October 10, 2016

VIDEO THUMBNAIL SELECTION BASED ON DEEP LEARNING

Weilong Yang

Min-hsuan Tsai

Tomas Izo

VIDEO THUMBNAIL SELECTION BASED ON DEEP LEARNING

ABSTRACT

Video thumbnails are often the first thing a viewer sees when browsing or searching for videos. A frame that is visually representative of the video is typically selected and used as a thumbnail representation of the video. Sometimes, such a thumbnail is not an adequate semantic representation of the video. Further, it is possible that such a thumbnail is not visually pleasing. This disclosure describes deep learning techniques to select video thumbnails that are visually attractive and reflect the content of a video. Thumbnails as described in this disclosure are attractive, improve a likelihood of user selection, and help users find relevant content easily.

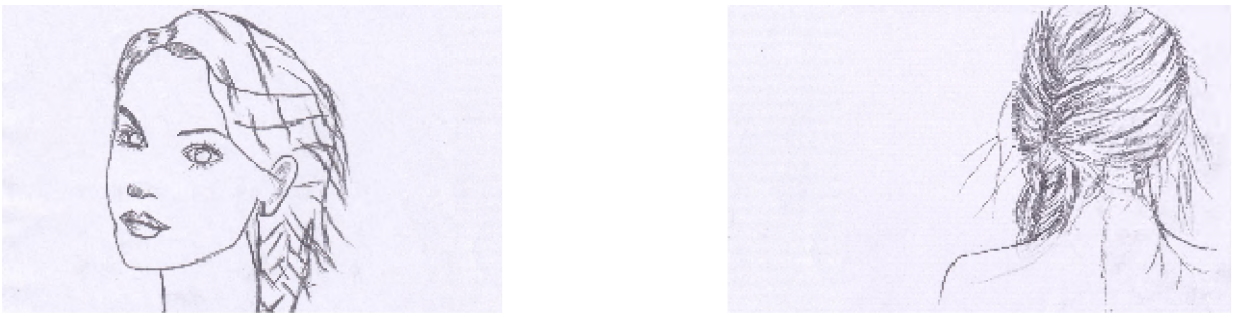
KEYWORDS

- Video thumbnails
- Deep Learning
- Semantic representation
- Photo aesthetics

BACKGROUND

When a user conducts a search, e.g., a web search, videos among the results are often represented by thumbnails in the search results. Websites that host video content include thumbnails for videos. A thumbnail that represents a video is typically selected from a large group of visually similar frames. In some scenarios, selection criteria for thumbnail images can result in thumbnails that are visually unappealing or that are not sufficiently representative of the content of the video. For example, Fig. 1 below shows two possible thumbnails for a video

sequence that includes a woman. Fig. 1(a) is a frontal still and Fig. 1(b) is a still shot from behind.



(a) (b)

Fig. 1: Thumbnail options

The still of Fig. 1(a), showing the face includes more detail than that of Fig. 1(b). The still of Fig. 1(a) is more likely to attract user interest, e.g., selection of the video via a click.

(a) (b)

Fig. 2: A video sequence that starts at a landscape scene (a), and gradually zooms in to plot characters (b)

Fig. 2 shows another example of thumbnails for a video sequence. The video starts at a landscape scene, as shown in Fig. 2(a), and zooms in to characters, as shown in Fig. 2(b). Due to slow zoom, a majority of frames in the video sequence resemble Fig. 2(a). Conventional techniques would therefore assign the frame in Fig. 2(a) as the representative thumbnail for the video. However, from a semantic standpoint, Fig. 2(b) which shows the characters of the plot, is likely more meaningful as a thumbnail.

DESCRIPTION

This disclosure describes techniques to enable selection of thumbnails that are semantically representative of a video and are visually pleasing. In order to derive semantically representative thumbnails, the video is annotated at several levels, e.g., frame, segment, entire video, etc. Annotation is performed using a machine learning technique, e.g., deep learning. Annotations at various frames are compared with video metadata, e.g., title, description, etc. A frame that shows high similarity between its annotations and the video's metadata is selected as the thumbnail.

In order to select a visually pleasing thumbnail, a machine learning model, e.g., deep learning, is trained to distinguish between visually pleasing frames and other frames. The training set of visually pleasing frames, known as positive data, is derived out of frames of videos and photos selected by human users. The training set of other frames, known as negative data, are frames and photos that are not selected by human users. Once trained, the machine learning model is able to select visually pleasing frames of any video. Training set includes only those frames and photos that have been permitted for use to train a machine learning model.

Thumbnail selection techniques of the present disclosure perform analysis of video content specifically upon user permission, e.g., from an owner of the video. Further, if image recognition is performed (e.g., to recognize whether a human face is present in a particular frame), such analysis is performed without specific reference to user profile data, unless user consent for use of data is obtained. If the user does not consent to analysis of a video, thumbnails are selected without applying these techniques, e.g., by selecting a first frame, a last frame, a relatively static frame, or a random frame of the video.

Selection of a visually pleasing thumbnail

The selection of a visually pleasing thumbnail is performed by training a machine learning model, e.g., deep learning, to distinguish between visually pleasing frames and other frames. Once trained, the machine learner can select visually pleasing frames. The training of the machine learner takes place in two phases.

In a first phase, several deep learning models ("phase 1 models") are trained to distinguish between visually pleasing and other frames. For the purpose of training, positive data used to train each deep learning model includes visually pleasing images selected by humans, e.g., video frames, thumbnails, photo album covers, or photographs that are selected by human users (with permission for such use of data). The negative data presented during training are images that are not selected by humans as pleasing, e.g., photographs or random frames from a video not selected by humans. Each deep learning model assigns a quality score to each image. The scores from the deep learning models are fused into a single score, for example, using a heuristically tuned linear combination of scores.

