July 17, 2018

Audiovisual to Sign Language Translator

Manikandan Gopalakrishnan

Audiovisual to sign language translator

ABSTRACT

Communication between conventionally-abled and hearing-challenged individuals can pose difficulties. Although sign language partially alleviates problems, it requires an interlocutor to know sign language. Sign language varies with region of the world, and while a conventionally-abled interlocutor can learn a particular sign language, it is difficult or impractical to learn more than one sign language.

This disclosure describes an audiovisual to sign language translator. The translator captures, with user permission, speech and video of a conventionally-abled speaker and translates it in real time to a video of the speaker acting out a sign language in an emotion-preserving manner. The sign language video can be presented to the hearing-challenged person, e.g., in a picture-in-picture format, such that sign language symbolisms appear in one window, while raw video of the interlocutor appears in another. The techniques of this disclosure enable a conventionally-abled speaker to communicate with a hearing-challenged individual without needing to learn a sign language.

KEYWORDS

- sign language
- natural language processing
- sign language translation
- machine learning
- machine translation
- video generation

BACKGROUND

Communication between conventionally-abled and hearing-challenged individuals can pose difficulties. Although sign language partially alleviates many difficulties, it requires an interlocutor to know sign language. Sign language varies with region of the world, and while a conventionally-abled interlocutor can learn a particular sign language, it is difficult or impractical to learn more than one sign language.

