by identifying representative *concrete* phrases from the text and searching for visuals that match these words. Left: Original text article. Right: Visuals selected by our system based on the concrete words (red) in each sentence. The text is automatically turned into voiceover speech, and the visuals are timed to appear whenever the first concrete word appears in a sentence.

ABSTRACT

We present a system that automatically transforms text articles into audio-visual slideshows by leveraging the notion of word concreteness, which measures how strongly a word or phrase is related to some perceptible concept. In a formative study we learn that people not only prefer such audio-visual slideshows but find that the content is easier to understand compared to text articles or text articles augmented with images. We use word concreteness to select search terms and find images relevant to the text. Then, based on the distribution of concrete words and the grammatical structure of an article, we time-align selected images with audio narration obtained through text-to-speech to produce audio-visual slideshows. In a user evaluation we find that our concreteness-based algorithm selects images that are highly relevant to the text. The quality of our slideshows is comparable to slideshows produced manually using standard video editing tools, and people strongly prefer our slideshows to those generated using a simple keyword-search based approach.

CCS Concepts

•**Information systems** → **Multimedia content creation;**

Author Keywords

Audio-visual slideshows; Text-to-video; Word concreteness

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
https://doi.org/10.1145/3313831.3376519

INTRODUCTION

Effective writing doesn't just tell a story; it paints a picture. It draws on familiar visual concepts to communicate imagery through text. In psycholinguistics, this evocative quality of language is often discussed in terms of *concreteness*, which measures how strongly a word or phrase is related to some perceptible concept. For example, the word "ladybug" is considered relatively concrete—it elicits the image of a small red insect, decorated in a distinct pattern of black dots. In contrast, the word "reassurance" is not very concrete; one may have a firm grasp of its meaning but no clear strategy for conveying it in a game of Pictionary.

Psycholinguists have tied the concreteness of language to the memorability of text, highlighting the important role that it plays in effective writing [20, 40]. However, for concrete language to be effective, it needs to refer to concepts that are familiar to the reader. This can pose a challenge: how is an author to know, for instance, whether readers have ever seen a ladybug before? One strategy is to augment text with images.

Visual content can be used to enhance written information in many ways: photos emphasize salient moments in news stories, maps illustrate directions, film brings scripts to life, and diagrams make it easier to understand educational articles. Audio-visual media can further support engagement with content by displaying images organized in time with specific portions of the text and can aid in longer term recall of details and emotional engagement with content [31, 8]. Through a formative study in which we compare informational articles presented in three different formats (text-only, text with images, and audio-visual slideshow), we find that people not only prefer audio-visual slideshows but also find the content easier to understand compared to the other two formats. Unfortu-

nately, creating such audio-visual multimedia content requires authors to work with imagery and sound, which can demand significant effort, time, and skill.

To address this problem, we present a system that leverages concreteness in written language to transform informational articles into audio-visual slideshows automatically (Fig. 1). The input to our system is an existing text article. We obtain imagery relevant to each sentence in the text by analyzing its grammatical structure, identifying the most concrete words, and then using these words as search terms to find relevant images from image search engines. We then apply text-to-speech [4] to convert the article into a voiceover and time-align the imagery to appear at the appropriate point within the resulting speech.

We demonstrate the effectiveness of our approach by generating audio-visual slideshows (ranging from 20 to 131 seconds) for short text articles (ranging from 3 to 23 sentences) from a variety of domains, including Wikipedia articles and pop-science articles. For these articles it takes between 2 and 10 minutes to generate audio-visual slideshows using our approach, with most of this time dominated by downloading images and rendering the slideshow with zoom and pan effects. In a user evaluation we find that our concreteness-based algorithm selects images that are highly relevant to the text. The quality of our slideshows is comparable to slideshows produced manually using standard video editing tools, and people strongly prefer our slideshows to those generated using a simple keyword-search based approach.

RELATED WORK

We build on two major areas of related work: (1) text-based video editing tools and (2) automatic visualization of text.

Text-based video editing tools

There are a number of commercially available tools that transform text into videos. These range from fully automatic workflows, such as Article Video Robot [1], which takes a web URL of a text article as input and outputs a slideshow containing templated graphics and images from the original text article, to hybrid tools like Wibbitz [6] and Magisto [5], which automatically arrange and edit user-provided video clips and images. These tools use generic clip art and animation or require the user to provide visual content for the text. In contrast, we take a fully automatic approach and obtain relevant images from the web and edit them into an audio-visual slideshow.

Instead of creating new videos, some tools leverage text documents to make the task of browsing and navigating existing videos easier. This makes possible formats, such as video digests [42], that provide skimmable text summaries of narration-heavy videos. Similarly, Visual Transcripts [46] automatically generate readable notes from lecture videos and their corresponding transcript. SceneSkim [41] aligns multiple documents, such as captions, scripts, and plot synopses, to make it easier to browse and search movie content by a broad set of metadata.

Text-based tools can also facilitate video editing. Some of these tools make use of metadata, such as user annotations, sentiment, and camera shot type, to enable authors to match video clips to the appropriate points in scripted dialogue [32] or voiceover narration [51]. Other work has explored letting users edit talking-head video by performing cut, copy, paste, and more recently, insert operations, on the text transcript to edit video [11, 21]. Some tools also analyze the input video to identify where, for example, the speaker is quiet or still and visualize this information in the transcript text to help users manually segment the video for editing [14, 49]. Another approach is to create scaffolded, template-based systems with less automation and more user involvement to help users create videos [29, 9]. Rather than selecting or editing salient portions of visual content already provided by the user, we propose a system that gathers new visual content for generating an audio-visual slideshow from text. We also use plain text that does not require structure or human annotations. Also closely related to our work, B-Script [25] helps video editors insert supplementary b-roll content into talking head videos using their transcript. B-Script uses a mixed initiative approach, in which the system suggests keywords in the transcript that the user can use to query and select the specific b-roll to insert. Our approach is fully automatic both for selecting the images and time-aligning them to the audio narration.

Automatic text visualization

Prior work has shown that adding visual media to text can facilitate better understanding of the text content in many ways. For example, combining images and text can disambiguate between different interpretations of the text and highlight contradictions through the use of juxtaposition [10].

