June 11, 2019

Search snippets that include rich media

Victor Cărbune

Thomas Deselaers

Search snippets that include rich media

ABSTRACT

Information pertinent to a search query can be in the form of rich media, e.g., video, audio, images, text, etc. However, search snippets or summaries currently comprise mainly text. This disclosure describes techniques that surface rich media snippets within search results. The search snippets provided using the described techniques include relevant segments of rich media, e.g., audio or video, summarized to sections that are pertinent to the search query.

KEYWORDS

- Search snippet
- Rich media snippet
- Media summary
- Scene recognition
- Speech recognition
- Machine learning
- Search indexing
- Granular metadata

BACKGROUND

Information pertinent to a search query can be in the form of rich media, e.g., video, audio, images, text, etc. However, to a large extent, search snippets currently comprise text. Rich media is not particularly well exposed to users with the same information density as text snippets. For example, it may be the case that a segment of a video is relevant to a user search query. If such a video is surfaced in a search result, the user still has to watch and skip through

other, non-relevant portions before they can view portions of the video that are relevant to their query.

DESCRIPTION

This disclosure describes techniques to surface rich media snippets within search results. Per the techniques, machine learning models are utilized to summarize rich media content (or segments thereof). The generated summaries are provided to search-indexing services. Summarized rich media, e.g., short audio, cropped images, automatically played videos, etc., appear in the search results when such media is pertinent to a search query. The techniques are implemented using the following components, described in greater detail in the later sections of this disclosure.

- **Machine learning models** to provide granular metadata to search-indexing services.
- **Search indexing and retrieval pipeline** to incorporate granular metadata provided by the machine learning models.
- **User interface** to surface rich-media snippets.

Machine learning models to provide granular metadata to search-indexing services

The machine learning models include image and audio/video understanding models that extract information about objects embedded in rich media content and the coordinates or points of time where such objects are exposed. For example, image recognition models identify objects within an image and encapsulate them within annotated bounding boxes. For sequence data such as audio or video, identified objects are encapsulated within bounding boxes as well as within time intervals.

Example machine learning models that can be applied on rich media content include:

- Optical character recognition (OCR) models that provide bounding boxes with

annotations around text that appears in images or video.

- Activity recognition models that can be run on whole images, or on chunks of video or audio.
- Face recognition models that annotate relevant parts of an image or video.
- Media captioning models that provide annotations to chunks of videos, e.g., “a dog is chasing a cat” from 0m14s to 1m15s; “kids are playing” between 5m15s and end; etc.

The output of the machine learning models includes granular metadata in a form that is suitable for processing by the search-indexing pipeline, e.g., segment-wise, descriptive annotations of content with bounding-box coordinates as necessary.