DESCRIPTION

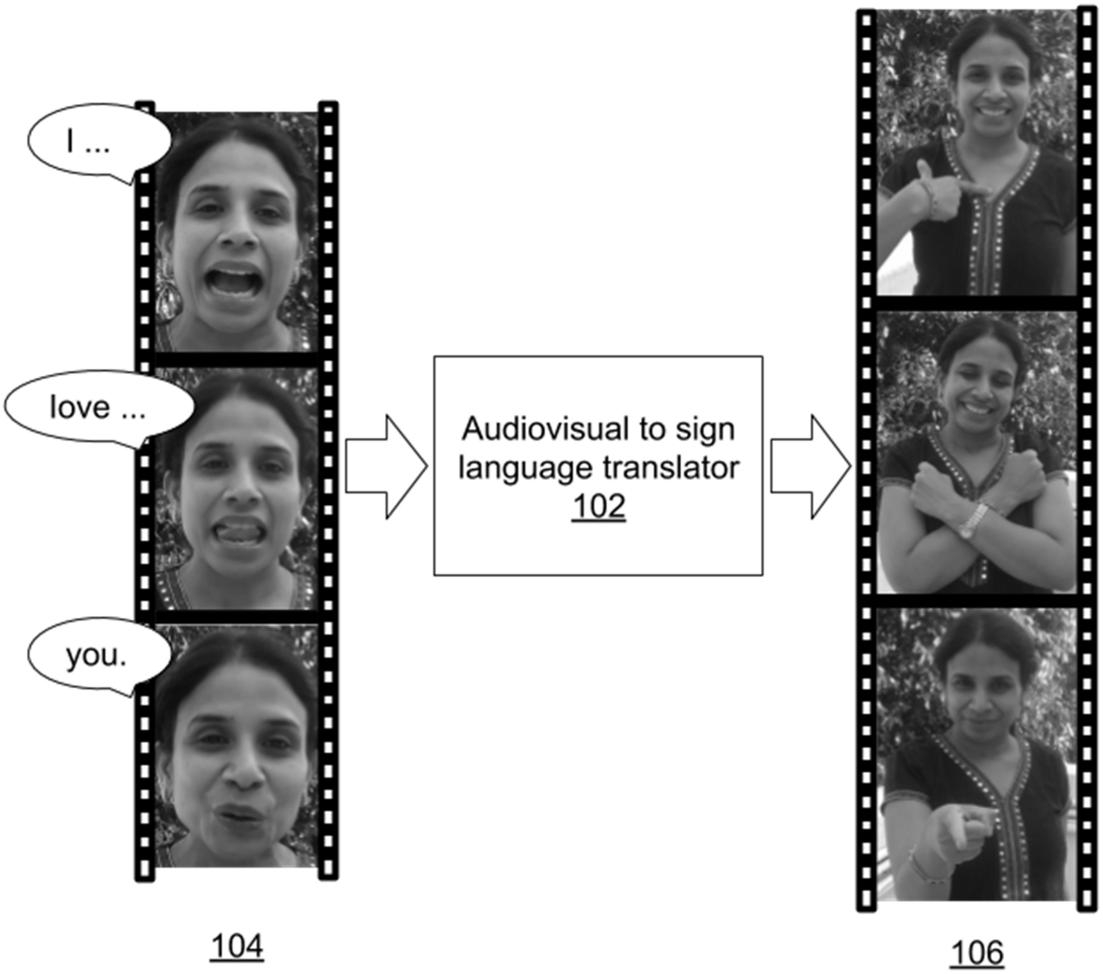

Fig. 1: Audiovisual to sign language translator

Fig. 1 illustrates an audiovisual to sign language translator, per techniques of this disclosure. With user permission, the audiovisual to sign language translator (102) accepts as input an audio and/or video feed comprising speech of a conventionally-abled individual (104). The translator produces as output a video comprising a sign language translation (106) of the input statement.

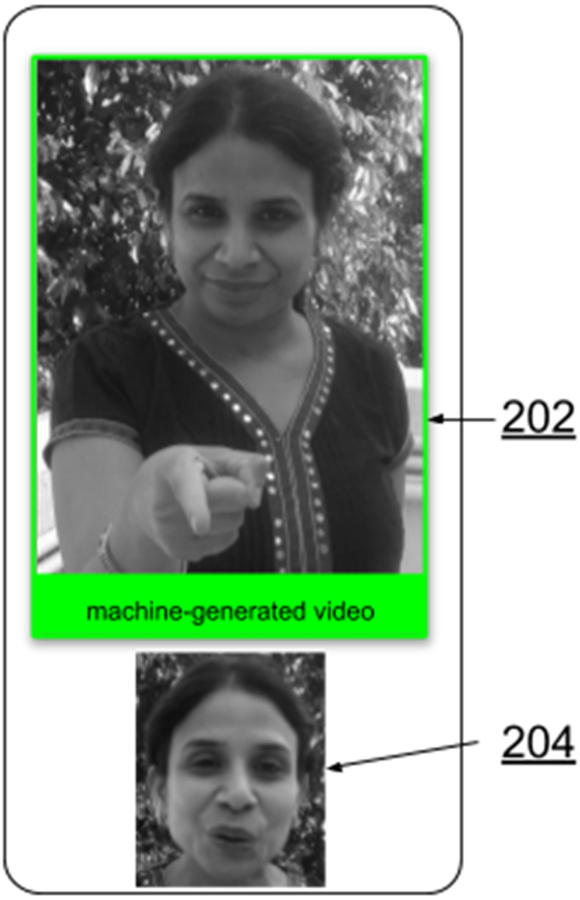

Fig. 2: Picture-in-picture

The video feed is presented to the hearing-challenged person, e.g., in a picture-in-picture format, as illustrated in Fig. 2. For example, the input audiovisual feed is presented in one window (204) while the translated (sign language) video feed is presented in another window (202) on a display screen. Such formats that display both video feeds at the output of the translator enables the hearing-challenged person to discern meaning from the sign language as well as from lip movements.

The generated video feed is clearly identified as being synthetically generated. For example, the generated video feed can be provided at a different frame rate than a regular video stream; displayed with a border or overlay the indicates that it is simulated; displayed with a different color and/or resolution; etc. Further, indications that the video feed is synthetically generated are presented in a prominent manner and may be included in the video feed itself, such that such indications cannot be turned off or removed from displayed or stored versions of the video feed.

