Crosscast: Adding Visuals to Audio Travel Podcasts

Haijun Xia
UC San Diego
haijunxia@ucsd.edu

Jennifer Jacobs
UC Santa Barbara
jmjacobs@ucsb.edu

Maneesh Agrawala
Stanford University
maneesh@cs.stanford.edu

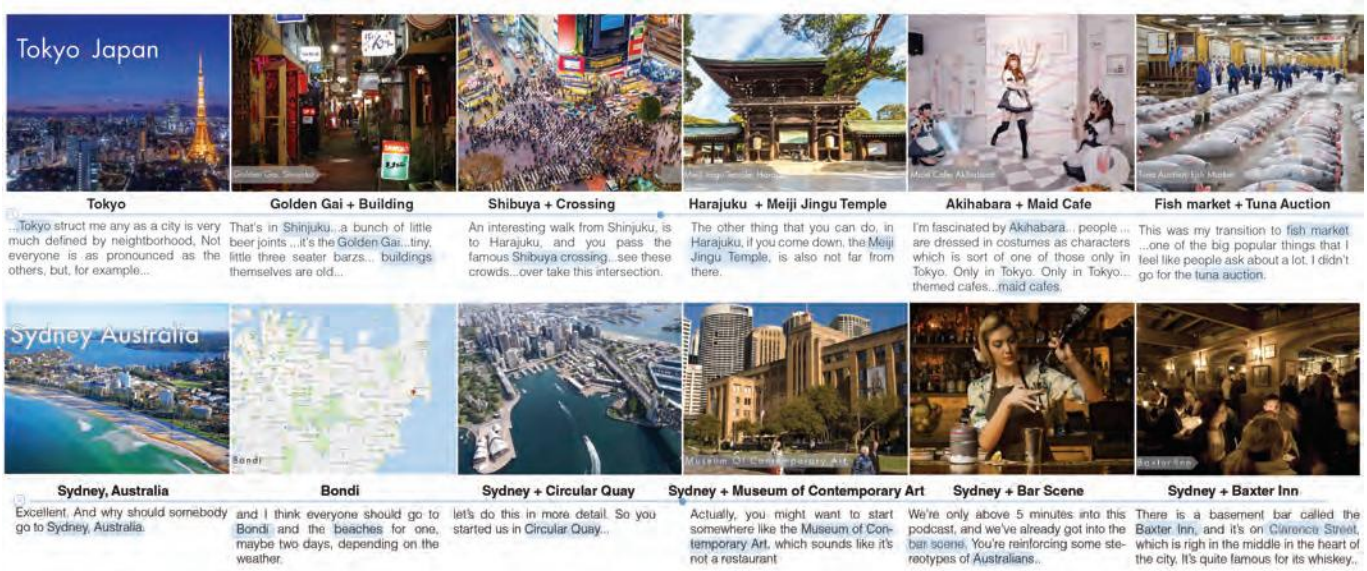

Figure 1. Given audio travel podcasts and transcripts (e.g. travel to Tokyo and Sydney), Crosscast automatically selects the most relevant locations and visual entities at any moment of a podcast, and queries and displays images to accompany the audio, enabling audio-visual travel storytelling experience.

ABSTRACT

Audio travel podcasts are a valuable source of information for travelers. Yet, travel is, in many ways, a visual experience and the lack of visuals in travel podcasts can make it difficult for listeners to fully understand the places being discussed. We present Crosscast: a system for automatically adding visuals to audio travel podcasts. Given an audio travel podcast as input, Crosscast uses natural language processing and text mining to identify geographic locations and descriptive keywords within the podcast transcript. Crosscast then uses these locations and keywords to automatically select relevant photos from online repositories and synchronizes their display to align with the audio narration. In a user evaluation, we find that 85.7% of the participants preferred Crosscast generated audio-visual travel podcasts compared to audio-only travel podcasts.

INTRODUCTION

Today, more than 20 million people listen to travel podcasts in the United States [30]. Such podcasts are a prominent source of information for listeners to discover new locations, hear about travel adventures, and gain inspiration for future trips. Travel podcasts describe personal travel stories, contain interviews with locals or travel experts, highlight notable landmarks, and provide general guidance on planning a route through a given region.

Despite the benefits of travel podcasts, for many people, the primary appeal of travel is rooted in the *visual experience* of new places. Travelers place high value on understanding where a particular location is in relation to places they already know as well as seeing notable or beautiful landmarks, objects, and scenery while on vacation [43], and most tourists travel with a camera or return home with visual documentation of their trip [6]. Because the visual experience is so important while engaged in travel, the ability to view maps, pictures, and videos of travel destinations plays a central role when people are researching and planning future travel [11]. While travel podcasts offer a wealth of verbal information, unlike many other forms of travel media (e.g. guidebooks, travel documentaries, travel v-logs), they lack visuals, including maps, images, and videos. The visual nature of travel suggests there are significant benefits to enhancing audio travel podcasts with relevant, high-quality visual content.

Adding visuals to travel podcasts aligns with existing forms of podcast consumption and efforts by podcast creators to expand the format. Podcasts are sometimes thought of as a format that people listen to while doing other things, but a recent survey showed that the majority of audio podcast listeners (70%) engage in focused consumption and solely listen to podcasts without performing other activities [30]. For such listeners, visuals could provide another channel for engagement.

UIST '20, October 20–23, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7514-6/20/10 ...$15.00.
https://doi.org/10.1145/3379337.3415845

Furthermore, adding visuals to a podcast does not enforce a change in experience. Listeners can choose to engage with the visuals, or listen to the podcast in the standard audio format.

Podcast creators in other domains such as investigative journalism [5] and lifestyle stories [12, 35] have started to explore adding visuals to podcasts to increase engagement and facilitate content comprehension. Our work aligns with these efforts to visually augment podcasts; however, we focus on the unique opportunity of visually enhancing travel podcasts with relevant images of the locations discussed in the audio.

Travel podcasts are often structured around descriptions of specific locations including geography, regions, landmarks, artifacts, and local establishments (restaurants, bars, and hotels). Manually adding images and maps that depict the corresponding locations poses many challenges. Podcast creators must (1) select specific locations and details they want to visualize at every moment of the audio, (2) curate appropriate images and maps of these locations and details, and (3) align the selected images and maps with the narration. Optionally, they may want to (4) add text labels that identify the content of each image and map. Each of these steps is time consuming. In our research, it took an author with professional training in video editing 8 hours to produce one complete sequence for a 30-minute travel podcast. On average, audio travel podcasts range from 30 to 60 minutes. Manually adding visuals to most audio travel podcasts, while valuable, requires significant effort.

We present Crosscast, an automated tool for adding visuals to travel podcasts. To develop Crosscast, we surveyed over 300 travel podcast episodes, as well as travel documentaries and TV shows, to determine a structure for identifying and synchronizing relevant travel visuals to audio narration. We instantiated this structure as a set of computational methods to automatically (1) identify the most relevant locations and visually important entities in travel podcast transcripts, using natural language processing (NLP) and text mining techniques, (2) query and select appropriate images to visualize the content, and (3) synchronize the timing of visuals with the speech.