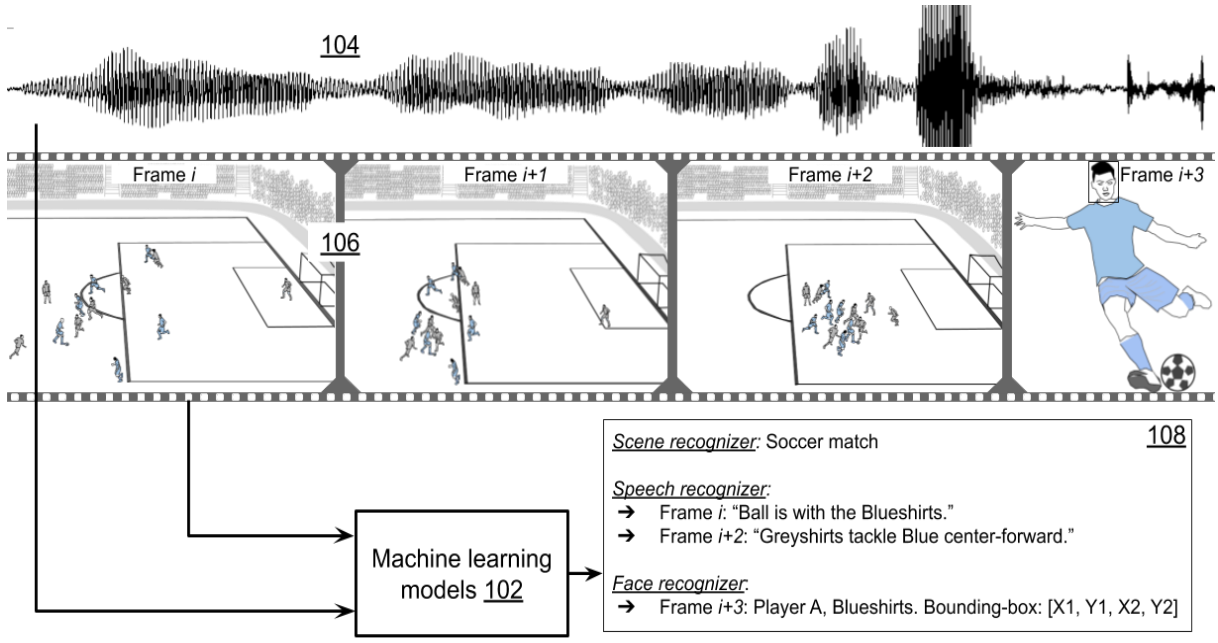

Fig. 1: Video annotation using machine learning models to produce granular metadata

Fig. 1 illustrates an example of annotations produced by machine learning models (102). Video content (106) that is accompanied by audio commentary (104) is analyzed using the machine learning models. In this example, a scene (or activity) recognition model recognizes the video as that of a soccer match; a speech recognition model recognizes audio commentary

relating to sporting action; and a face recognition model recognizes a player and the frame and bounding box within which the player appears (108).

Search indexing and retrieval pipeline to incorporate granular metadata provided by the machine learning models

Annotations, metadata, and other signals generated by the machine learning models are incorporated with traditional signals and document annotations that are used for searching. A jointly-ranked index is created that corresponds to a user query at a document level (e.g., unsegmented content) followed by an index into a specific match to the relevant content segment.

For learned search-ranking models, metadata generated by the machine learning models need not be in human-readable form. Rather, the output of the content analysis models can be directly used as an embedding, e.g., an abstract machine learned representation can be provided and utilized directly for search ranking. An advantage of directly using embeddings is that the models can be trained end-to-end while easily leveraging pre-trained domain models.

The search-indexing and retrieval component enables the direct exposure of media snippets to the user. For example,

- for a query “dogs running in the park on a sunny day”, a search result can include metadata such as (video id: 12345, segments: [0m2s, 0m50s], [5m34s 5m50s], shortened-video: ..., metadata: ‘dogs, park, ...’);
- for a query “zoomed-in view of Pleiades”, a search result can include metadata such as (image id: 0132, relevant region: (t: 100 r: 200 b: 300 l: 100), metadata: ‘stars of Pleiades cluster’); etc.

User interface to surface rich-media snippets

Rich media snippets can be exposed within search results as follows:

- On large screen devices, e.g., personal computer, laptop, or tablet screens, text snippets are split into a multi-column representation of variable length; media snippets of high relevance are hovered over and played.
- On small screens, e.g., mobile devices, a multi-column representation is indicated by presenting left-swipe or right-swipe options that enable the exposure of rich media content.

In this manner, the techniques described herein provide search-result pages that have multiple types of relevant information, e.g., text, images, video, audio, etc. The search-result pages thus surface rich media information easily and in a transparent manner at the topmost level.

Aside from surfacing rich media snippets, the techniques also improve the quality of search by providing information extracted from rich media by trained machine learning models to the search engine. The quality of regular search improves from such access to rich media content and from granular media content retrieval. To further improve rich media search results, when permitted by the search user, user interaction with the returned content can be used as a signal to refine the search snippet matching techniques described herein.

The techniques apply to search on various platforms, e.g., search on mobile devices, photo-media search, e-commerce/product search, general search queries, etc. For example, per the techniques, a relatively specific product query “how does a robot vacuum cleaner perform on thick carpets?” can result in a snippet video of a robot vacuum cleaner in operation on a thick carpet.

CONCLUSION

This disclosure describes techniques that surface rich media snippets within search results. The search snippets provided using the described techniques include relevant segments of rich media, e.g., audio or video, summarized to sections that are pertinent to the search query.

REFERENCES

[1] Chechik, Gal, and Samy Bengio. “Relevance-Based Image Selection.” U.S. Patent Application 14/687,116, filed April 15, 2015.