December 12, 2017

Captions Based On Speaker Identification

Andrew Gallagher

Terrance McCartney

Zhonghua Xi

Sourish Chaudhuri

CAPTIONS BASED ON SPEAKER IDENTIFICATION

ABSTRACT

Disclosed herein is a mechanism for generating and providing captions based on speaker identification. In some instances, the mechanism can be used to determine intervals where a single-speaker is speaking within particular image frames to assist the task of manual captioning or manual transcription. In some instances, the mechanism can be used to provide an awareness or indication of speaker turn-changes in captions, where a particular word or phrase can be grouped by particular speaker. In some instances, the mechanism can be used to provide an awareness or indication of speaker position and identity information corresponding to the speaker.

BACKGROUND

Video content providers (e.g., television channels, video streaming services, movie providers, etc.) often provide captions with the video content they provide. For example, captions often indicate words spoken by people or characters in the video content, or other audio content such as laughter, music, and applause. However, it can be difficult to automatically generate captions when multiple people or characters speak within a given segment. Additionally, in an instance where multiple people or characters speak within a scene, it can be difficult for a viewer to read the captions.

DESCRIPTION

A video content provider (e.g., a television channel, a video hosting or streaming service, and/or any other suitable video content provider) can use the mechanisms to analyze a video to provide speaker information useful for captioning the video. For example, the speaker

information can indicate that a first speaker and a second speaker are talking within a segment, which can allow a user providing manual captions for the video to more easily provide captions for the first speaker and the second speaker. As another example, when providing the captions to a viewer of the video, the speaker information can provide the captions in a manner that indicates the character associated with the caption. As a more particular example, captions associated with the first speaker can be placed in a position near the image of the first speaker, in a first font or color associated with the first speaker, and/or provided in any other suitable manner.

FIG. 1 illustrates an example method for generating speaker information. The method can be performed by a system that provides captions for video content, such as a server that hosts and streams video content to user devices.