We demonstrate the effectiveness of Crosscast by adding visuals to a variety of travel podcasts that cover different regions of the world (e.g. Japan, Australia) and use different podcasting styles (e.g. conversational, pre-scripted). Figure 1 shows 2 automatically generated results. We also conducted a user evaluation to evaluate the audio-visual qualities of Crosscast-generated podcasts in comparison to audio-only travel podcasts, and audio-visual podcasts created manually by a professional video editor. We found that participants ranked Crosscast-generated podcasts higher than audio-only podcasts 85.7% of the time and in 32.1% of cases, Crosscast-generated podcasts were rated higher than both audio-only and manually generated audio-visual podcasts.

This paper contributes the identification of travel podcast structure and the design of algorithmic methods that automatically add visuals to travel podcasts based on their structural properties. Crosscast benefits podcast consumers by providing additional visual experience, and podcast creators as the automated results can serve as rough cuts for their video production.

RELATED WORK

Our research draws on prior work in transcript-based audio and video editing systems as well as automatic generation of visual travel resources and other visual content.

Transcript-based Audio and Video Editing

Much research in media production support tools seeks to reduce the difficulty of producing audio or video stories through transcript analysis and editing. Prior transcript-based audio editing tools use time-aligned text transcripts of spoken audio to automatically group similar sentences, highlight repeated words, and maintain synchronization between multiple speakers [33], support automatic alignment of music with spoken audio [32, 31], or enable linked editing between script writing and audio recording and editing [36]. Transcript-based video production systems analyze time-aligned video transcripts to identify points for inserting [14] or removing footage [1], allow for vocally-annotating raw footage [27, 41], or enable the synthesis of short segments of talking-head video of puppets [3] and people [4]. Other systems use script transcript analysis to select relevant video clips [18], or leverage linguistic structures to create corresponding graphical structures [45]. We too seek to reduce difficulties in creating audio-visual stories through automated analysis of spoken audio and therefore also rely on time-aligned transcripts. However unlike prior transcript-based video generation systems, we aim to support the creation of travel-specific visual narratives through a fully automated process and thus rely on text analysis methods which identify relevant geographic locations without any human labeling.

Automatic Generation of Visual Travel Resources

In the area of computational tourism support, several systems generate visual content to aid travelers. Grabler et al. [9] explored the automatic generation of tourist maps which highlights landmarks and other visually distinctive building. Prior works analyze image metadata from internet photo sharing sites, and user-generated travelogues to create visual summaries of specific geographic areas [34] and tourist destinations [26], or generate sparse 3D models of highly-photographed landmarks [37]. Crosscast also queries images from online repositories, however unlike prior work, we generate a temporal visual output with audio narration. We substantially expand on prior work on travel podcast visualization [20] with a system that generalizes to different travel podcasts by different creators. We also evaluate our location and keyword selection via a comparison to gold-standard labeled podcasts, and evaluate the quality of our results through a user study.

Automatic Generation of Visual Content From Text

Significant work has explored the automatic generation of visual content from text. Prior work has explored supplementing relevant visual content to text stories based on the similarity of the story text to the annotation of images or video clips [16, 17, 44]. Similarly, built on large training datasets of visual content with text annotations, research has explored the generation of images [13], 3D scenes [2], and video sequence [22] from text with neural nets. Leake et al. explored leveraging word concreteness, which measures how closely a word is related some perceptible concepts, to automatically create slideshows from

text input, by composing images of the most concrete words [19] . Our work is different because we focus on the design of algorithms that can compute the relevance of locations and location features in real-time with the input text.

TRAVEL PODCAST STRUCTURE

Effective visual communication requires that the content and format of the graphics of any visual media should correspond to the content and format of the content to be conveyed [42]. Visual travel media, including travel shows, documentaries, and travel blogs, conform to this core principle with video footage or photographs corresponding with the adjacent audio narration or textual description respectively With visual correspondence as our primary guideline, developing a system for audio/visual travel stories required identifying a structure within travel podcasts that determines both topical relevance and visual synchronization.

We surveyed more than 300 episodes of travel podcasts as well as popular travel TV shows and documentaries, and noted common patterns hosts used when describing relevant entities with concrete visual qualities. Our analysis highlighted two categories: *locations*, and *visually significant entities (VSEs)*. Locations are geographical areas that contains one or more entities of interests to travelers. VSEs are entities whose visual appearance has been discussed in detail, including landmarks, historical artifacts, and food. We noted that hosts discussed multiple VSEs and locations simultaneously, alternated between mentioning different locations and VSEs, and sometimes referred to VSEs and locations that were not relevant to the central narrative or dialog. We therefore identified the following structural properties that could be used to determine the relevant image to display at any point in a travel podcast:

- ***Freshness***: More recently mentioned locations or VSEs are more relevant than those that have already been mentioned earlier in the podcast.
- ***Geographic specificity***: Geographically specific locations are more relevant (e.g. Tokyo is more relevant than Japan).
- ***Subject relevance***: Explicit mentions of a location or VSE as the subject of a sentence indicate higher relevance and should be prioritized over others in the same sentence. In this example, "Koenji is the birthplace of Tokyo punk music", Koenji is more relevant as it is the subject of the sentence, whereas Tokyo is a modifier of punk music.
- ***Detail relevance***: Locations and VSEs described in notable detail are more relevant.
- ***Distractions***: Locations or VSEs outside the geographical region of the podcast should be disregarded as these are often used for quick comparison, but are less relevant to the primary geographic location being discussed in the podcast.

In our survey of travel videos, documentaries, and TV shows, we also identified a few common lower-level strategies for synchronizing visuals of locations and VSEs with the narration.

- Images of a location or VSE should appear when that location or VSE is explicitly discussed in the podcast and should be identified onscreen with a text label.
- Maps should appear when a location is mentioned for the first time and when several locations are mentioned in a sequence for geographical description. The locations should be marked on the map within broader regions to indicate their relative geographic relationship.
- Transitions should match the topic changes in the audio content and occur at a rate that gives the viewer sufficient time to absorb the details of an image and read the label, but short enough to avoid monotony. In practice we found an interval between 3.5s and 7.5s achieves these criteria.
- In cases where one location is talked about for longer than the maximum transition interval, additional images of the location and VSE should be displayed.
- In cases where locations and VSEs are only briefly discussed in the podcast (i.e. for less than 3.5s), they should be ignored to avoid jarring visual transitions.

ALGORITHMIC METHODS

Given an input audio travel podcast, our goal is to automatically add maps and photographs to it that depict the specific locations and visually significant entities (VSEs) (e.g. landmarks, historical artifacts, food, etc.) being discussed. Our approach involves 6 steps: (1) We acquire a text transcript for the podcast and time-align it to the audio. (2) We label all locations and VSEs in the transcript using NLP and geographic information lookup techniques. (3) We construct the geographical hierarchy among the locations, and identify the primary region of the podcast. (4) We select the most relevant location and VSE for each word in the transcript using scoring functions that account for the structural properties presented in the Travel Podcast Structure section. (5) We compose the names of the top scoring location and VSE into a search query to acquire relevant visual imagery from an image search engine. (6) Finally, for each word in the transcript we select the map if the location is mentioned for the first time, or an image from the set of relevant candidates while ensuring that the map or image is onscreen long enough for viewers to absorb details, but short enough to avoid monotony. Text labels were also added when locations and VSEs are mentioned for the first time to facilitate comprehension.

In the remainder of this section, we describe each of these steps in detail using the running example of a Condé Nast Traveler podcast about Tokyo [40], final result shown in Figure 1.