Prior work has focused on techniques for using images to summarize or enhance text for different applications. One area of work in using images for summarization focuses on synthesizing a single summary picture for a text using multiple images [60]. This method defines the notion of "picturability" as the extent to which a word can be drawn or a good image can be found to represent the word. This work trains a model to predict word picturability based on a small labeled dataset (500 words). We use the notion of "concreteness" defined on a much larger database (40K words), and we aim to illustrate the entire content of the text rather than its summary. Combining text and images has been used to generate news summaries by computing the image saliency and relevance scores for concepts, pictures, and sentences together [34]. Similar tools for enriching textbooks with images have also been developed by extracting and illustrating the main concepts from each section of the book [7]. Multimodal summaries for complex sentences have been generated by identifying main entities in Wikipedia articles about people and events and finding relevant imagery within Wikipedia [53]. Instead of extractively summarizing an article, our approach augments the entire text with images from the web. Many prior approaches for combining text and images focus on static outputs (i.e., text combined with photos), whereas we generate audio-visual slideshows.

Also closely related, Videolization [27] takes Wikipedia articles as input and produces videos based on knowledge graphs. This system is specific to Wikipedia and requires the input

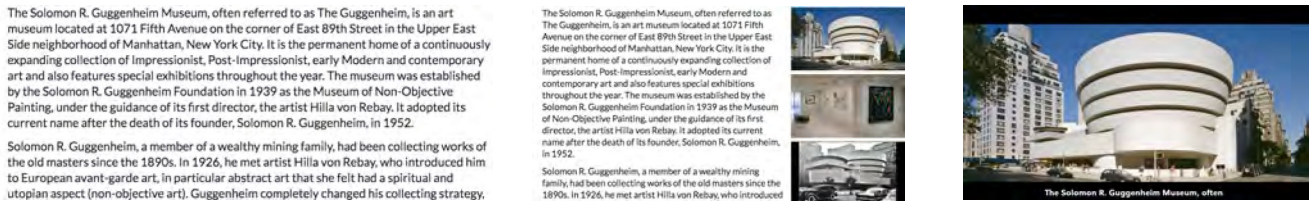

Figure 2. Three formats of the "Solomon R. Guggenheim Museum" example. For the text format breaks occur at each paragraph. For the text and images format images accompany paragraphs of text. The slideshow is a 2-minute narrated video.

to adhere to the article structure enforced by Wikipedia. In contrast, our system leverages linguistic structure to generate the output and thus can construct a slideshow from a wider variety of unstructured text documents. Our system analyzes source text documents for psycholinguistic attributes (e.g., concreteness) and queries images from the web to illustrate each sentence in the text. Although we focus on informational texts, our approach is general and can be applied to a wide range of texts. Our work focuses on general text articles and creates full slideshows from text without additional user input.

FORMATIVE STUDY

Prior work on the effectiveness of multimedia content provides various guidelines about when and how to use different media formats, such as text, text with images, video, and animation, to best convey information [8, 10, 31, 36, 37, 52]. However, prior work provides mixed evidence about the advantages (e.g., memorability, learning, engagement) and disadvantages (e.g., distraction, confusion, improper pacing) of each format. Therefore, we conducted a separate formative study to learn how the presentation format of informational articles impacts a viewer's understanding and preferences.

We compared three formats: (1) text-only, (2) text with images, and (3) audio-visual slideshows (Figure 2). We collected text excerpts (200-350 words in length) from three informational articles published in Wikipedia and How Stuff Works [58, 54, 15], covering topics from different areas (i.e., food, place, health) that could be both entertaining and educational to a broad audience: *Solomon R. Guggenheim Museum* [58], *Jams, Jellies, and Preserves* [54], and *Weighted Blankets* [15].

The text only format simply contains the text excerpt from the article. To create the audio-visual slideshow format, we recruited 15 participants on UserTesting.com who indicated they knew how to use video editing software. We provided them with audio narration of the text excerpts, which was generated using a text-to-speech service [4], and asked them to create slideshows by incorporating relevant images from publicly available image search engines. For our study we selected the best example on each topic to use as the manual comparison for that topic. To generate the text with images format, we used the images that participants chose for the slideshows, and placed them in a column to the right of the text (Figure 2, middle).

Finally, to evaluate the three formats, we recruited 120 participants on Amazon Mechanical Turk. Each participant saw all three article formats, but with a different topic for each article.

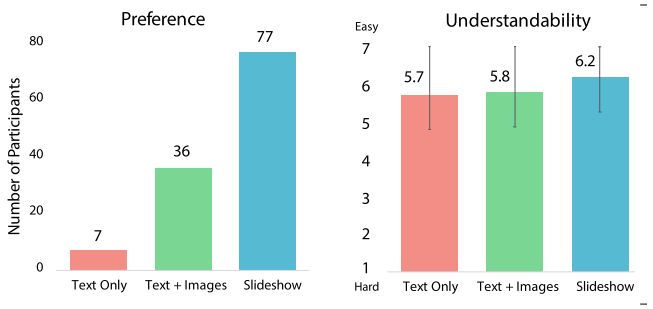

Figure 3. Results from the formative study. Viewers preferred the slideshow format over text-only and text with images. Viewers also found content presented in slideshows easiest to understand.

We randomly assigned the ordering of the formats and topics across participants. We asked participants about preference, understandability and recall of facts in the article.

Figure 3 summarizes the result of our formative study. The majority of the participants preferred the slideshow format over the text with images or text-only format. When we asked participants in which of the three formats they would like to consume informational articles, 77 (of 120) chose slideshow, compared to 36 who chose text with images and 7 who chose text-only. The most common reasons for favoring slideshows were ease of viewing and having multiple sources of information (i.e., audio, text, and images) concurrently. Some viewers also commented that slideshows were entertaining to watch. Those who favored text with images liked being able to set their own pace and being able to go back and forth between different parts. Those who favored the text-only version liked the simplicity of the format and did not feel the need for supplementary images for the given topics.

As shown in Figure 3b, viewers also rated slideshows easiest to understand (M = 6.2, SD = 1.0), followed by text with images (M = 5.8, SD = 1.3), and then text-only (M = 5.7, SD = 1.5). The differences were statistically significant ($p < 0.01$). Viewers rated the images presented in both the text with images format and slideshows highly relevant to the content (M = 5.2 and 6.0, respectively, on a 7 point Likert scale).

These results suggest that audio-visual slideshows can make information easier and more entertaining to consume. People also find audio-visual slideshows aid their understanding of the content, especially when the images are relevant. We use these study findings to motivate our automatic algorithm that generates high-quality audio-visual slideshows from text articles.