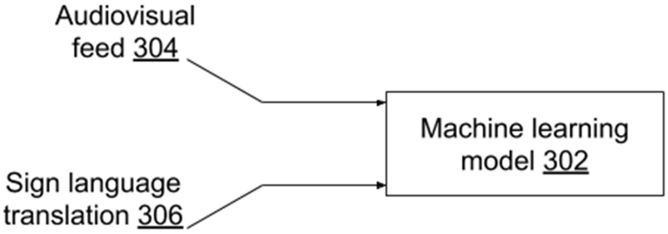

Fig. 3: Training the translator

Fig. 3 illustrates preprocessing or training of the audiovisual to sign language translator. A machine learning model (302) is provided an audiovisual feed (304) comprising a spoken language and a corresponding sign language translation (306). The machine learning model associates corresponding phrases in both input feeds, and upon sufficient training, can generalize. The machine learning model may be a multi-layer neural network, e.g., a long short-term memory (LSTM) neural network. Other types of models, e.g., recurrent neural networks, convolutional neural networks, support vector machines, random forests, boosted decision trees, etc., can also be used. The preprocessing stage results in a map between the sign language and words or phrases of the spoken language, and the construction of a sign language repository.

The preprocessing training can be done for any combination of spoken language (e.g., English, French, etc.) and sign language (e.g., American sign language, Chinese sign language, etc.)

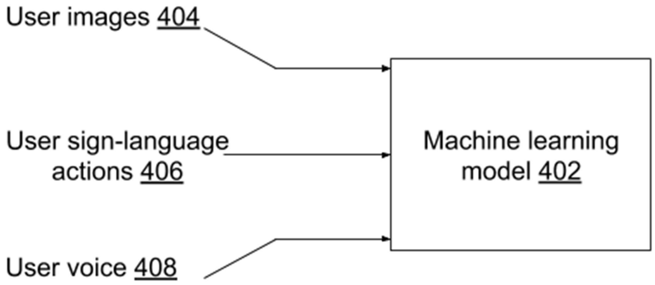

Fig. 4: Personalizing the translator

With user consent and permission, the translator can be personalized. Fig. 4 illustrates personalizing the machine learning model such that an audiovisual feed of a particular conventionally-abled individual is translated to a sign language as acted out by that individual. The machine learning model (402) accepts as input examples of user images (404) under various angles and emotions, user actions (406) corresponding to symbols of a sign language, and examples of user speech (408) expressed under various emotions, modulations and pitch. The user-personalization stage produces a repository of associations between images, sign language actions, and emotive speech of a given user. A machine-learning model can similarly build a repository for any user.

In operation, a conventionally-abled user communicates with the hearing-challenged individual via the translator. The video stream of the conventionally-abled user is contextually processed, using repositories created during preprocessing training and personalization, to create a simulated video of the conventionally-abled user acting out sign language corresponding to spoken speech, e.g., as illustrated in figures 1 and 2. Emotions of the conventionally-abled speaker can also be reflected in the simulated video. The simulated video is presented to the hearing-challenged individual, e.g., in picture-in-picture format, such that the hearing-challenged individual has the perception of directly communicating with the conventionally-abled person.

In this manner, the techniques of this disclosure enable a user to communicate with a hearing-challenged person in an emotion-preserving and real-time manner, without requiring the user to learn a sign language.

Alternatively, spoken language may be translated to sign language by first transcribing speech to text, then translating text to sign language. However, such text-based translation does not convey emotions adequately, e.g., clipped or loud voice due to anger, or wobbling speech due to sadness, is lost during the transcription to text. Additionally, a text-based sign language translator generally results only in pictographic (hieroglyphic) output, whereas true sign language is richer, and is capable of conveying nuanced meaning via facial expression and body movements.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes an audiovisual to sign language translator. The translator captures, with user permission, speech and video of a conventionally-abled speaker and translates it in real time to a video of the speaker acting out a sign language in an emotion-preserving manner. The sign language video can be presented to the hearing-challenged person, e.g., in a picture-in-picture format, such that sign language symbolisms appear in one window, while raw video of the interlocutor appears in another. The techniques of this disclosure enable a conventionally-abled speaker to communicate with a hearing-challenged individual without needing to learn a sign language.