Application of Audio and Video Processing Methods for Language Research

Przemyslaw Lenkiewicz[1], Peter Wittenburg[1], Oliver Schreer[2], Stefano Masneri[2], Daniel Schneider[3], Sebastian Tschöpel[3]

[1]Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[2]Fraunhofer-Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany
[3]Fraunhofer IAIS Institute, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Przemek.Lenkiewicz, Peter.Wittenburg}@mpi.nl
{Oliver.Schreer, Stefano.Masneri}@hhi.fraunhofer.de
{Daniel.Schneider, Sebastian.Tschoepel}@iais.fraunhofer.de

Abstract

Annotations of media recordings are the grounds for linguistic research. Since creating those annotations is a very laborious task, reaching 100 times longer than the length of the annotated media, innovative audio and video processing algorithms are needed, in order to improve the efficiency and quality of annotation process. The AVATecH project, started by the Max-Planck Institute for Psycholinguistics (MPI) and the Fraunhofer institutes HHI and IAIS, aims at significantly speeding up the process of creating annotations of audio-visual data for humanities research. In order for this to be achieved a range of state-of-the-art audio and video pattern recognition algorithms have been developed and integrated into widely used ELAN annotation tool. To address the problem of heterogeneous annotation tasks and recordings we provide modular components extended by adaptation and feedback mechanisms to achieve competitive annotation quality within significantly less annotation time.

Introduction

In psycholinguistics research very large audio and video corpora are available in order to investigate topics of interest. To evaluate the huge amount of media in the most efficient manner, meaningful annotations for the entire collection are required. One of the aims of the AVATecH project (Wittenburg et al. 2010) is to implement algorithms that allow for the automatic and semi-automatic creation of pre-annotations for the corpora, hence reducing the time needed to perform the manual annotation task. Analysis of the content created by Max Planck researchers is a difficult task due to two factors: 1) the size of the media corpora is very significant, reaching 70 TB presently; 2) the recordings are of very high diversity of languages, conditions and situations. This means that effective methods for automated processing of such content are not widely available or don't exist at all. The developed automated annotation algorithms will play a twofold role: 1) they would allow a dramatic decrease of time necessary to perform these tasks, which are normally very laborious; 2) automation of some parts of the process can significantly increase the uniformity of the annotations created worldwide by different researchers, which would contribute to consistency of the available language data. In this paper we describe the algorithms that operate on video recordings and present the initial results that we could obtain with them.

Audio and Video Analysis Algorithms

All algorithms have been created with the aim of performing well on recordings of any language and different acoustic and light conditions. They have also been designed to work without user interaction (except for the initial configuration by the means of few numerical parameters) to allow batch processing on multiple videos. The implementation was performed using a highly modular structure, so that future automatic annotators can be easily integrated in the current framework, using as input the results provided by the previous detectors.

Audio Segmentation

For linguistic annotation, segmentation on the utterance level is of high importance, but hard to achieve automatically without errors. This recognizer provides a fine-granular segmentation of the audio stream (Cheng 2010) into homogeneous segments, e.g. between speakers or at other significant acoustic changes. The user can control the granularity of segmentation by tuning a corresponding feedback parameter.

Speech detection

This recognizer is able to label audio segments containing human speech, regardless of the language of the recording. The user is allowed to manually provide a small amount of speech and non-speech samples in order to adapt the model to the given data, which leads to a more robust detection.

Speaker clustering

A language-independent speaker clustering recognizer is able to find segments spoken by the same person within a given recording. (Biatov and Kohler 2006; Biatov and Larson 2005; Reynolds 1995). The results can be used for removing the interviewer in a recording, or for extracting material from specific speakers from a recorded discussion. For optimization of the detection performance we use manual user input, e.g., the number of speakers or speaker audio samples.

Vowel and pitch contour detection

The pitch contour detector can allow researchers to graphically specify pitch contours and search for similar patterns. The detector can tag segments in audio recordings and annotate with pitch and intensity properties such as for example minimum, maximum, initial or final f0 frequency, or volume. The detector invokes PRAAT to

calculate f0 and volume curves of the input over time. Those are then used to find characteristic segments and annotate them.

Shot/cut detector and keyframes extractor

A shot is defined as the set of video frames that were continuously recorded with a single camera operation and represent therefore the basic unit of a video. This recognizer is able to detect such shots and label them. All further described algorithms provide result for given shot and therefore rely on the results of the shot/cut detection. Sub-shots are defined as a sequence of consecutive frames showing one event or part thereof taken by a single camera act in one setting with only a small change in visual content.