Step 1: Acquire Time-Aligned Transcript

To obtain high-quality transcripts of input travel podcasts, we use a crowd-sourcing transcription service `rev.com`. We could alternatively use an automatic speech recognition tool [15] for transcription; but, we have found in practice that automatic tools produce many errors when applied to conversational speech, which can be mitigated in the future with the increasing performance of speech recognition. After obtaining transcripts, we use the forced alignment approach of Rubin et al. [33] to time-align the transcripts to the raw audio.

Step 2: Label All Locations and VSEs

We next identify all the locations (i.e. countries, states, cities, sub-districts and neighborhoods that cover one contiguous

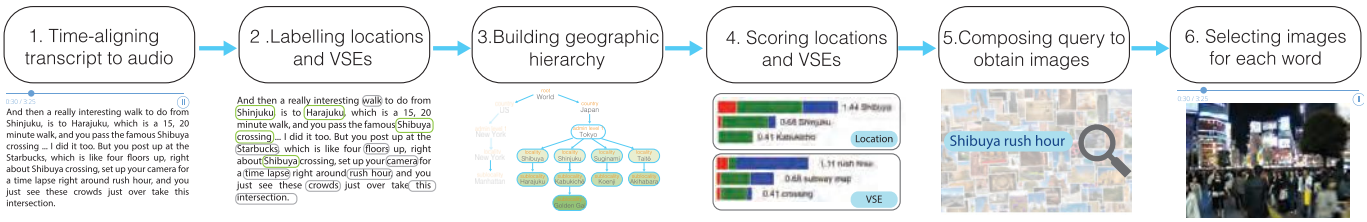

Figure 2. Crosscast receives audio travel podcasts and text transcripts as input and automatically label and score location and visually significant entities, and combines them as search query to retrieve relevant images and time-align to the transcript.

area) and VSEs (i.e. generic nouns such as, *shopping, lake, fashion, pizza, people* as well as specific point locations, such as *Eiffel Tower, Tsujiki Fish Market, Guggenheim Museum*).

We start by processing the transcript using the Google NLP toolkit [7] which labels each noun with one of eight category labels; *location, organization, event, person, work of art, consumer good, other* and *unknown*. We treat all nouns, except those labeled *location*, as VSEs. We have found that the Google NLP Toolkit labels some generic geographic nouns (such as *lake, road, restaurant*) as well as landmarks (such as *Eiffel Tower*) as locations, whereas in our work the *location* category must only include geographic regions. Thus, we further process each word labeled as *locations* by the Google NLP Toolkit, using the Google Places API [8].

We treat the location word as a query and the Places API returns a list of real-world places that best match the query. It also reports whether the resulting places are geographic regions or a point locations. When the query term is a generic geographic noun (e.g. *lake*), the Places API cannot identify a single real-world place with a name that exactly matches the query word and instead returns specific places (e.g. *Lake Ontario, Green Lake*) which fail to exactly match the query term. In such cases we label the query term a *VSE*. In contrast when the query word refers to a real-world place (e.g. *Paris, Eiffel Tower,*), the Places API returns an exact match. In such cases we further check if the query word is a geographic region and if so we label it as a *location*. Otherwise if it is a point location we label it as a VSE.

The following excerpts from the Tokyo podcast show the final labeling. Locations are in green and VSEs are in blue.

struck me as a that is very much defined by ...when you go into , the , , of ...

... this neighborhood of ... it's all low , it's like hanging over the . But it has this really tight knit artistic that has kept it that .

Step 3: Build Geographic Structure

Geographic regions are hierarchical in the sense that neighborhoods are contained within cities which are contained within states and so on. As we proceed up the hierarchy, each location, or administrative level, covers a larger area and is less specific. We use this hierarchy as part of our relevance computation in Step 4 and therefore as we extract each location we lookup its hierarchy using the Google Places API. For example, the location *Koenji* returns the following hierarchy *(country: Japan,*

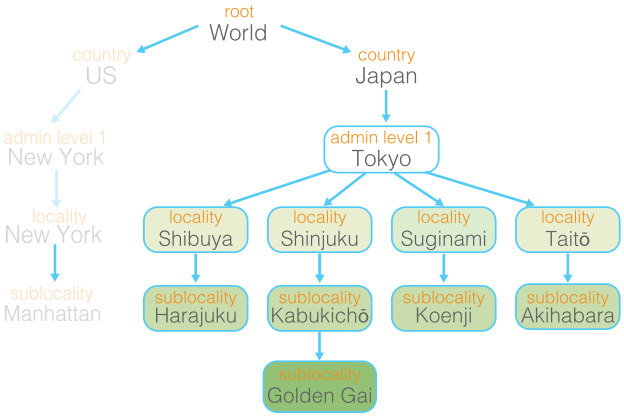

Figure 3. Geographical tree constructed from location entities in the transcript.

1st administrative level: Tokyo, locality: Suginami, sublocality: Koenji). After acquiring the geographical hierarchies for all locations in a podcast, we aggregate them into a geographical tree in which locations near the root are more general and those at the leaves are most specific (Figure 3). Note that looking up some VSEs such as point locations (e.g. *Eiffel Tower*) in the Places API also returns a geographic hierarchy. For each such VSEs we add its immediate parent region to the set of location words in the transcript and we include its hierarchy in the aggregated geographical tree.

Travel podcasts usually focus on presenting a region at one primary administrative level, and while they frequently present multiple locations at lower administrative levels within the same subtree, they rarely present locations in other subtrees. For example, while a podcast about Tokyo may mention multiple neighborhoods within Tokyo, it will only occasionally mention the country of Japan or other countries, cities and neighborhoods outside of Tokyo (e.g. New York, Kyoto). We identify the primary administrative level using a voting scheme as follows. Each time a location is mentioned in the transcript, we add one vote to each of its parent administrative levels in the tree. We then treat the most specific level receiving the most votes as the primary administrative level p. It is the most specific administrative level that contains the majority of locations mentioned in the entire podcast.

Step 4: Select Most Relevant Locations and VSEs

Our goal in step 4 is to identify the most relevant location and VSE for each word in the transcript. More specifically, we consider each transcript word along with a local context

Figure 4. Local context window of a transcript word. We have found that setting an appropriate size n for the sliding local context window is essential in this relevance computation. We empirically tested a variety of different sizes and found that a small window ($\leqslant$ 50 words) usually did not contain enough information to produce good results, whereas a large number of words ($\geqslant$ 200) often introduced noise from earlier parts of the transcript. We found that a 100 word window gave the best results for the podcasts we tested.

window of 100 words that precede it (Figure 4) and initially consider all of the locations and VSEs that lie within the window as relevant candidates. We then score each such candidate with respect to the relevance properties we presented in the Travel Podcast Structure section, using slightly different scoring functions for locations and the VSEs. Finally we set the top-scoring candidates as the most relevant location and VSE for the transcript word. We first describe how we score locations and then explain how we score VSEs.

Scoring Locations

We compute a location relevance score S_{loc} for each candidate location l that appears within the local context window of each transcript word t as a combination of four terms

$$S_{loc}(t,l) = S_{fsh}(t,l)(w_{geo}S_{geo}(l) + w_{sbj}S_{sbj}(l) + w_{dtl}S_{dtl}(l)),$$

where S_{fsh} scores the freshness of location l with respect to transcript word t, S_{geo} scores the geographic specificity of l, S_{sbj} scores the subject relevance of l, S_{dtl} scores the detail relevance of l, while the weights w_{geo}, w_{sbj} and w_{dtl} control the strength of these three terms. Since freshness, or the recency with which a location is mentioned is the most important measure of relevance, we treat S_{fsh} as a weight on the sum of the other three terms. The weights w_{geo}, w_{sbj} and w_{dtl} allow us to balance the contributions of the three other terms and we empirically set $w_{geo} = 1$, $w_{sbj} = 0.8$ and $w_{dtl} = 2$ with a trial and error approach.