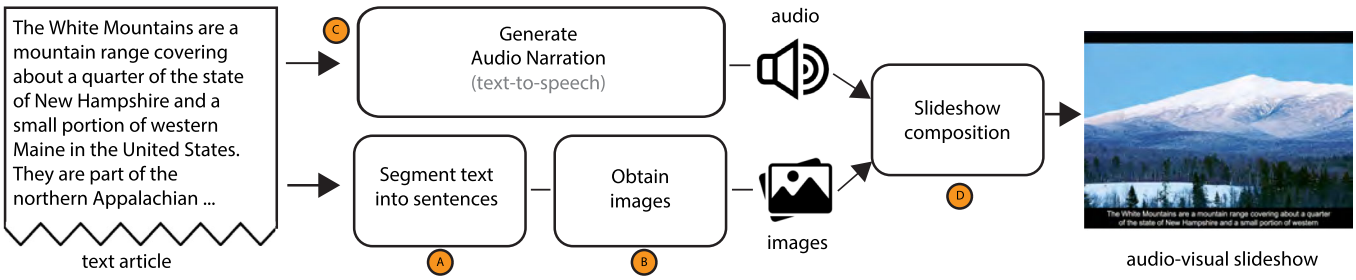

Figure 4. Our system takes a text article as input and generates an audio-visual slideshow. (Step A) First, we parse the text and segment it into sentences. (Step B) Then, we obtain images for each sentence by computing a search query and using it with an image search engine. (Step C) We also generate audio narration using text-to-speech. (Step D) Finally, we compose the audio-visual slideshow by time-aligning the images with the narration and adding finishing effects.

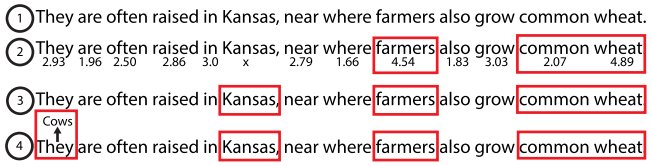

Figure 5. Three sub-steps are used to construct the search query (red boxes) for an input text sentence. Compute concreteness score for each word, and select all words and noun phrases with concreteness > 4.5 (sub-step 1). Add named entities (sub-step 2). Replace pronouns using co-reference resolution (sub-step 3).

METHODS

The goal of our system is to generate an audio-visual slideshow from a text article that represents the concrete information in the text. Our process consists of four major steps (Fig. 4). First, we break the text into sentences (step A). Since each sentence usually contains its own independent idea, we choose a single image for each sentence (step B). We use text-to-speech to generate audio narration (step C). Finally, we time-align the audio narration to the images and add final effects to compose the audio-visual slideshow (step D).

The main contribution of our paper is our concreteness-based method for obtaining appropriate images to represent each sentence. First, we describe this part of our algorithm in detail, and then, we explain how we compose the final slideshow from the images and the audio narration.

Obtaining Images for Text Using Concreteness

Our goal is to obtain images from image search engines that best represent the concrete information contained in each sentence. To do so, we must compute an image search query. There are different ways to generate a search query, such as searching using the entire sentence or using a keyword selection algorithm. Our method uses the notion of *concreteness* to select the words in the sentence that contain the main perceptual information. In our evaluation, we compare our approach with a keyword-based selection method.

There are three sub-steps to compute the image search query for each sentence (Fig. 5). We explain each step using an example sentence from a text article about cows:

> They are often raised in Kansas, near where farmers also grow common wheat.

Sub-step 1: Concreteness

First, we look up the concreteness score for each word in the sentence in a concreteness lexicon that contains human-rated scores for 40K common words and compound phrases [13]. In this dataset words are rated on a scale of 1 to 5 based on how easily a word can be perceived by one of the five human senses. We select words with concreteness values above a threshold τ, to form the initial search query. Empirically we find that setting $\tau = 4.5$ produces good results.

Although we obtain concreteness scores for each word in a sentence, in some cases, we need to consider certain phrases together, such as noun phrases and compound nouns. In our example sentence, the words "common" and "wheat" should be considered together as "common wheat," where the adjective "common" is used to describe a type of "wheat." We use the spaCy dependency parser [24] to identify noun phrases and compound nouns. If any word in the noun phrase or compound noun is above the concreteness threshold, we add the whole phrase to the search query. The search query for our example sentence includes {"farmers", "common wheat"}, since "farmers" has a concreteness score of 4.54, and "common wheat" has a concreteness score of 4.89 (Fig. 5, sub-step 1).

Sub-step 2: Named Entities

Although the concreteness lexicon covers a large number of words, it omits many specific people, places, and organizations, in which case the concreteness score is not defined. However, such named entities usually represent a concrete idea so we include them in the search term. We use the spaCy named entity recognition tags to identify words that refer to people, places, and organizations [24]. In our example sentence, the word "Kansas" is not in the concreteness lexicon, but we add it to the search query because it is a named entity. The search query then becomes: {"Kansas", "farmers", "common wheat"} (Fig. 5, sub-step 2).

Sub-step 3: Pronoun Replacement

Text articles often use pronouns to refer to concrete objects across multiple sentences. In our example sentence the word "they" is used to refer to "cows," which appears in an earlier sentence ("Cows are abundant in the Midwest"). To determine what the pronoun is referring to, we use the pronoun coreference resolution method implemented by NeuralCoref [59]. This data-driven NLP approach selects the most

likely reference-word in the text for each occurrence of a pronoun.

If the concreteness score of the reference-word is higher than the concreteness threshold, we add it to the search query for the sentence that contained the pronoun. In our example we add the word "Cows," which replaces the pronoun "They." The final search query for our example sentence is: {"Kansas", "farmers", "common wheat", "Cows"} (Fig. 5, sub-step 3).

Special Cases

In certain special cases we augment or adjust the search query to retrieve relevant images. If the search query includes duplicate words, we keep only a single occurrence. If a search query contains a single word, we add the article title to provide context and reduce ambiguity. When there are no concrete words in a sentence, the search query may be empty. In this case, we do not show a separate image for this sentence in the slideshow; instead we continue to show the image from the previous sentence. In rare cases in which the search query for the first sentence of an article is empty, we pull the image from the nearest sentence in the article that has a valid search query. Figures 6 and 7 show several other examples of search queries obtained using our method.