Global motion detection

Accurate motion analysis allows distinguishing between different types of video content and it can be used to segment a video in order to select only the parts, which are relevant for the researchers. For example, the presence of zooms and motion inside of a scene are usually the most interesting, while shots containing just panning and a low amount of internal motion are usually of little interest and can be usually discarded without further analysis. The algorithm developed performs a frame-based analysis and detects when global motion (pan, tilt, zoom in or zoom out) occur inside a shot. For each frame in the video a motion vector map is computed using the Hybrid Recursive Matching (HRM) algorithm (Atzpadin, Kauff, and Schreer 2004).

Skin color estimation

Due to the peculiarities of the dataset in the underlying application scenario, there isn't a unique set of skin color parameters which can achieve good results in the entire dataset and therefore typical approaches that make use of a training set to collect the parameters for skin detection on the entire dataset cannot be applied (Terrillon, Shirazi and Fukamachi 2000; Vezhnevets, Sazonov and Andreeva 2003). The algorithm created uses both the temporal information provided by the change between one frame and the next and the spatial information provided by the fact that skin color pixels tend to cluster in well defined regions. This skin color estimator does not need a training dataset but rather estimates the YUV ranges identifying skin color for each frame in each video. To successfully estimate the skin color the algorithm retrieves six parameters, the mean value and the size of the range for the luminance component Y and the color components U and V. These parameters define a subset (a parallelepiped) in the YUV color space, and all the pixels in the image inside this subset are marked as skin.

Hands and heads tracking

The algorithm works at first by segmenting the image in skin vs. non-skin pixels, using the information provided by the skin color estimator. The subsequent step in the detection process involves the search of seed points where the hands and heads regions most likely occur. Histograms along the horizontal and vertical directions compute the number of pixels with luminance and color values within the desired interval; the pixels where a maximum occur in both the directions are selected as seed points (Figure 1 left). A region-growing algorithm is then applied to the seed points in order to cluster together all the skin pixels in the neighborhood. Each region is approximated by an ellipse, characterized by the position of the center, its orientation and the length of its axes and for tracking purposes each of them is assigned a label (Figure 1 right). The tracking is performed by analyzing the change in position and orientation of the ellipses along the timeline, assigning labels based on position of the regions in the current and previous frames.

User interaction

The expected data is very heterogeneous and in some cases baseline recognizers can perform poor with no additional adaptation. Furthermore the researchers cannot accept annotation errors, e.g., a segment that is wrongly labeled as no-speech but has speech in it (false negative). Therefore the analysis components support adaptation and feedback-loop mechanisms. By adaptation mechanism we mean that the researcher is able to give examples of aspects he likes to detect, e.g., samples of a speaker for automatic speaker detection or sample segments without speech for the automatic detection of speech. By feedback-loop mechanism we mean strategies where the user runs a recognition process at first, then gives feedback about the quality of the result and then runs the process with the updated information again. For example, this could be applied for the speaker identification process: The user adapts the recognizer before running the component the first time by selecting some examples of the speaker, then runs the recognizer, and then verifies a number of segments and the recognizer would use this response to adapt the algorithm before running the process again.

Experiments

We have performed a range of tests to assess how our methods can increase the effectiveness of researchers' work. By the measure of effectiveness we consider the difference between the time necessary to create annotation for given media with and without our algorithms. This value is not easily calculated, as the time necessary for annotating a time unit of media depends on factors like: the purpose of the recording and contents of the media; what exactly from the contents needs to be analyzed and annotated; the person performing the annotation process and their expertise. Also the level of applicability of our methods can be different for different scenarios, therefore resulting in different amount of help they can offer.

Figure 1: Left – a video frame with skin color pixels marked and histograms of those in two dimensions, X and Y. Right – example of ellipses approximating the skin areas in given image.

In order to estimate the usefulness of our methods we have created a scenario in which a researcher had to perform a number of annotation tasks, to answer different linguistic research questions. The tasks have been chosen to represent a very common set of actions taken by a researcher annotating his recordings and included: 1) marking utterances of all speakers in the recording 2) marking the size of the gesture space of a recorded person; 3) marking where speech overlaps with gesturing; 4) marking specific gesture or behavior through the entire recording, like nodding, raising your arms from rest to the level of the body, etc.; 5) marking when gesturing action happens and segmenting them into stroke, hold and retreat. These tasks have been first performed manually by several researchers and the time necessary to carry them out was measured and averaged. In order to measure the possible decrease of the annotation time with the use of our algorithms, the following steps have been taken:

Annotating utterances of speakers in the recording has been performed automatically using the Speaker Clustering recognizer.

The size of the gesture space was measured, using the maximum and minimum x and y coordinates of the hands. Furthermore, recognizers offer the possibility to get more complex values than by user observation, like for example: the size of the entire gesturing space is 300 by 400 pixels, but 90% of all gesturing happens in a 150 by 150 pixels subspace.