Note that if the same location l is mentioned multiple times within a local context window, it is likely to be more relevant. We therefore compute $S_{loc}(t,l)$ as the sum of the location score for each mention. Thus, for each transcript word t we obain a corresponding score $S_{loc}(t,l)$ for each unique location l in the context window. We compute each term of the location score as follows.

Freshness Score $S_{fsh}(t,l)$

A location l can appear at different positions within the local context window. A location that is mentioned closer to the current transcript word t was discussed more recently—it is fresher— and therefore more relevant. Thus, we define the freshness score as

$$S_{fsh}(t,l) = g(d(t,l),\sigma), \tag{1}$$

where $d(t,l)$ is the distance in number of words between t and l and the function g sets how rapidly the freshness falls off as the distance between t and l increases. We use a Gaussian falloff so that

$$g(x,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \tag{2}$$

and the standard deviation parameter σ controls the width of the Gaussian. We empirically set σ to $\frac{1}{2}$ the size of the local context window.

Geographic Specificity Score $S_{geo}(l)$

When multiple locations appear in the same local context window but are at different administrative levels in the geographic hierarchy, the more specific location (lower in the hierarchy) is usually more relevant. Consider the following context window which include multiple locations (highlighted in green):

..just for the sheer size of [ ] *... I was expecting, what I saw in* [ ] *and* [ ] *and even even* [ ]*, but this neighborhood of* [ ]*, even though it's so close to the center of the city...*

In this case *Tokyo* is a less specific location and therefore less relevant than sub-districts *Shibuya, Shinjuku* and *Harajuku* which in turn are less specific than the neighborhood *Koenji*. Our geographic specificity score is therefore based on the depth of the location l in the geographic tree we compute in Step 3 of the algorithm.

Sometimes podcasts mention locations outside the primary administrative level for comparison. In this context window from our example podcast, the host compares neighborhoods in *Tokyo* to those in *New York*

You also have these companies in [ ]*... It's a shopping neighborhood ..Very upscale. It's like our* [ ] *or* [ ] *...whereas* [ ] *feels more like* [ ]*..you also find very different but equally intense architectures...*

In a podcast about *Tokyo*, images of *New York* can be confusing. To avoid distracting viewers with locations that are not relevant to the podcast, our geographic specificity score penalizes locations that fall outside the geographic subtree rooted at the primary administrative level p (as we computed in Step 3).

We define the geographic specificity score for a location l as

$$S_{geo}(l) = subtree(p,l)\frac{d_{geo}(l,p)}{h_p} \tag{3}$$

where $subtree(p,l)$ is a sign function indicates whether l belongs to the subtree of primary administrative level p, $d_{geo}(l,p)$ is the number of edges in the geographic tree between l and p, and h_p is height of the subtree rooted at p.

Subject Relevance Score $S_{sbj}(l)$

When a location is presented as the subject of a sentence, it is

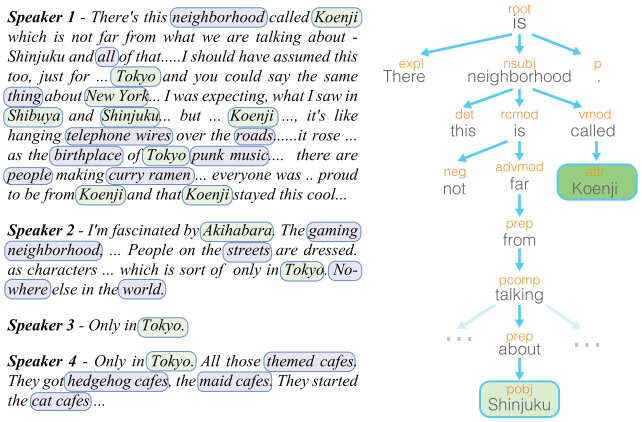

Figure 5. left: transcript of a clip of Tokyo podcast; right: syntactic tree of the first sentence of speaker 1.

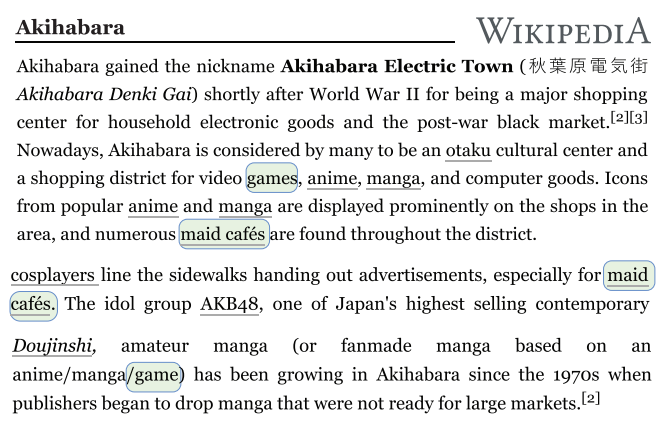

Figure 6. Wikipedia page of "Akihabara", circled words are relevant details that can be matched with entities from the transcript.

usually more relevant than a location presented as an object. Consider the following context window,

..there is one neighborhood I have to plug, because I wrote a piece about it. There is this neighborhood called ____ *which is not far from what were are talking about -* ____ *and all of that. It's like two stops on the train west..*

Two locations are mentioned explicitly, with *Koenji* serving as the subject, and *Shinjuku* serving as an object for comparison. In this case *Koenji* is the more relevant location.

To compute the subject relevance score, we first identify the longest complete sentence containing a location within the local context window and we recover its syntactic tree structure using the Google NLP toolkit [7]. This syntactic tree (also known as dependency tree) describes the grammatical relationships among the words of a sentence with the root corresponding to the main verb of the sentence, and the edges between the words indicating syntactic relationships (Figure 2). The closer a word is to the root in the syntactic tree, the more directly the word is related to the main verb, and the more relevant the word is to the subject of the sentence.

We define the subject relevance score for a location l as

$$S_{sbj}(l) = 1 - \frac{d_{sbj}(l,r)}{h} \tag{4}$$

where $d_{sbj}(l,r)$ is the distance (i.e. number of edges) between l and the root of the syntactic tree r, and h is the height of the syntactic tree.

Detail Relevance Score $S_{dtl}(l)$

In discussing a location, people often describe location specific details instead of referring to the location by name. For example, in the following local context window,

I'm fascinated by ____ *... People on the streets are dressed in* ____ *as characters from ..* ____ *..which is.. one of those only in* ____ *. Nowhere else in the world. Only in* ____ *. Only in* ____ *. All those* ____ *. They got* ____ *. The* ____ *...* ____ *...*

even though the location *Tokyo* is mentioned multiple times, many of the details, such as *costumes, games, themed cafes* are about *Akihabara* . It is *Akihabara* that is the most relevant location for this window. Our detail relevance score is designed to increase the relevance of a location when it is being discussed in terms of its specific details rather than by name. Our approach uses *term frequency inverse document frequency (TF-IDF)* [21], a numerical measure of the importance of a word to a document in a corpus. We use TF-IDF to assess how much each non-location word w in the local context window W can serve as a location specific detail.