In a second phase illustrated in Fig. 3 below, an emulated machine learner model (302), e.g., an emulated deep learning model, is trained using several images such that the score produced by the emulated model for an image (304) is close to the fused score (308) from the phase 1 models (306a-c) for the same image (304). A measure of closeness between fused score and the emulated model score is the L2 loss (310). The emulated model is more compact compared to the phase 1 models, such that it requires less computation to compute the score of an image.

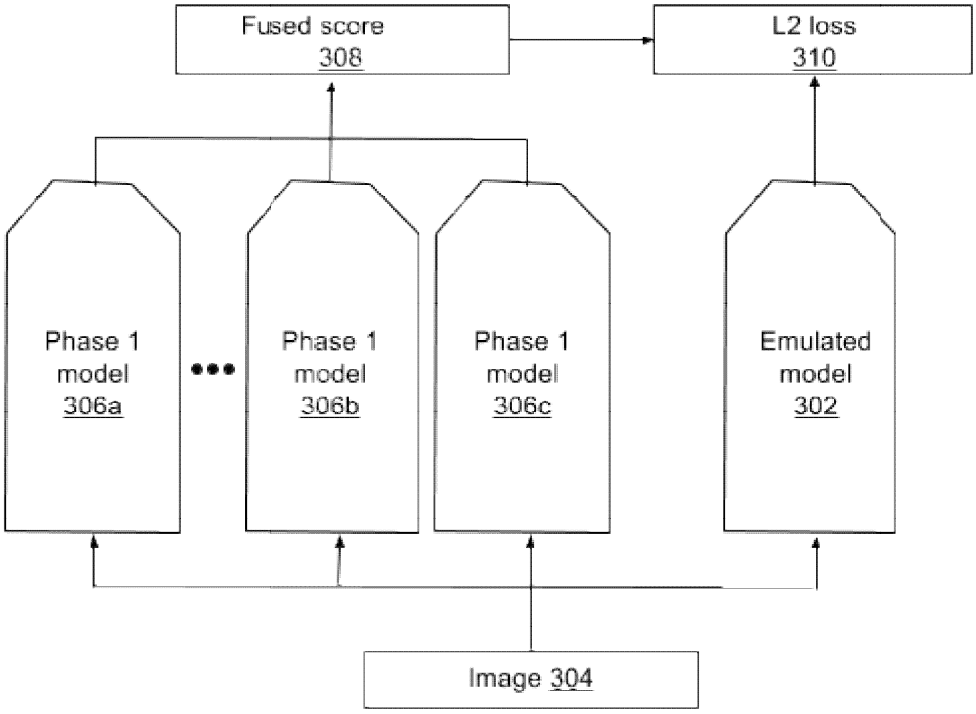


Fig. 3: Training an emulated model

Selecting a semantically representative thumbnail

Fig. 4 below illustrates an example by which a semantically representative thumbnail is selected. A frame sampler (404) samples a video (402) to obtain a number of frames (406). A frame annotator (408) annotates each frame with annotations (410a-d), e.g., of objects and concepts, along with a confidence score for each annotation. An aggregation model (412) aggregates the annotations in order to produce one or more semantically relevant labels (414) for the video. A representativeness scorer (416) compares the video label and video metadata (418) to respective annotations of each frame (410a-d) to determine a frame that is similar in annotation to video label and/or metadata. A frame that bears high similarity in annotation to the video label and/or metadata is selected as a semantically representative frame (420).

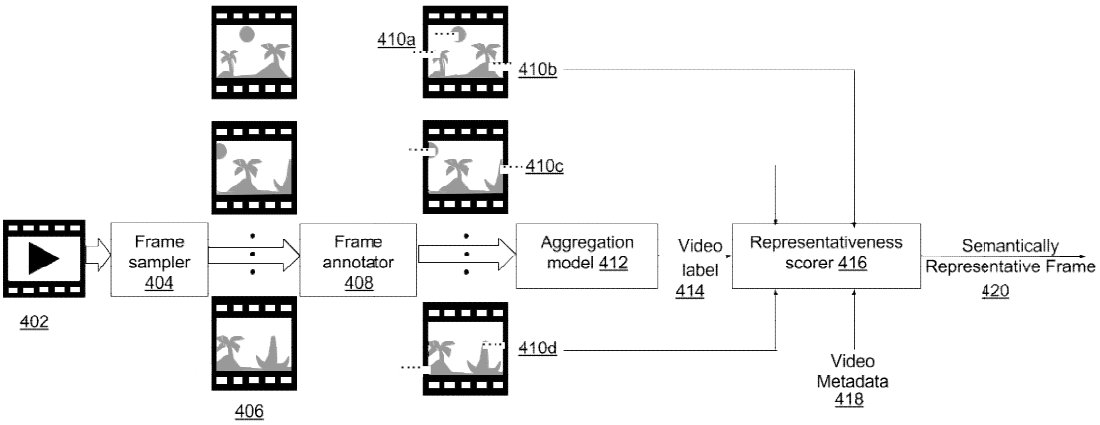


Fig. 4: Selecting a semantically representative frame as a thumbnail

Diversity in thumbnails

A video-hosting service may provide *n* thumbnails for the creator of the video to pick from. In this case, the video is partitioned into *n* non-overlapping segments. For each segment of video, thumbnail frames are selected based upon their visually pleasing and semantic-representativeness scores.

CONCLUSION

This disclosure describes machine learning techniques to select thumbnails from video. The techniques automatically select frames that are visually appealing and semantically representative of the content of the video as thumbnails. Such thumbnails attract the viewer and improve the likelihood of a viewer selecting the thumbnail, e.g., to play the video.