Image Selection

Once we have the search query, we use an image search engine to obtain images for the slideshow. We have experimented with several image search engines and decided to use Bing Image Search [2] in our implementation. A good slideshow contains images that are not only relevant, but also of high quality. To obtain high-quality images, we apply several filters in Bing Image Search to select images that have a minimum resolution of 480 x 360px and a horizontal aspect ratio that is close to 4:3, the target aspect ratio of our output slideshows. We also filter out charts, diagrams, and images that contain text since these types of images may contain information that conflicts with the text narration [23]. We also use filters to only retrieve photographs (vs. graphic designs). To avoid watermarks, we remove stock images by filtering URLs of common stock photo websites. Finally, we select the top search result as the representative image for the sentence.

Slideshow Composition

Once we obtain the images to use, we compose the slideshow by time-aligning the images with the audio narration and adding finishing effects.

Audio Narration

We use the Google Cloud Text-to-Speech service [4] to generate the audio narration from the input text. Since the text-to-speech service does not provide timing information, we re-process the output audio through Google Speech-to-Text [3], which returns per-word time-stamps. The speech-to-text result may contain transcription errors. To find optimal alignment between the input text article and the transcript, we use the Needleman-Wunsch algorithm [38]. The output of this step is audio narration of the input text with a time-stamp for each narrated word.

Input Text	# Sent.	Mean Search Set Size	Audio Dur. (sec)	Source
Autobahn	3	2.33	20	Wikipedia [50]
Border Collie	5	2.60	30	Wikipedia [55]
Evergreen	12	1.92	93	Wikipedia [57]
Food Allergies	5	3.80	25	HowStuffWorks [35]
Hawaiian Luau	9	1.78	66	GoHawaii [22]
Jelly, jam,& preserves	19	2.32	79	HowStuffWorks [54]
Mayonnaise	5	2.80	28	Wikipedia [56]
Moldy bread	11	1.55	43	HowStuffWorks [30]
Guggenheim	11	3.64	118	Wikipedia [58]
Sunburn	20	1.95	102	HowStuffWorks [47]
Symmetry	15	1.33	83	HowStuffWorks [44]
String Cheese	23	2.00	125	HowStuffWorks [48]
Weighted Blankets	16	1.94	131	HowStuffWorks [15]

Table 1. For each of the examples here we describe the length of the input in sentences, the number of search terms we select, and the time of the output video.

Time-aligning images to the narration

We time-align each sentence in the narration with the corresponding image. However, instead of placing the image transition at the beginning of a sentence, we display the image when the first word that is part of the search query for that sentence is mentioned because showing an image too early can be confusing.

We also avoid displaying images for a very short period of time because such transitions can be jarring, and people may not have enough time to understand the image. If the image duration is less than 2 seconds, we remove the image and extend the duration of the previous image.

Composition and Effects

To improve the visual quality and clarity of the slideshows, we apply video effects and add captions. First, we resize and crop large images to fit the slideshow dimensions (960 x 720px). To resize the image, we scale the image, keeping the original aspect ratio, and then we crop to fit the target size. For cropping, we use python-smart-crop [19], which uses facial bounding boxes and feature detection to select the best framing of an image for a desired output size. Finally, for all image clips that span longer than three seconds, we apply a zoom or pan effect. To do this, we detect the largest facial bounding box or the most salient region in the image using OpenCV [12]. If a face is detected, we zoom toward the largest face. Otherwise, if the salient region is in the center or occupies more than half of the frame, we zoom in toward the center of that region. If the salient region is to the left, right, top, or bottom of the frame, we pan from the opposite direction toward the salient region. The zoom or pan speed is based on the duration of the image clip. Finally, we add captions at the bottom of the screen that display the original text. This adds an extra modality of information and makes the slideshows more accessible.

RESULTS

We have used our system to generate results for a variety of different text articles (Table 1), including Wikipedia entries and articles from HowStuffWorks. These inputs vary in length from 3 to 23 sentences long, and the resulting slideshows are 20 to 131 seconds in length. Our system takes between 2 and 10 minutes to produce these results, and the vast majority

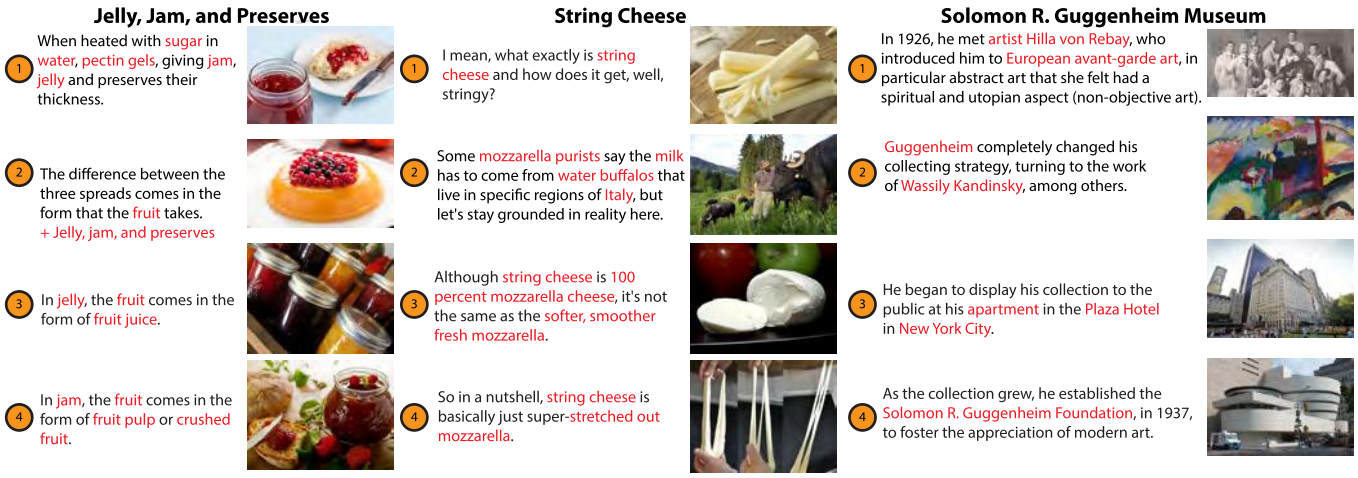

Figure 6. We use our system to generate videos for a variety of different texts. We show excerpts of results for articles on "Jelly, jam, and preserves" and "Weighted blankets" from *How Stuff Works* and an article on "The Solomon R. Guggenheim Museum" from *Wikipedia*. Our system identifies the concrete words (shown in red) in the sentences and retrieves relevant images for these different types of text.