Specifically, for a location l we treat its Wikipedia entry e_l as a source of location specific details (Figure 6). We have found that Wikipedia pages usually provide good quality documentation of the demographics, history, geography, culture and other notable details of a location. We then compute the detail relevance score for l as

$$S_{dtl}(l) = \sum_{w \in W} TF(w, e_l) IDF(w, C) \tag{5}$$

where C is the corpus of Wikipedia entries for all locations that appear in the transcript as well as 100 popular locations recognized in travel websites[1]. The term frequency term $TF(w, e_l)$ measures the frequency with which a non-location word w appears in the Wikipedia entry for location l with the inverse document frequency term $IDF(w, C)$ measures the frequency with which w appears in the corpus of location entries. Thus, non-location words w that are descriptive of location l much more than other locations will produce relatively high TF-IDF scores and thereby increase the detail relevance of location l.

Note that a high TF-IDF score for a word w with respect to the location l implies that w is an important detail for the location. In our example, these relevant detail words include *costumes, games, themes cafes* produce high TF-IDF scores compared to less relevant detail words like *streets*. For each location l we maintain a list the most relevant detail words w with TF-IDF score ≥ 0.2.

[1] https://traveleye.com/top/best-travel-destinations

Scoring Visually Significant Entities

Our relevance score for visually significant entities S_{vse} is similar to our location relevance score S_{loc}. For each candidate VSE v that appears in the local context window of transcript word t we compute

$$S_{vse}(t,v) = S_{fsh}(t,v)(w_{spc}S_{spc}(v) + w_{sbj}S_{sbj}(v) + w_{dtl}S_{dtl}(v)).$$

The only new term is S_{spc}, which computes a VSE specificity score. The freshness term S_{fsh}, the subject relevance term S_{sbj} and the detail relevance term S_{dtl} are computed exactly as we did for location, but with respect to v rather than a l. For example, in computing S_{dtl} we consider the Wikipedia entry e_v for v as the document. We empirically set $w_{spc} = 1$, $w_{sbj} = 0.4$ and $w_{dtl} = 3$.

Specifcity Score $S_{spc}(v)$

In general VSEs do not appear in geographical tree of locations and therefore we designed the VSE specificity term S_{spc} as a replacement for the geographic specificity term S_{geo} when scoring VSEs. Consider the following local context window,

.. is tall, for the most part. There are short parts of it. But, is also the place where exists. And that of like, on the other side of the . And it's this teeny, tiny very old , , that is in the of all that,

Here *Yakitori Alley* is a specific VSE, whereas other VSEs such as *sort* and *feeling* are generic nouns. To determine the specificity of a VSE v, we look up whether there is a Wikipedia entry e_v for the VSE, and consider it *identifiable* if there is.

We also check whether the VSE might be a distraction by examining how closely it is associated with the primary location p (as identified in Step 3 of our algorithm) discussed in the podcast. In the following local context window

... what I've heard was with the , how everything is in and English now. And apparently ... in for the ... I was expecting to be pulling out my translate every five minutes...

both *Olympics* and *Google* are identifiable VSEs. In this case *Olympics* is relevant because when the podcast was recorded the 2020 Summer Olympics were going to be hosted in *Tokyo* and Tokyo is the primary location of the podcast. This fact is documented in the Wikipedia entry e_v for the VSE *Olympics* as well as the Wikipedia entry e_p for the primary location *Tokyo*. Both the VSE and the primary location appear in both Wikipedia entries. In contrast the Wikipedia entry for the VSE *Google*, and the primary location *Tokyo* are completely disjoint. Therefore, for identifiable VSEs, we compute the degree to which the Wikipedia entries for the VSE v and the the primary location p overlap as VSE specificity score

$$S_{spc}(v) = TF(v,e_p)IDF(v,C) + TF(p,e_v)IDF(p,C), \quad (6)$$

where corpus C is the same as the location corpus we used in computing S_{dtl} for locations. The first TF-IDF term measures how much the VSE v appears in the Wikipedia entry for the primary location p while the second TF-IDF term measures how much p appears in the Wikipedia entry for v. VSEs that are not identifiable receive 0 as their specificity scores.

Figure 7. (left) a query term with location and visual entity returns generic images. (right) adding the detail of the visual entity returns more accurate images.

Step 5: Compose Query and Acquire Relevant Images

At the end of Step 4, for each transcript word t we have computed the top-scoring location l_{top} and top-scoring VSE v_{top}. Our goal in Step 5 is to acquire a relevant set of high-quality images for each transcript word using l_{top} and v_{top}. Our approach is to first form an image search query, and then filter the resulting image set to ensure that only the most relevant and highest quality images remain.

For each transcript word t, we start by composing l_{top} and v_{top} into an initial image search query. In practice however, we have found when the relevance scores S_{loc} and S_{vse} for l_{top} and v_{top} respectively are below a threshold of 0.2, the top location and VSE often do not match the content discussed in the video. We observed that this usually happens when podcast hosts start to talk about tangential content, where no locations, VSEs, or their details are mentioned or discussed. Thus, if $S_{loc}(t,l_{top}) < 0.2$ we replace l_{top} with the primary location p of the podcast in the search query. If $S_{vse}(t,v_{top}) < 0.2$ we remove v_{top} from the search query.

Consider for example, the following local context window.

..that you only really get in ... I think you and I have talked about this. At places, you know the , another that people might not be used to is the . The . Right? Yeah... Confused the out of me. Yeah. I saw them, but I didn't actually do it. .

Here, *Tokyo* is the top-scoring location l_{top} and *ramen* is the top-scoring VSE v_{top}. However, even though the initial search query *(Tokyo + Ramen)* retrieves appropriate images of *Ramen* (Figure 7 left), we have found that we can further improve the query by adding additional detail words. As noted in Step 4, we capture a list of relevant detail words when we compute the detail relevance score S_{dtl} for l_{top} and v_{top}. We append the word with the highest detail relevance score for VSE v_{top} if it exceeds 0.4. In this above example, our algorithm appends *machine* to the search query, as *machine* has been mentioned several times in the Wikipedia entry for *ramen*, and the search query becomes *(Tokyo + ramen + machine)* (Figure 7 right).

We use the Microsoft Bing Image Search Engine [23] to obtain top 100 high-resolution images with the formed query. We utilize the search filters provided by Bing to search for photos and eliminate illustrations, graphic designs and other synthetic images. For best visual quality, Crosscast does not filter images by copyright licenses, but it is trivial to specify license constrains with image search engines. To eliminate photos with watermarks, we filter out images from stock photography websites (e.g. alamy.com, istockphoto.com, etc.) based on whether the image url contains one of these website addresses.

Figure 8. Short segment is combined with previous segment, and long segments are even split to show multiple images.

For each transcript word t this process produces a set of 20 to 30 high-quality photographs ranked by relevance to the query. Note that in practice adjacent transcript words often generate identical search queries and therefore to improve performance we only run the image search and filtering steps when we encounter a new search query.

Step 6: Select Visuals for Each Transcript Word

The search query for each transcript word tends to change only when the sliding window includes a new location or VSE. We call such changes transition points. We start by segmenting the podcast in time at the words corresponding to these transition points. At the start of each transition point we select either the maps of the locations or the highest ranked photograph from our set of high-quality images produced in Step 5. While the resulting sequence of visuals is guaranteed to be relevant to the underlying podcast, this approach does not account for the length of time the photographs are onscreen.