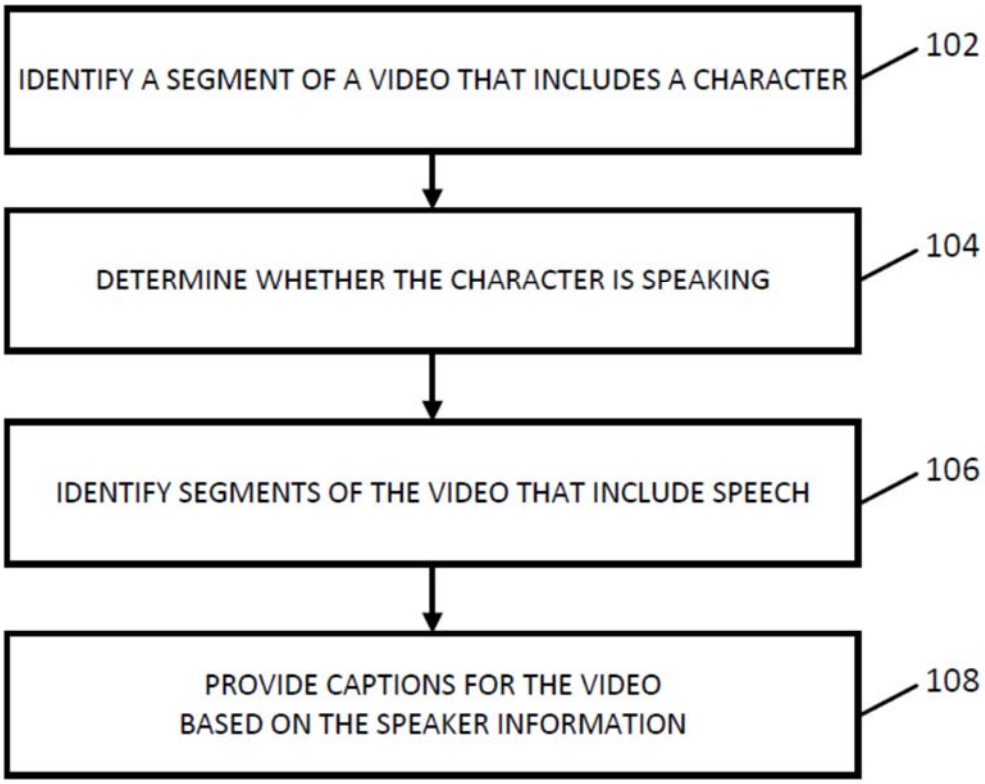


FIG. 1

At step 102, the server can identify a segment of a video that includes a character. The character can be, for example, a fictional character in the video (e.g., a character in a movie or television show, etc.), an animated character, a person appearing in the video and/or any other suitable type of character or person. The server can identify the segment using any suitable image recognition technique. For example, the server can identify one or more adjacent frames of the video that include the same face and/or body of a person. The segment of the video can be stored in any suitable format. For example, the segment can be stored as a group of frames (e.g., adjacent frames) of the video that include the same face and/or body of the person. In some instances, the group of frames can be a cropped version or portion of each frame that includes only the identified face and/or body of the person.

At step 104, the server can determine whether the character included in the identified segment is speaking using machine learning techniques. For example, the segment of the video can be used as an input to a deep neural network that provides as an output a classification of whether the character in the input segment is speaking during the segment. In some instances, the deep neural network can be trained on any suitable training set that incorporates any suitable features, such as movement of pixels in a mouth area of a face included in the segment across frames of the video provided as input, movement of any suitable parts of a body across frames of the video provided as input, and/or any other suitable features. In some instances, the deep neural network can classify the input into any suitable classifications, such as speaking, not speaking, laughing, yawning, eating, etc. Additionally, in some instances, the server can determine a probability value or confidence value associated with the classification that can indicate a likelihood that that the character is engaging in the selected classification during the segment of the video.

At step 106, the server can identify segments of the video that include speech. For example, the server can analyze audio content corresponding to the video to identify time periods or frames that include speech (e.g., between 1:02 and 1:04, between frames 30 and 50, and/or any other suitable segments). The server can identify the segments of the video that include speech using any suitable machine learning or audio fingerprinting techniques. For example, the server can analyze the audio and assign a probability that a segment includes speech using a machine learning algorithm.

As another example, the server can identify a speaker in the segment of the video. As a more particular example, the server can identify that Speaker 1 and Speaker 2 are both speaking during a segment that spans the time frame 1:02-1:04 of the video. In some instances, the server can identify the speaker using any suitable technique(s). For example, the server can compare audio content with the visual information of steps 102 and 104 to compare words spoken in the audio content with faces included in the visual information. As a more particular example, the server can determine whether it is likely that a particular word was spoken by a particular face based on a position or movement of a mouth. As another example, the server can determine if an audio fingerprint corresponding to the speaker's voice in the audio content matches visual biometric data corresponding to a particular face or body. As a more particular example, the server can determine if there is a match between the audio fingerprint and an image of the face based on a database or key that provides an example speech segment spoken by the character corresponding to the face, and the server can determine if an audio fingerprint corresponding to speech in the current segment of the video matches the example speech segment. As yet another example, the server can use any suitable natural language processing techniques to determine if a group of spoken words identified in the segment of the video are likely to come from one