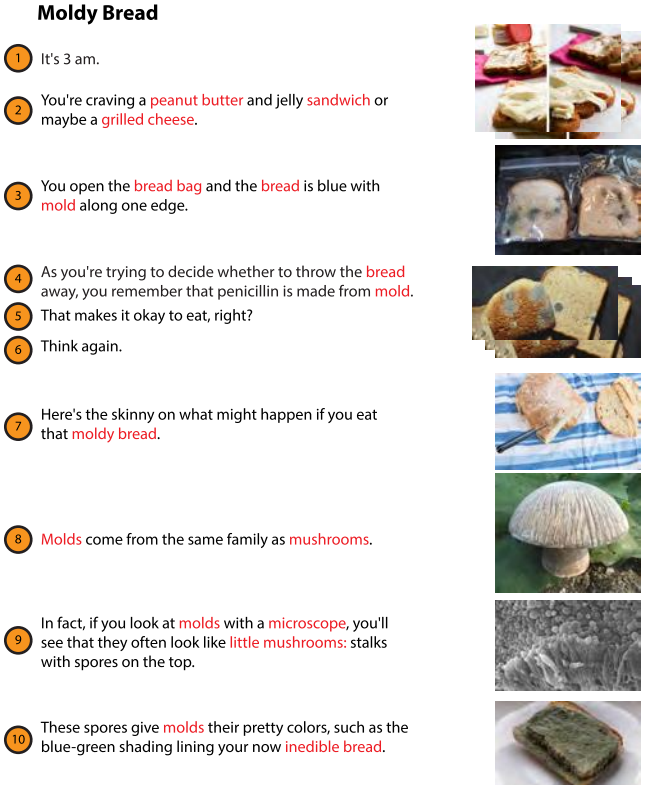

Figure 7. An excerpt from a *How Stuff Works* article about "Moldy bread." The words in the search query appear in red. For conversational sentences (1, 5, 6) that don't contain concrete words, our slideshow pulls the image from the next sentence (in the case of sentence 1) or holds the prior image on screen.

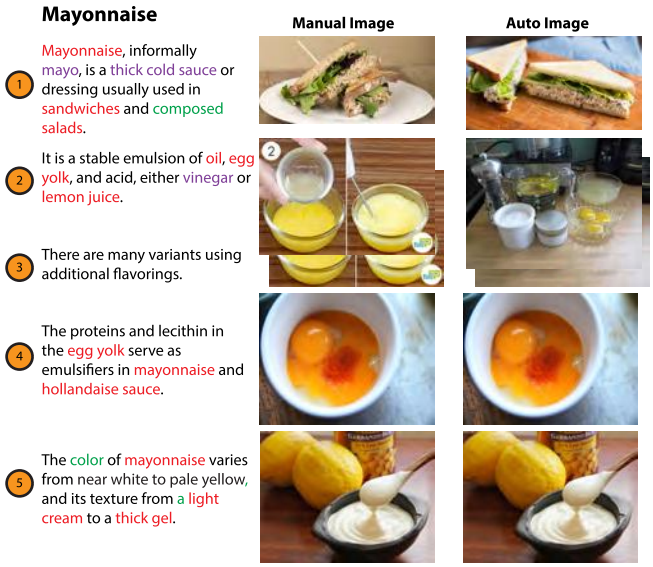

Figure 8. Comparison of manual (from one human annotator, P3) and automatic search queries. Red words appear in both the manual and automatic search queries. Green words only appear in the manual query, and purple words only appear in the automatic query. Although individual words may differ between the manual and automatic search queries, the images returned by the searches are relevant in both cases.

of this time is spent downloading the images from Bing and applying zooming and panning effects to the slideshow clips.

Examples of the resulting visuals for several of these texts are shown in Figures 1 and 6–8. In all of these examples, our system extracts the concrete words from the sentences while excluding more abstract words. Consider the "Hawaiian Luau" article (Fig. 1). Our algorithm extracts search queries that include the concrete words and phrases, such as "taro leaf," "meal," and "Hawaiian Islands." The search query for the third sentence is the name "King Kamehameha II" which does not appear in the concreteness lexicon, but is detected by the named entity parser. Our example on "The Solomon R. Guggenheim Museum" (Fig. 6, right) also heavily relies on named entity recognition to retrieve images for specific people, such as Hilla von Rebay and Wassily Kandinsky, and places, such as The Plaza Hotel and New York City.

These examples also show many instances in which a sentence contains multiple different concrete words. Our system generally retrieves images that show at least one of these words, and in many instances, it finds images depicting several of them. For example, in the "Mayonnaise" example (Fig. 8, right, sen-

Article	# Sent.	Mean F1, P1	Mean F1, P2	Mean F1, P3
Autobahn	3	0.58	0.56	0.46
Border Collie	5	0.75	0.50	0.62
Food Allergies	5	0.77	0.60	0.47
Mayonnaise	5	0.72	0.53	0.86
Moldy Bread	10	0.88	0.68	0.76

Table 2. For each sentence, we compare the words in the manual and automatic search query and report the mean F1 score for each article for three human annotators (P1, P2, and P3).

tence 2), the search query includes “oil,” “egg yolk,” “vinegar” and “lemon juice,” and our system retrieves images showing all of these ingredients. Our concreteness metric is able to pick up on details that let us select highly relevant imagery. For example, in our “String Cheese" example (Fig. 6, center), in sentence 4, our search query includes “string cheese," “stretched out," and “mozzarella," and the corresponding image result shows cheese stretched out.

We also encounter situations in which there are no concrete words in a sentence, such as in the 1st (“It’s 3 am.”), 5th (“That makes it okay to eat, right?”), and 6th (“Think again”) sentences in the “Moldy Bread” example (Fig. 7). In these cases, our system either holds a prior image, if there is one (sentences 5 and 6), or uses the following image when a prior image doesn’t exist (sentence 1). Even though the term “moldy bread” occurs repeatedly across multiple sentences, our system shows a variety of images related to this term, since the term is almost always paired with different concrete words in different sentences. This balance between showing relevant yet diverse imagery results in compelling visual slideshows.

In other scenarios, there is just a single concrete term in the sentence, which can lead to lack of specificity. In these situations, our system adds additional context by including the title of the article in the search query. For example in “Jelly, Jam, and Preserves," in the second sentence the only concrete word is fruit (Fig. 6, left). Just retrieving images for the word “fruit” would yield images that are too general to fit the sentence. Our system adds the title of the article to add more context and search for images that relate fruit to jelly, jam, and preserves.