We found that images and maps should remain onscreen between 3.5s and 7.5s to ensure that viewers have enough time to absorb the visual informaiton and to avoid monotony. To achieve this, we first remove any transition point at the start of a segment that is less than 3.5s in length and combine the short segment with its longer, earlier neighbor. In this case we show maps and images relevant to the earlier segment combined length of time. Whenever a segment is longer than 7.5s we evenly split it into shorter segments until each one is shorter than 7.5s, and then cycle through the maps and the ranked list of photographs produced in Step 5, as shown in Figure 8. Text labels are also added to facilitate comprehension if the location and VSE appear for the first time.

When describing a location, it is helpful to see the map of location to understand its geographical orientation within the surrounding area. When a top-scoring location entity or VSE of a location point is mentioned for the first time, Crosscast displays Google Map of location's parent region and marks the position of the location on the map. In some cases, podcast hosts describe a location's orientation by describing its spatial relationship to other locations. To facilitate the visual comparison among locations, when more than 3 locations are mentioned within a segment, Crosscast displays Google Map with all the mentioned locations marked on the map.

RESULTS

Crosscast can automatically generate audio-visual travel podcast with the input of podcast audios and their time-aligned transcript. Figure 1 and 9 show the example frames of the audio-visual podcasts, their videos can be found in the supplemental material. As shown, Crosscast automatically adds sequences of images and maps relevant to podcast content. We found that top-scoring locations and VSEs occasionally do not match the location and VSE being discussed in the transcript. To gauge how accurately Crosscast extracts the most relevant locations and VSEs, we conducted the following evaluation.

Technical Evaluation

We quantitatively measure the performance of our automatic location and VSE scoring algorithms by comparing their output against a gold standard selections made by experts. We gathered a corpus of 12 representative travel podcasts from 8 travel podcast hosts [10, 24, 25, 28, 29, 38, 39, 40]. The entire talking time of all the podcasts was 211.08 minutes, and the transcripts contain 27624 words in total, with 1214 location entities, and 2204 entities of other types.

Gold Standard Labels

To acquire gold standard labels of both locations and VSEs for these travel podcasts, we recruited two participants to independently label the most relevant locations and VSEs for all the travel podcasts. Labelers were asked to select locations and VSEs that worth shown for each sentence. They then worked together to form a consensus on the labels. This process produced the gold standard labels for our evaluation.

Algorithmic Methods Performance

We compare our algorithmic methods against a baseline condition, which always selects the newly mentioned location and visual entities as the most relevant for any moment of the podcast. We also examine the usefulness of each feature through an ablation evaluation where we remove a single term at a time and examine its performance. To compute matches to the gold labels generated for each sentence, we aggregate the different labels within each sentence generated by Crosscast and compute the accuracy. Overall, our algorithm achieves an average precision of 0.89 and an average recall of 0.77 for location selection with and average $F_1(location) = 0.82$. It achieves an average precision of 0.77 and an average recall of 0.95 for VSE selection with and average $F_1(keywords) = 0.85$. The F_1 scores of all comparison condition are shown in Figure 10.

The comparison shows our methods significantly outperform the baseline condition. The ablation of each term all leads to decrease of performance, among which the ablation of the freshness causes the largest decrease. While the weights of the scoring terms were set empirically, the combination we selected, although not optimal, performed well in the evaluation. In our tuning process, we found that one set of weights that work well for some podcasts may underperform for others, likely because of the different styles of the podcasts. This suggests that for optimal performance, the weights could adapt to different podcast styles if sufficient training data are provided.

USER STUDY

We conducted a user study with 28 participants (13 female, aged from 22 to 51) to evaluate the image relevance, quality of synchronization, image relevance, and overall quality for representative clips from Crosscast-generated audio-visual travel podcasts in comparison to clips from *audio-only* and expert *manually generated audio-visual* travel podcasts for the same audio content. Participants' familiarity with travel podcast was not considered during recruitment as we intended to gauge the generated results with general population.

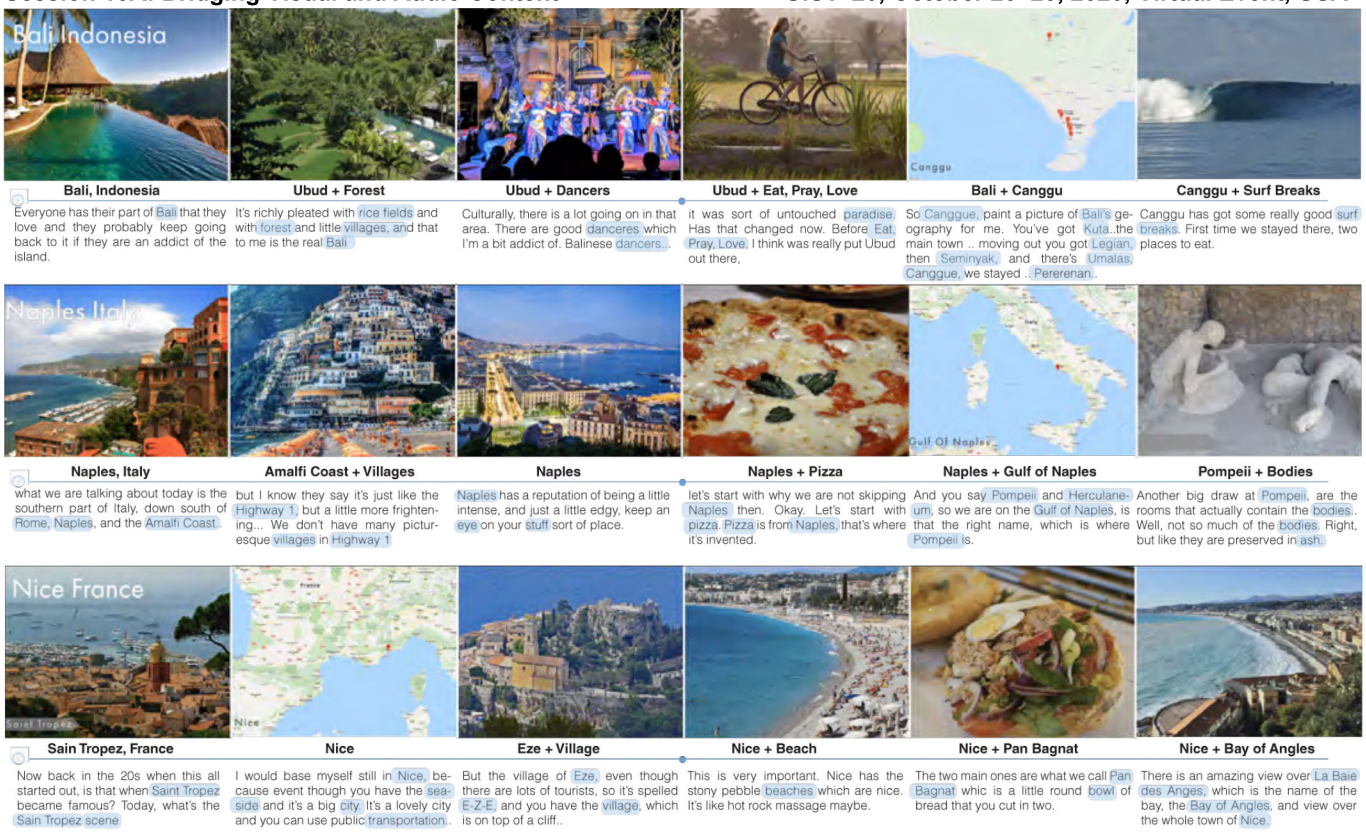

Figure 9. Results automatically generated by Crosscast for travel podcasts about Naples and French Riviera.