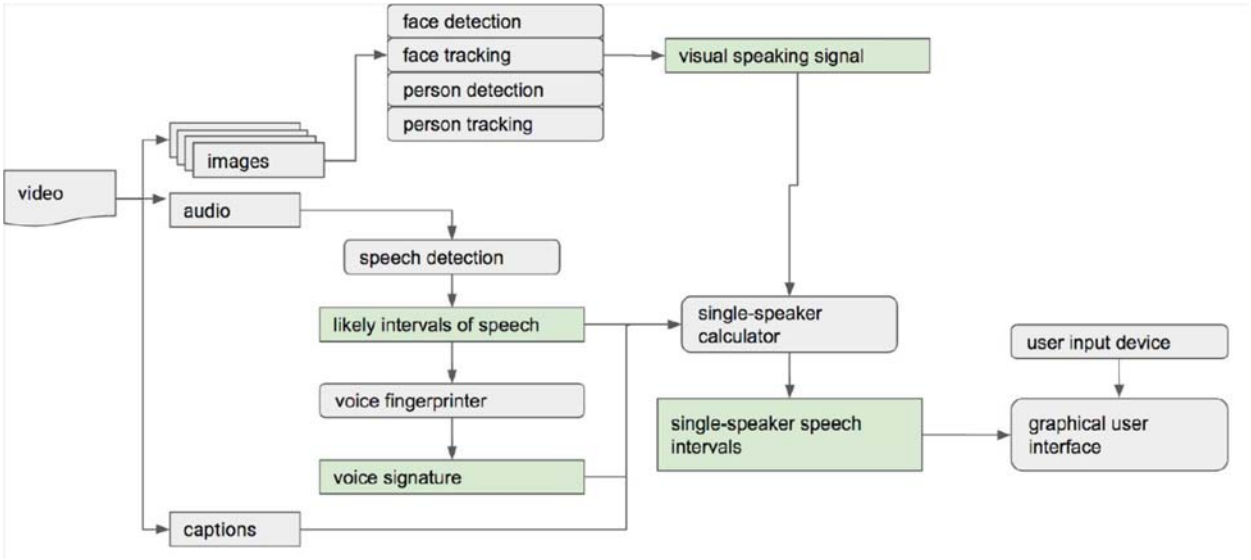
speaker or multiple speakers. As a more particular example, in instances where the segment contains the words “How are you I’m fine,” the server can determine that the words were likely spoken by a first speaker saying “How are you,” and a second speaker saying “I’m fine.”

Note that the server can perform steps 102 and 104 in parallel with step 106 of method 100. For example, the server can analyze video content to identify speaking characters while analyzing the audio content to identify intervals that contain speech. In some such instances, at step 108, the server can combine the results of steps 104 and 106 to provide captions for the video based on the results.

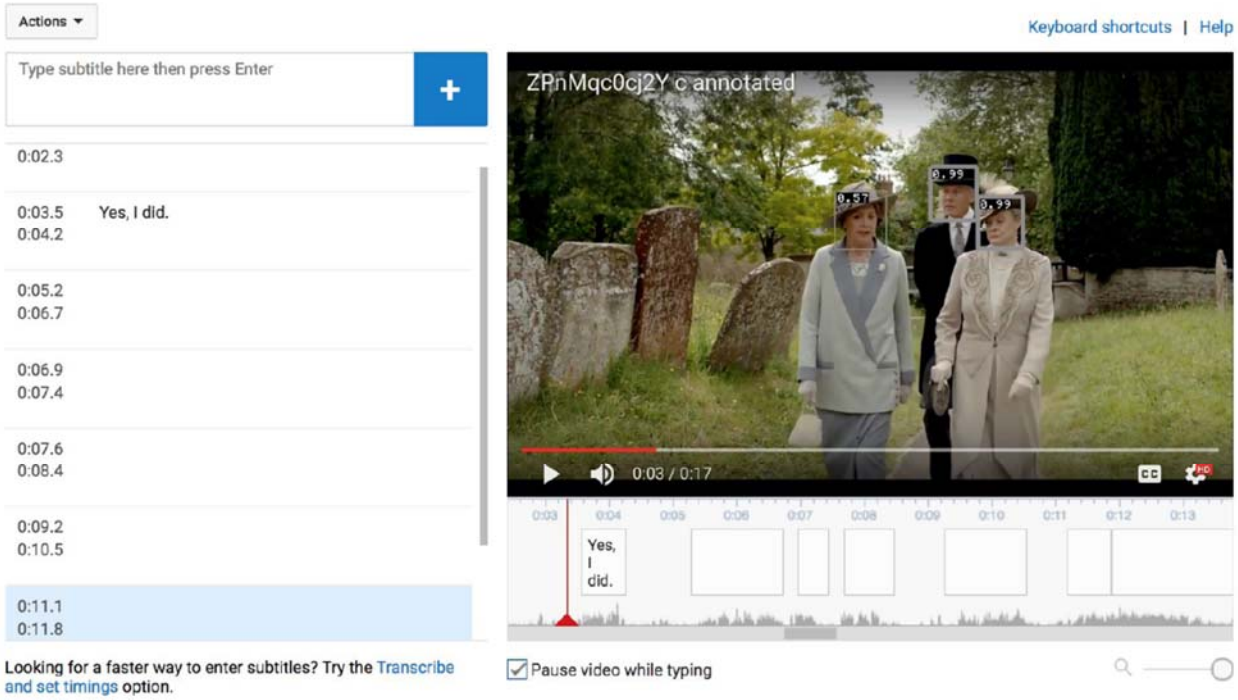
In some instances, at step 108, the server can use the speaker information to provide a user interface to a user entering manual captions for the video. For example, the user interface can show each identified segment of the video and a text box to enter captions for each identified speaker. As another example, the user interface can indicate each of the segments identified at 106 as containing speech. As yet another example, the user interface can allow the user entering manual captions for the video to edit the speaker information determined at steps 102-106, by resizing an identified speech segment to be longer or shorter, changing an identified speaker to a different speaker, and/or edited in any other suitable manner.