Although we focus on producing slideshows, our technique can be used to generate other types of formats using media from different sources. Specifically, our algorithm that pairs text with relevant images based on concreteness can be used to arrange text with images in a static document, as in the examples created for our formative study (Fig. 2).

EVALUATION

We evaluate the effectiveness of our audio-visual slideshow generation system in two ways. (1) To understand how well our system identifies the appropriate image search query, we compare its search queries to manually generated search queries. (2) To evaluate the overall quality of our slideshows, we ask users to compare slideshows generated by our system to manually created slideshows as well as slideshows created using a simple keyword-search based method.

Comparison of Automatic and Manual Search Queries

The most important part of our slideshow production algorithm is the selection of search queries based on word concreteness. To evaluate the quality of these search queries we compare

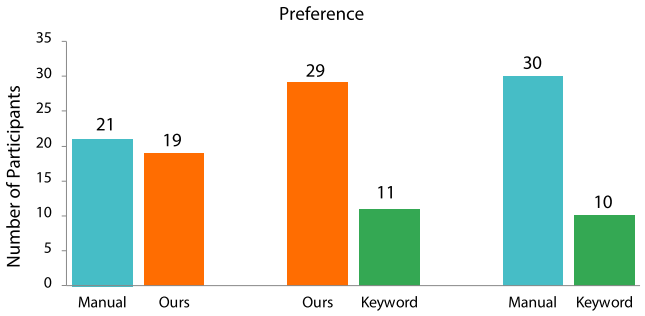

Figure 9. Participants strongly preferred our slideshows over the keyword-based version. They did not have a strong preference between our results and manually created slideshows. This suggests that our algorithm produces slideshows that are as good as manually created ones, and better than the keyword-based approach.

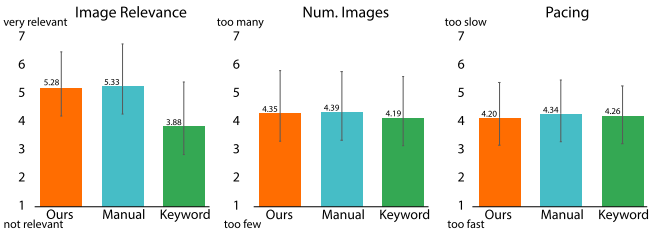

Figure 10. Participants rated the images in our slideshows as more relevant to the content than the keyword-based approach, and similar to the manually created slideshows. For all three conditions, participants were generally pleased with the number of images and the pace of the slideshows.

them to manually generated search queries. For five different text articles, three external annotators (P1, P2, and P3) independently examined each sentence and identified relevant search terms contained within the sentence without knowledge of the automatically generated search query. Fig. 8 shows the manual and automatic search queries for the Mayonnaise article for P3, while Table 2 shows the average F1 overlap scores between the manual and automatic search queries across the five articles. In general, we obtain significant overlap with an average F1 score of 0.74 for P1, 0.57 for P2, and 0.63 for P3. For comparison, a random query selection would have a F1 score of 0.21.

In addition to comparing the manual and automatic search queries, we compare the image search results (Fig. 8). We find that even in cases where the search terms differ, the resulting images may not differ in meaningful ways. For instance, in the second sentence, the automatic search query contains one additional word—“vinegar”— that is not in the manual search query. The inclusion of this additional word results in the search engine returning a different image, but both versions show mayonnaise ingredients.

User Study and Feedback

We conducted a user study to assess the quality of our output slideshows compared to slideshows generated using a keyword-search based approach (*keyword*) that does not take concreteness into consideration and to manually constructed slideshows created in our formative study (*manual*). We also experimented using full sentences as our search term but found

the results inferior to our approach. Full sentence examples are included in the supplementary material.

For keyword-based search, we used Rapid Automatic Keyword Extraction (RAKE) [45] to extract keyphrases from the text. RAKE is similar to TF-IDF [26] in that it uses the frequency and co-occurrence of words within the text to extract important parts of the text, but it extracts phrases rather than keywords and is more appropriate for the short excerpts of text in our examples. We use the top-ranked keyphrase in each sentence as the image search query and time the image with the narration according to the start of the keyphrase.

We recruited 120 participants on Amazon Mechanical Turk to evaluate the different slideshows. In an early pilot we found that asking participants to compare three different slideshows (*manual*, *keyword*, and *ours*) was difficult. Users mentioned that by the time they saw the third slideshow they did not remember much about the first one. To avoid this problem, we instead asked each participant to compare a pair of slideshows on the same article. Each pairing (*ours* vs. *keyword*, *manual* vs. *keyword*, and *ours* vs. *manual*) was evaluated by 40 participants. After watching the slideshows, participants answered questions about the quality of each slideshow and their preference between the two. The study took about 4 minutes, and participants were compensated with $0.50.

Figure 9 and 10 summarize the results. Participants strongly preferred our slideshows over the keyword-based version with 29 people preferring *ours* (vs. 11 *keyword*). They also strongly preferred the *manual* over *keyword* (30 vs. 10). Participants did not have a strong preference between *ours* and *manual* (19 vs. 21). This suggests that our approach produces slideshows that are as good as manually created ones and better than the keyword-based approach.

We also asked the participants to rate the relevance of the images in each slideshow, as well as the number of images and the pace of the slideshows using a 1-7 Likert scale (Figure 10). Participants rated the images in *ours* (M = 5.28, SD = 1.31) and *manual* (M = 5.33, SD = 1.52) as more relevant to the content than images in *keyword* examples (M = 3.88, SD = 1.60). Our relevance results suggest that our image selection algorithm based on concreteness obtains images that are highly relevant to the content, as good as those selected manually by people, and better than those selected by a keyword-based approach. Participants were generally satisfied with the number of images and the pace of the presentation, and there was no significant difference for all three versions of the slideshows.

LIMITATIONS AND FUTURE WORK

Concreteness can be applied to a wide range of domains, including poetry [28]. We applied our approach to several texts from classic literature, such as William Wordsworth's poem "I Wandered Lonely as a Cloud" (Fig. 11). Despite having a different grammatical structure from informational articles, the poem contains concrete words in each verse, and our system can produce a convincing audiovisual slideshow. Visualizing poems and other creative texts presents additional challenges and is an interesting avenue for future work.