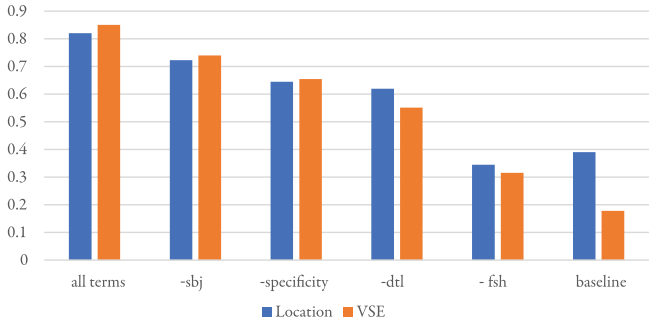

Figure 10. F1 scores for all comparison conditions.

User Study Methodology

We asked participants to play two sets of three clips sampled from travel podcasts about Berlin and Tokyo, respectively. Each set consisted of one audio-only clip, one manually generated audio-visual clip, and one Crosscast-generated clip. We used audio sampled from the beginning or middle of travel podcasts and restricted clip length to approximately 2 minutes to avoid user fatigue. One paper author with formal training in video editing manually created clips with images queried from the same search engine (i.e. Bing) and the video format was restricted to fading transitions between static images with one text heading per image. Search keywords, image selection, image order, and transition timing were left to the discretion of the manual editor. The video editor took an average of 4 hours to produce one complete sequence of images and titles for a 15 minute podcast whereas running Crosscast on the same podcasts took 12 seconds to generate. Note that neither the manual clips nor the Crosscast clips include maps as we ran the study before we had incorporated maps into Crosscast.

The user study randomized the order participants were shown audio-only, manually generated, and Crosscast-generated clips within each set. After viewing each clip, we asked participants to rate the following questions, "the images were relevant to the audio content"; "the image transitions were well aligned with the audio content", and "the images are high quality and consistent" on a 7-point Likert scale *(1-strongly disagree to 7-strongly agree)*. After viewing each set of three clips, we asked participants to rank all three clips in order of preference *(1-most preferred to 3-least preferred)*.

User preference for audio-visual travel podcasts

Participants ranked Crosscast-generated clips higher than audio-only clips (85.7%), and manually-generated clips higher than audio-only (92.9%). This result suggests that the addition of visual content is preferred by the participants. In the majority of cases, participants preferred manually generated clips over Crosscast-generated clips (67.9%). This result is not a surprise, and suggests that additional work is required to match the quality of hand-generated audio-visual travel podcasts.

The quantitative results suggest that people view visuals as a useful addition to audio podcasts. This sentiment was reflected in participants' qualitative responses. Participants described how visuals improved the experience of listening to the audio podcast. For example, P15 commented *"the images give so much more detail about the place that I can't get from the audio"*, and P17 noted that *"I wouldn't have guessed what Golden*

Gai looks like just by listening, the visuals make me want to visit it." Furthermore, most people felt Crosscast-generated clips improved the experience of listening to the audio content, noting that they found them "interesting" compared to the audio-only clips.participants also noted that even when images transitions were slightly off and images were a bit redundant, they were still "more helpful" than audio alone.

Image Relevance

For Crosscast clips, 82.1% participants responded favorably to the statement that the images were relevant to the audio content (a score of 5 or higher), and 89.3% for manually generated clips. Crosscast relies on the order of results returned by the search engines for relevant images without understanding the content of the images, whereas the human-expert can select images that can best depicts the underlying content. Yet, not all participants responded favorably even for the manually created clips, which suggests that selecting relevant images is an ambiguous task even for human editors.

Audio-Visual Synchronization

For Crosscast clips, 75% responded favorably to the statement that the image transitions were well aligned with the audio, and 89.3% favorably for the manually generated clips. Participants who disliked the audio-visual synchronization in Crosscast, commented that visual transitions tend to lag behind the topic transitions in the audio. While humans can easily recognize the transition of topic in the audio content, Crosscast relies on informative words, modeled on travel podcast structural properties, to trigger transitions, which often slightly lags behind the most appropriate transitions moments. Similar to image relevance, three participants felt unsatisfied with transition timing in the manually generated clips, which suggests the ideal synchronization is subject to individual preference.

Image Quality and Consistency

82.1% of participants responded favorably to the statement that the images are high quality and consistent for Crosscast clips, and 92.9% for manually created clips. Participants attributed their positive decisions about Crosscast clips to aesthetic factors like "high-resolution", "sharper", and "brighter". However, they also took issue with Crosscast clips that presented relevant but repetitive content, stating that they would rather see greater variety than "redundant images" or "the same thing" with different framing. While Crosscast seeks to avoid showing the same images, completely ruling out redundant images require understanding of the image content which Crosscast is not capable of. Human-expert, on the other hand, can easily recognize and select diverse and high-quality images.

Summary

Overall our study suggests that automatically generated Crosscast image sequences are a desirable addition to travel podcasts, and an improvement over audio alone. Manually generated audio-visual podcasts outperformed Crosscast-generated podcasts, but they also required substantial time to produce. Crosscast increases engagement and informational value of travel podcasts while reducing the effort required to source and synchronize location images with audio.

LIMITATIONS

Our technical and user evaluations demonstrate the advantages and benefits of the Crosscast system, but also suggest limitations and opportunities for improvement.

Working with travel podcasts that are not about locations. Some travel podcasts do not focus on describing the visual details of particular locations, but instead cover other travel related topics, such as transportation (e.g. what is the best airline), accommodation, safety, cost, and so on. Since Crosscast is designed to select most relevant location and visually significant entities of the location, it falls short on providing relevant images. Extending Crosscast to work for such non-location based travel podcasts is an open direction for future work.

Better handling of multiple locations. Another challenge for Crosscast is managing transitions between images when multiple locations are mentioned quickly one after another, or compared to one another. Crosscast usually cannot show all the locations in such cases because of the minimum timing constraints. An alternative might be to use a split screen when two or more locations are mentioned in quick succession, or to perhaps present all the locations on an overview map.

FUTURE WORK

Crosscast opens a few larger directions for future work.

Generalizing to other formats. Crosscast is designed to add visuals to travel podcasts. But travel information is often consumed in other formats such as magazine articles, blog posts, etc. It may be possible to extend the techniques used in Crosscast to add photos and maps to such travel articles.

Generalizing to other podcast topics. Automatically curating images can be beneficial for other podcasts topics, such as news, history and design. However, curating the best visual content requires focus on the specific nature of the content. We have found the geographic specificity of location is useful for retrieving relevant images for travel podcasts, which is not applicable to other domains. Nevertheless, our approach of first identifying the most relevant query elements, and then designing algorithmic methods to extract such elements is a generic process that is applicable to other domains.

Generalizing to other visual styles. In user evaluation, we constraint expert-made clips to similar style of Crosscast for comparison. In practice, domain experts create visual content of diverse and compelling styles. One future direction is to learn from best examples created by experts and enhance Crosscast with different types of visual content and effects.