For example, as shown in the figure below, image information, audio information, and caption information (if present from, for example, speech recognition) can be extracted from a video. From the image information, face and/or person detection can be performed, where faces and/or people can be tracked across multiple image frames of the video (e.g., face detection and face tracking, person detection and person tracking, etc.). For each face or person that is detected, the pixels within the image frames can be analyzed to classify whether the person or face is speaking in each frame and a visual speaking signal is generated that indicates the

likelihood that the face or person is speaking at the corresponding time. In addition, the audio information can be analyzed to produce intervals that are likely to contain speech, where a single-speaker calculator inputs the visual speaking signal and the likely intervals of speech to produce single-speaker speech intervals. Accordingly, the single-speaker speech intervals can be identified in a graphical user interface that allows a user via a user input device to manually enter a transcript.



An illustrative example of a user interface including single-speaker speech intervals that are each available to receive a label or transcript from a user via a user input device is shown below.



In some instances, at step 108, the server can use the speaker information determined at steps 102-106 to change an appearance of captions (generated either manually or automatically) provided to a viewer of the video. For example, in some instances, sentences within a scene that are spoken by different speakers can be split into different lines when provided to the viewer. As a more particular example, in instances where a stream of speech to be captioned is “How are you? I’m fine,” the server can use the speaker information (e.g., visual information indicating whether a first character speaks in the scene from step 104, audio information indicating identities of one or more speakers during the scene from step 106, and/or any other suitable speaker information) to determine that “How are you?” was spoken by a first speaker and “I’m fine” was spoken by a second speaker. The server can then cause the provided caption to be split

such that “How are you?” is presented on a first line and “I’m fine” is presented on a second line to indicate a change in speakers.

An illustrative example of a user interface including single-speaker captions, where spoken words can be grouped into captions according to the speaker turn changes is shown below.

Additionally or alternatively, in some instances, the server can use the speaker information to change a position or appearance of the captions when providing the captions to a viewer of the video, as shown in Figure 2.

