Query-by-Emoji Video Search

Cappallo, S.; Mensink, T.; Snoek, C.G.M.

DOI
10.1145/2733373.2807961

Publication date
2015

Document Version
Final published version

Published in
MM'15

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Link to publication

Citation for published version (APA):
https://doi.org/10.1145/2733373.2807961
Query-by-Emoji Video Search

Spencer Cappallo† Thomas Mensink†‡ Cees G. M. Snoek†‡
†University of Amsterdam ‡Qualcomm Research Netherlands
{cappallo, tmensink, cgmsnoek}@uva.nl

ABSTRACT
This technical demo presents Emoji2Video, a query-by-emoji interface for exploring video collections. Ideogram-based video search and representation presents an opportunity for an intuitive, visual interface and concise non-textual summary of video contents, in a form factor that is ideal for small screens. The demo allows users to build search strings comprised of ideograms which are used to query a large dataset of YouTube videos. The system returns a list of the top-ranking videos for the user query along with an emoji summary of the video contents so that users may make an informed decision whether to view a video or refine their search terms. The ranking of the videos is done in a zero-shot, multi-modal manner that employs an embedding space to exploit semantic relationships between user-selected ideograms and the video’s visual and textual content.

1. INTRODUCTION
The internet is perpetually inundated with new videos, but the methods for interfacing with this exponentially growing data remain static. There is a need to investigate new, alternative approaches for exploration of large video datasets, with a focus on interaction that translates seamlessly to new mobile platforms such as smartwatches. We present Emoji2Video as a new means of interaction and representation for large databases of multi-modal video and text.

Emoji2Video allows users to query any dataset of visual media using emoji. Emoji present a unique opportunity for visual representation and exploration. As a set of intuitive iconography, emoji are able to transcend language, age, and culture boundaries. Furthermore, the use of ideograms maintains a strictly visual language of interaction between search construction and video results. In this way, using emoji helps to decrease the semantic gap between the query grammar and the visual media. Furthermore, the native support of emoji from companies like Google and Apple makes them uniquely well-positioned as a means for video search via smart watches and phones, with a square form factor that integrates seamlessly into touch interfaces.

Emoji are becoming increasingly popular as a form of communication, and have steadily gained the attention of major companies. Microsoft’s Bing has begun supporting search-by-emoji, while Instagram has recently investigated users’ use of emoji within their network. Our demo uses emoji as a set of ideograms, but the underlying architecture is zero-shot and therefore trivially adaptable to any arbitrary set of ideograms. Emoji are simply an ideal test case due to their pervasive presence, which renders users instantly familiar with the range of ideograms available. Furthermore, the popularity of emoji suggests that they have sufficiently descriptive semantics for a wide range of scenarios. These properties make them an ideal test bed

2http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji
Users construct emoji queries through a web app interface which are used to return a ranked list of videos. Thumbnails of the videos are shown along with at-a-glance emoji representations of video content. Users can then either select a video for viewing or refine their query. An exploded view of several frames from the video can also be viewed, along with their respective frame-level emoji representations.

2. EMOJI2VIDEO SEARCH ENGINE

Our technical demo system utilizes emoji-based search queries to return videos from the VideoStory46k dataset [2]. This dataset consists of 46,000 videos harvested from YouTube, and as such is a reasonable proxy for the breadth and nature of content on video-sharing websites. The system is multi-modal, relying on both the frames of the videos and the video titles to generate searchable emoji-based representations. This is an extension of our previous work [1] on zero-shot emoji representation to a video search application.

Our approach relies on a word2vec [3] semantic embedding space, within which we can relate visual, textual, and emoji concepts. Visual evidence is collected through the application of deep-learning based concept detectors [4] on the video frames, and textual evidence is harvested from the title text of the YouTube videos.[1]

2.1 User Experience

In our demo, users are presented with a search interface similar to those commonly found on the mobile web. However, instead of users typing text strings to compose the search query, they are presented with a panel of emoji to choose from. As users compose and refine their query, videos are ranked according to this query. The results are displayed in an ordered list with a thumbnail of each video, along with an emoji-based summary of the video contents. The emoji representation is generated using the entirety of the video along with its title, and so may capture information about the content of the video that is not present in a single thumbnail image. This at-a-glance visual summary allows users to make an informed decision about which of the returned videos they want to view.

As the user refines their query, suggestions for other emoji to include are made by grouping the emoji within the semantic space. When a user has found a video they are interested in, they can choose to view that video, or see an expansion of some frames of the video along with their emoji representations. An overview of the core experience along with the system’s results for an example query is given in Figure 2.

2.2 Video Summarization

Emoji-based representations are calculated on a frame-level across the video. At the frame-level, a smaller weight is applied to evidence from the video’s title than to the local, frame-specific visual information. The mean of the frame-level representations is used to calculate the video-level emoji representation. In this way, the video-level representation provides an ideogrammatical summary of the visual content across the video. The demo is run as a web app, and could therefore be accessed by any internet-capable devices networked to the server, providing compact emoji summarization of videos for devices where screen real estate may be at a premium.

Our demo presents a new way of interacting with large numbers of videos. Emoji2Video allows us to leverage the increasingly familiar emoji ideograms to provide an interface for quickly exploring video datasets in a form that can be used on smart watches, smart phones, and by users of any language, age, or culture.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COM-MIT.

3. REFERENCES