Exemplary action of interests has been described and detected. Namely, the action of raising one's hands from the resting position has been described as the situation when the hands significantly increase their y coordinate in short time and when they reach values that are similar to the position of the head. This action is relatively simple to describe and detect, but manual annotation for such is nevertheless very time consuming. We are working currently on a more compound scheme for pattern description, which would allow detecting much more complex actions in the recordings.

The gesturing activity has been described and detected in the recording. The description was based on the three phases that normally take place in a gesture: the stroke, which was characterized by a fast acceleration of the hand; the hold, which was described as a period of time following the stroke, with very little hand movement; and finally the retreat, which was defined as a steady movement of the hand, after the hold phase. This description naturally doesn't cover all the situations of

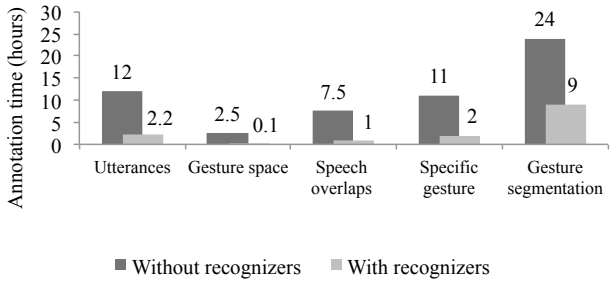

Figure 2: Comparison of annotation time for different tasks, with and without recognizers.

gesturing in human behavior and therefore it will not successfully detect all the gestures and all the phases in the recordings (manual corrections of the resulting annotation are necessary).

The above-described recognizers have been executed with an exemplary recording and the results of previously described algorithms. As the effect, appropriate annotations have been created. Afterwards they have been subjected to evaluation and manual corrections by a researcher.

Figure 2 presents the time necessary to carry out all the mentioned tasks. When the task has been performed with the help of recognizers, the necessary time stands for correcting the results obtained from them in order to make them useful for the researcher. It is possible to see that all the tasks took much shorter time to perform with the help of recognizers. Marking utterances required some corrections of the boundaries of the annotations, as well as splitting and merging some of the results of algorithms. Marking the size of the gesture space has been performed very effectively and almost no feedback from the user was necessary. Detecting specific action and detecting overlaps of speech and gesturing took more of user feedback, requiring respectively 1 and 2 hours of corrections. Detecting and segmenting the gestures, being the most complex tasks, required significant amount of corrections. However, all test cases have proven to save a lot of time of the researchers.

Conclusions

The specification and implementation of the recognizers described in this document has been performed in a very close contact with linguist researchers and according to the needs they have specified. After testing the relative effectiveness of our methods and witnessing the significant decrease of time necessary for annotations, we can say that our goals have been chosen correctly and our methods have proven very useful. As our next steps we are planning to further develop the recognizers and improve their efficiency and also work together with linguists to develop new recognizers that would create new types of annotations, for different research questions. Also we are planning to make the recognizers available for other annotation tools and also as web services. We believe this work would contribute significantly to the quality of linguistic data stored at the Language Archive of MPI and possibly in other locations.

Acknowledgments

AVATecH is a joint project of Max Planck and Fraunhofer, started in 2009 and funded by MPG and FhG. Some of the research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA).

References

Atzpadin, N., Kauff, P., Schreer, O., "Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing", Transactions on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications, Vol.14, No.3, 2004, pp. 321-334.

Biatov, K., and Kohler, J., 'Improvement speaker clustering using global similarity features', in Proceedings of the Ninth International Confer- ence on Spoken Language Processing, (2006).

Biatov, K., and Larson, M., 'Speaker clustering via bayesian informa- tion criterion using a global similarity constraint', in Proceedings of the Tenth International Conference SPEECH and COMPUTER, (2005).

Reynolds, D.A., 'Speaker verification using adapted gaussian mixture models', Speech Communication Journal, 17(1-2), (1995).

Shih-Sian Cheng, Hsin-Min Wang, and Hsin-Chia Fu, 'BIC-based speaker segmentation using divide-and-conquer strategies with appli- cation to speaker diarization', IEEE transactions on audio, speech, and language processing, 18(1), 141–157, (2010).

Terrillon, J.C., Shirazi, M.N., Fukamachi H., and Akamatsu, S., "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images", in Proceedings of the International Conference on face and gesture recognition, 2000, pp. 54-61.

Vezhnevets, V., Sazonov V., and Andreeva, A., "A Survey on Pixel-Based Skin Color Detection Techniques", in Proceedings of the GraphiCon 2003, 2003, pp. 85-92.

Wittenburg, P., Auer, E., Sloetjes, H., Schreer, O., Masneri, S., Schneider, D., Tschopel, S.: Automatic annotation of media field recordings: Presentation held at the 4th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH 2010, Lissabon, Portugal, 16. August 2010.