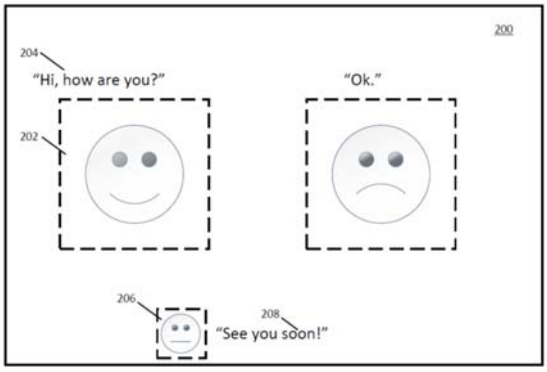
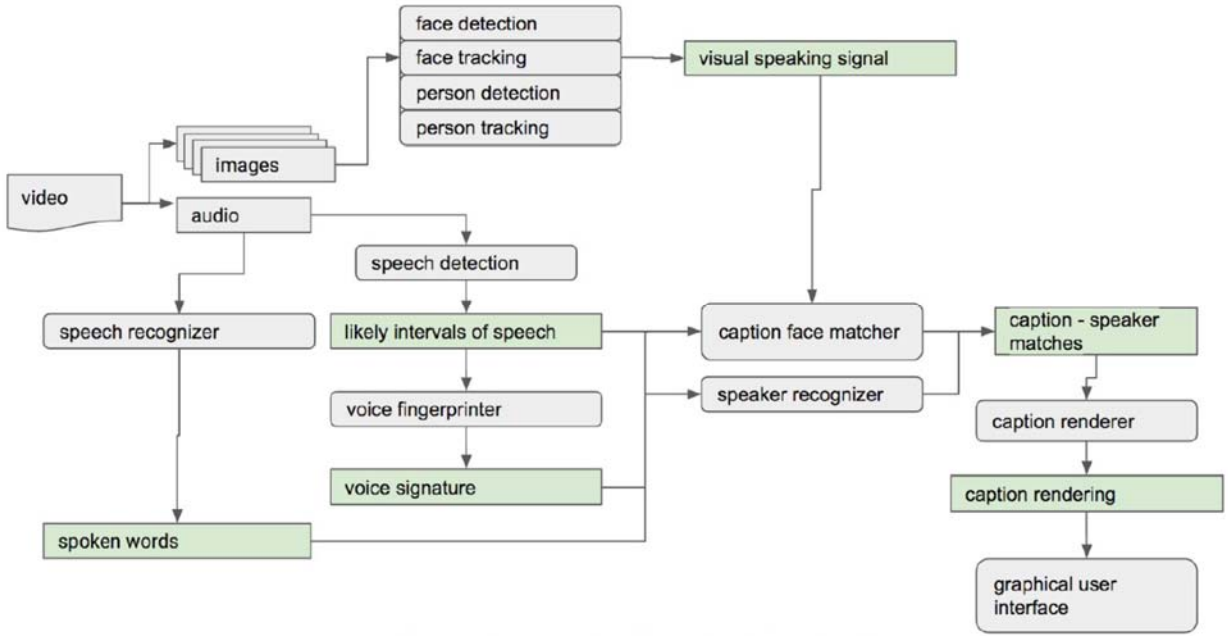


FIG. 2

For example, as shown in Figure 2, a caption 204 corresponding to speech spoken by a first character 202 can be positioned near first character 202. As another example, in some instances, captions corresponding to speech by different characters can indicate the corresponding character using font style. As a more particular example, captions can indicate speaker using font color by presenting each caption in a color corresponding to the speaker. Additionally or alternatively, as shown in Figure 2, captions corresponding to speech of characters not visually present in a particular scene can be indicated in any suitable manner. For example, as shown in Figure 2, a face icon 206 can be presented that indicates an identity of the character (e.g., by showing an image of the character's face) in connection with a caption 208. Additionally or alternatively, a font style associated with caption 208 can indicate that the character is not visually present in the scene, for example, by presenting caption 208 in an italic text, in a gray font color, and/or with any other suitable appearance.

This approach is also shown, for example, in the figure below.



An illustrative example of a user interface including a speaker-aware caption that infers which person is speaking in the image frame or frames and positions the caption proximal to the speaker in the image frame is shown below. Also shown in this illustrative example of speaker-aware captions, the user interface can include a speaker attribution marker (e.g., a line) that directs the user's attention by connecting a caption to a particular speaker.

In instances where the speaker is off-screen, an illustrative example of a user interface including a speaker-aware caption that includes an icon indicating the identity of the speaker is shown below.

Accordingly, a mechanism for generating and providing captions based on speaker identification is provided.