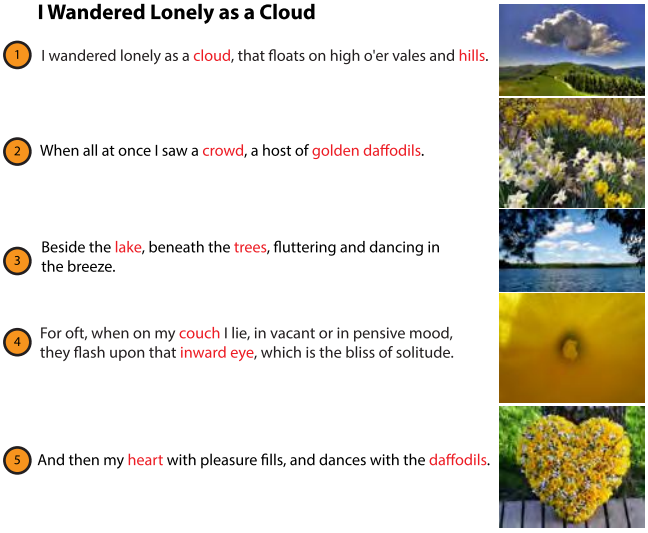

Figure 11. William Wordsworth's poem "I Wandered Lonely as a Cloud" uses concrete language, which enables our system to produce an audiovisual slideshow for the poem.

While our system does not explicitly filter for copyrighted images, we encourage people to use our system in a way that respects the rights of content creators. For example, users can apply our algorithm with additional image search filters to exclude copyrighted images or obtain the licensing for the selected images.

While our approach is general enough to apply to a wide range of texts, we encountered examples that did not work well. Narratives about very specific objects or people, particularly those that are not famous or public, do not work well because there are not enough publicly available images to illustrate these texts. Highly technical and scientific articles also pose a challenge because they contain words that are too obscure to appear in the concreteness lexicon. They are also misidentified with part-of-speech tags and named-entity recognizers. In addition, colloquialisms and extensive use of metaphorical language can lead the algorithm to choose images that may incorrectly represent the meaning of the text.

Automatically understanding more nuanced interpretations of text is an active area of research in the natural language processing community. Applying these techniques to retrieve better images to pair with text could be particularly applicable to domains like fiction and lyrical poetry.

Although our results focus on searching for static images, the methodology of computing a search query from a text using concreteness can be used to search for other media, such as video clips. Composing content with video clips would entail additional challenges, including trimming and timing the video content.

Finally, we use word concreteness, but there are other similar measures that can be used to select portions of text to visualize. For example, specificity, familiarity, and imageability have also been explored in connection with concreteness and text comprehension [33, 18, 16]. Imageability, in particular, is closely related to concreteness. Imageability has been defined as "the extent to which the item evokes a mental image," and

concreteness has been defined as "the extent to which it can be experienced by the senses" [43]. These definitions have overlap in their notion of evoking a sensory response, and there are theoretical works that seek to explain how the brain processes concrete and imageable words and how these concepts are correlated [39, 17]. Combining or comparing different measures for finding relevant imagery is an interesting problem for future work.

CONCLUSION

Text articles are abundant on the web today. Many of these articles contain descriptive language designed to evoke visual imagery. We have demonstrated an automated approach for converting such articles into audio-visual slideshows by identifying concrete words in each sentence, converting text to speech, and automatically selecting relevant images. There is a wide range of potential applications for our system, including illustrating audiobooks, educational materials, speeches, and other scenarios in which reading may not be preferred. We believe audiovisual slideshows can serve as informative, multimodal, and accessible alternatives to text articles.

ACKNOWLEDGMENTS

We thank Abe Davis, Urvashi Khandelwal, and Dan Jurafsky for their helpful comments and suggestions on this project. Our research is supported by The Brown Institute for Media Innovation.

REFERENCES

[1] 2019. Article Video Robot. (2019). https://www.articlevideorobot.com/.

[2] 2019. Bing Image Search. (September 2019). https://www.bing.com/images/

[3] 2019a. Google Cloud Speech to Text API. (Mar 2019). https://cloud.google.com/speech-to-text

[4] 2019b. Google Cloud Text to Speech API. (Mar 2019). https://cloud.google.com/text-to-speech

[5] 2019. Magisto. (2019). https://www.magisto.com/.

[6] 2019. Wibbitz. (2019). https://www.wibbitz.com/.

[7] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 2011. Enriching Textbooks with Images. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 1847–1856. DOI: http://dx.doi.org/10.1145/2063576.2063843

[8] Patricia Baggett. 1979. Structurally equivalent stories in movie and text and the effect of the medium on recall. *Journal of Verbal Learning and Verbal Behavior* 18, 3 (1979), 333 – 356. DOI:http://dx.doi.org/https://doi.org/10.1016/S0022-5371(79)90191-9

[9] Tom Bartindale, Guy Schofield, and Peter Wright. 2016. Scaffolding Community Documentary Film Making Using Commissioning Templates. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2705–2716. DOI: http://dx.doi.org/10.1145/2858036.2858102

[10] John Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

[11] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Trans. Graph.* 31, 4, Article 67 (July 2012), 8 pages. DOI: http://dx.doi.org/10.1145/2185520.2185563

[12] Michael Beyeler. 2015. Open-CV Saliency. https://github.com/mbeyeler/opencv-python-blueprints/blob/master/chapter5/saliency.py. (2015).

[13] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46, 3 (2014), 904–911.

[14] Juan Casares, A. Chris Long, Brad A. Myers, Rishi Bhatnagar, Scott M. Stevens, Laura Dabbish, Dan Yocum, and Albert Corbett. 2002. Simplifying Video Editing Using Metadata. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '02)*. ACM, New York, NY, USA, 157–166. DOI: http://dx.doi.org/10.1145/778712.778737

[15] Nathan Chandler and Kate Kershner. 2019. Do Weighted Blankets Help With Sleep? (2019). https://health.howstuffworks.com/mental-health/sleep/basics/do-weighted-blankets-help-sleep.htm

[16] Max Coltheart. 1981. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A* 33, 4 (1981), 497–505. DOI: http://dx.doi.org/10.1080/14640748108400805

[17] Louise Connell and Dermot Lynott. 2012. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition* 125, 3 (2012), 452 – 465. DOI: http://dx.doi.org/https://doi.org/10.1016/j.cognition.2012.07.010

[18] Pasquale A Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano F Cappa. 2010. Beyond the abstract—concrete dichotomy: mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior research methods* 42, 4 (2010), 1042–1048.

[19] Epixelic. 2017. python-smart-crop. https://github.com/epixelic/python-smart-crop. (2017).