CONCLUSION

Adding visuals of relevant locations and visually significant entities can substantially improve people's engagement with audio travel podcasts, but manually sourcing, aligning, and labeling travel images with podcast audio is challenging and laborious. We have presented Crosscast, a system that automatically creates audio-visual podcasts using NLP techniques combined with geographic semantics. Crosscast enables podcast creators *or* listeners to quickly and easily produce visuals for travel podcast audio, making it possible for a wider range of people to visualize travel destinations.

REFERENCES

[1] 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

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5828–5839.

[3] Ohad Fried and Maneesh Agrawala. 2019. Puppet Dubbing. In *Proceedings of the Eurographics Symposium on Rendering (EGSR '19)*. Eurographics Association, 10.

[4] Ohad Fried, Ayush Tewari, Zollhöfer Michael, 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.* (Aug. 2019).

[5] Daniel Funke. 2017. With 1.7 million monthly downloads, the Reveal podcast wanted to give its listeners more. The solution: an SMS chatbot. (2017). https://bit.ly/2lU0yMx

[6] Brian Garrod. 2009. Understanding the Relationship between Tourism Destination Imagery and Tourist Photography. *Journal of Travel Research* 47, 3 (Feb 2009), 346–358. DOI: http://dx.doi.org/10.1177/0047287508322785

[7] Google. 2019a. Google NLP Toolkit. (2019). https://cloud.google.com/natural-language/.

[8] Google. 2019b. Google Places API. (2019). https://developers.google.com/places/web-service/intro.

[9] Floraine Grabler, Maneesh Agrawala, Robert W. Sumner, and Mark Pauly. 2008. Automatic Generation of Tourist Maps. *ACM Trans. Graph.* 27, 3, Article 100 (Aug. 2008), 11 pages. DOI: http://dx.doi.org/10.1145/1360612.1360699

[10] Peter Greenberg. 2019. Peter Greenberg WorldWide. (2019). https://petergreenberg.com.

[11] Ulrike Gretzel and Kyung Hyan Yoo. 2008. Use and Impact of Online Travel Reviews. In *Information and Communication Technologies in Tourism 2008*, Peter O'Connor, Wolfram Höpken, and UlrikeEditors Gretzel (Eds.). Springer Vienna, 35–46.

[12] The Guardian. 2019. Why we want to make podcasts better. (2019). https://www.theguardian.com/info/2019/jun/12/why-we-want-to-make-podcasts-better.

[13] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7986–7994.

[14] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J. Mysore. 2019. B-Script: Transcript-based B-roll Video Editing with Recommendations. *CoRR* abs/1902.11216 (2019). http://arxiv.org/abs/1902.11216

[15] IBM. 2016. IBM Speech to Text Service. https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/speech-to-text/. (2016). Accessed 2016-12-17.

[16] Dhiraj Joshi, James Z Wang, and Jia Li. 2004. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*. ACM, 119–126.

[17] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Ranking and retrieval of image sequences from multiple paragraph queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1993–2001.

[18] 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

[19] Mackenzie Leake, Hijung Valentina Shin, Joy O. Kim, and Maneesh Agrawala. 2020. Generating Audio-Visual Slideshows from Text Articles Using Word Concreteness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. DOI: http://dx.doi.org/10.1145/3313831.3376519

[20] Jihyeon Janel Lee, Mitchell Gordon, and Maneesh Agrawala. 2017. Automatically Visualizing Audio Travel Podcasts. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 165–167. DOI: http://dx.doi.org/10.1145/3131785.3131818

[21] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[22] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. 2017. Attentive semantic video generation using captions. In *Proceedings of the IEEE International Conference on Computer Vision*. 1426–1434.

[23] Microsoft. 2019. Microsoft Bing Image Search. (2019). https://azure.microsoft.com/en-us/services/cognitive-services/bing-image-search-api/.

[24] Flight of Fancy. 2019. Flight of Fancy. (2019). http://www.traveller.com.au/.

[25] Extra Pack of Peanuts. 2019. Extra Pack of Peanuts. (2019). https://extrapackofpeanuts.com.

[26] Yanwei Pang, Qiang Hao, Yuan Yuan, Tanji Hu, Rui Cai, and Lei Zhang. 2011. Summarizing tourist destinations by mining user-generated travelogues and photos. *Computer Vision and Image Understanding* 115, 3 (2011), 352 – 363. DOI:http://dx.doi.org/https://doi.org/10.1016/j.cviu.2010.10.010 Special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics.

[27] 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

[28] America's Nation Park Podcast. 2019a. America's Nation Park Podcast. (2019). https://nationalparkpodcast.com/.

[29] Indie Travel Podcast. 2019b. Indie Travel Podcast. (2019). https://indietravelpodcast.com/.

[30] Edison Research. 2019. The Podcast Consumer 2019. (2019). https://www.edisonresearch.com/the-podcast-consumer-2019/.

[31] Steve Rubin and Maneesh Agrawala. 2014. Generating Emotionally Relevant Musical Scores for Audio Stories. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 439–448. DOI: http://dx.doi.org/10.1145/2642918.2647406

[32] Steve Rubin, Floraine Berthouzoz, Gautham Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: Musical Underlays for Audio Stories. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 359–366. DOI: http://dx.doi.org/10.1145/2380116.2380163

[33] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based Tools for Editing Audio Stories. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 113–122. DOI: http://dx.doi.org/10.1145/2501988.2501993

[34] Stevan Rudinac, Alan Hanjalic, and Martha Larson. 2011. Finding Representative and Diverse Community Contributed Images to Create Visual Summaries of Geographic Areas. In *Proceedings of the 19th ACM International Conference on Multimedia (MM '11)*. ACM, New York, NY, USA, 1109–1112. DOI: http://dx.doi.org/10.1145/2072298.2071950

[35] Sarah Schmalbach. 2018. Uncovering the potential of mobile audio: a new experimental player, and a new show. (2018). https://bit.ly/2orYuco.

[36] Hijung Valentina Shin, Wilmot Li, and Frédo Durand. 2016. Dynamic Authoring of Audio with Linked Scripts. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 509–516. DOI: http://dx.doi.org/10.1145/2984511.2984561

[37] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo Tourism: Exploring Photo Collections in 3D. In *ACM SIGGRAPH 2006 Papers (SIGGRAPH '06)*. ACM, New York, NY, USA, 835–846. DOI: http://dx.doi.org/10.1145/1179352.1141964

[38] Rick Steves. 2019. Rick Steves Europe. (2019). https://www.ricksteves.com.

[39] Amateur Traveler. 2019a. Amateur Traveler. (2019). https://amateurtraveler.com.

[40] Condé Nast Traveler. 2019b. Condé Nast Traveler. (2019). https://www.cntraveler.com/.

[41] 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

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

[43] John Urry. 1992. The Tourist Gaze "Revisited". *American Behavioral Scientist* 36, 2 (1992), 172–186. DOI:http://dx.doi.org/10.1177/0002764292036002005

[44] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, and Ariel Shamir. 2019. Write-a-Video: Computational Video Montage from Themed Text. *ACM Trans. Graph.* 38, 6, Article Article 177 (Nov. 2019), 13 pages. DOI:http://dx.doi.org/10.1145/3355089.3356520

[45] Haijun Xia. 2020. Crosspower: Bridging Graphics and Linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. ACM, New York, NY, USA. DOI: http://dx.doi.org/10.1145/3379337.3415845