[20] K. Fliessbach, S. Weis, P. Klaver, C.E. Elger, and B. Weber. 2006. The effect of word concreteness on recognition memory. *NeuroImage* 32, 3 (2006), 1413 – 1421. DOI:http://dx.doi.org/https://doi.org/10.1016/j.neuroimage.2006.06.007

[21] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based Editing of Talking-head Video. *ACM Trans. Graph.* 38, 4, Article 68 (July 2019), 14 pages. DOI: http://dx.doi.org/10.1145/3306346.3323028

[22] Go Hawaii. 2019. Luau of Hawaii. (2019). https://www.gohawaii.com/hawaiian-culture/luau

[23] Hoffstaetter. 2016. Python Tesseract. https://github.com/madmaze/pytesseract. (2016).

[24] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).

[25] Bernd Huber, Hijung Valentina Shin, Bryan Russel, Oliver Wang, and Gautham J. Mysore. 2019. B-Script: Transcript-based B-roll Video Editing with Recommendations. *To Appear in CHI 2019* (May 2019). DOI:http://dx.doi.org/10.1145/3290605.3300311

[26] Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* (2004).

[27] Murat Kalender, M. Tolga Eren, Zonghuan Wu, Ozgun Cirakman, Sezer Kutluk, Gunay Gultekin, and Emin Erkan Korkmaz. 2018. Videolization: knowledge graph based automated video generation from web content. *Multimedia Tools and Applications* 77, 1 (01 Jan 2018), 567–595. DOI: http://dx.doi.org/10.1007/s11042-016-4275-4

[28] Justine T Kao and Dan Jurafsky. 2015. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)* 12 (2015).

[29] Joy Kim, Mira Dontcheva, Wilmot Li, Michael S. Bernstein, and Daniela Steinsapir. 2015. Motif: Supporting Novice Creativity Through Expert Patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1211–1220. DOI: http://dx.doi.org/10.1145/2702123.2702507

[30] Karen Kirkpatrick. 2018. What If You Eat Moldy Bread? (Aug 2018). https://science.howstuffworks.com/science-vs-myth/what-if/what-if-eat-moldy-bread.htm

[31] Matthew J. Koehler, Aman Yadav, and Michael Phillips. 2005. What is Video Good For? Examining How Media and Story Genre Interact. *Journal of Educational Multimedia and Hypermedia* 14, 3 (2005), 249–272.

[32] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational Video Editing for Dialogue-driven Scenes. *ACM Trans. Graph.* 36, 4, Article 130 (July 2017), 14 pages. DOI: http://dx.doi.org/10.1145/3072959.3073653

[33] Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[34] W. Li and H. Zhuge. 2014. Summarising News with Texts and Pictures. In *2014 10th International Conference on Semantics, Knowledge and Grids*. 100–107. DOI:http://dx.doi.org/10.1109/SKG.2014.34

[35] Linnea Lundgren. 2006. Understanding Food Allergies. (Apr 2006). https://health.howstuffworks.com/diseases-conditions/allergies/food-allergy/information/food-allergies-ga.htm

[36] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.

[37] N Hari Narayanan and Mary Hegarty. 2002. Multimedia design for communication of dynamic information. *International journal of human-computer studies* 57, 4 (2002), 279–315.

[38] Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48, 3 (1970), 443–453.

[39] Allan Paivio. 2014. *Mind and its evolution: A dual coding theoretical approach*. Psychology Press.

[40] Allan Paivio, Mary Walsh, and Trudy Bons. 1994. Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 5 (1994), 1196.

[41] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 181–190. DOI: http://dx.doi.org/10.1145/2807442.2807502

[42] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 573–582. DOI: http://dx.doi.org/10.1145/2642918.2647400

[43] John TE Richardson. 1975. Concreteness and imageability. *The Quarterly Journal of Experimental Psychology* 27, 2 (1975), 235–249.

[44] Dave Roos. 2017. Why Do We Get So Much Pleasure From Symmetry? (Nov 2017). https://science.howstuffworks.com/why-do-get-so-much-pleasure-from-symmetry.htm

[45] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*. 1 – 20. DOI: http://dx.doi.org/10.1002/9780470689646.ch1

[46] Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Frédo Durand. 2015. Visual Transcripts: Lecture Notes from Blackboard-style Lecture Videos. *ACM Trans. Graph.* 34, 6, Article 240 (Oct. 2015), 10 pages. DOI:http://dx.doi.org/10.1145/2816795.2818123

[47] Sarah Siddons. 2009. How to Treat Sunburn. (2009). https://health.howstuffworks.com/skin-care/beauty/sun-care/how-to-treat-sunburn.htm

[48] Meg Sparwath. 2019. How Does String Cheese Get Stringy? (Aug 2019). https://recipes.howstuffworks.com/food-science/string-cheese.htm

[49] Hariharan Subramonyam, Wilmot Li, Eytan Adar, and Mira Dontcheva. 2018. TakeToons: Script-driven Performance Animation. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 663–674.

[50] Cherise Threewitt. 2019. Is the Autobahn Headed to California? (Mar 2019). https://auto.howstuffworks.com/is-autobahn-headed-to-california.htm

[51] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. QuickCut: An Interactive Tool for Editing Narrated Video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 497–507. DOI: http://dx.doi.org/10.1145/2984511.2984569

[52] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International journal of human-computer studies* 57, 4 (2002), 247–262.

[53] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 43–52.

[54] Kathryn Whitbourne. 2019. What Is the Difference Between Jelly, Jam and Preserves? (2019). https://recipes.howstuffworks.com/question84.htm

[55] Wikipedia contributors. 2018a. Border Collie — Wikipedia, The Free Encyclopedia. (2018). https://en.wikipedia.org/wiki/Border_collie

[56] Wikipedia contributors. 2018b. Mayonnaise — Wikipedia, The Free Encyclopedia. (2018). https://en.wikipedia.org/wiki/Mayonnaise

[57] Wikipedia contributors. 2019a. Evergreen — Wikipedia, The Free Encyclopedia. (2019). https://en.wikipedia.org/wiki/Evergreen

[58] Wikipedia contributors. 2019b. Solomon R. Guggenheim Museum — Wikipedia, The Free Encyclopedia. (2019). https://en.wikipedia.org/wiki/Solomon_R._Guggenheim_Museum

[59] Thomas Wolf. 2018. NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks. https://github.com/huggingface/neuralcoref. (2018).

[60] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. 2007. A text-to-picture synthesis system for augmenting communication. In *AAAI*, Vol. 7. 1590–1595.