SAIVT Entity Extraction Webdemo Enhanced Final Year Project – Technical Paper

Francesco Ferraioli

Speech and Audio Laboratory, Queensland University of Technology, Brisbane, QLD, Australia

francesco.ferraioli@connect.qut.edu.au

Abstract

This paper will discuss the work done during my final year project at SAIVT regarding the enhancements to the SAIVT webdemo. The paper will begin by providing an overview of what the system does currently and all its various subsystems including speaker recognition and face detection. It will delve into the process by which all these subsystems work, what techniques do they use and what other techniques are present to perform the same task. In addition it will discuss what enhancements have been made to the webdemo as a result of the work undertaken as my final year project.

1. Introduction

The SAIVT webdemo is a website which showcases a lot of SAIVT's work. SAIVT has taken numerous news related video's and has run them through many different script that they have engineered to extract entities from videos. These script include a combination of various speaker extraction, face extraction and optical character recognition scripts as well as speaker and face clustering to link the video's together as best as possible. I will now discuss some of my findings on the various types of techniques that SAIVT uses to extract entities from the videos.

2. Entity Extraction Techniques

2.1 Speaker Recognition

Speaker recognition is defined as not only detecting when an individual is speaking but moreover the identification of the person, the individual, who is speaking. Over the years there has been much research in the area of the use of machines to automate speaker recognition [1]. There have been a vast quantity of algorithms and techniques implemented for automated speaker recognition, even over coming major difficulties like background noise (i.e. recognizing the speaker in a noisy environment), some more robust than others [1].

A method used for speaker recognition, and one that is highly researched and used at SAIVT is that of Joint Factor Analysis (JFA) Modeling. This technique involves combining the appropriate estimation of speaker variability and session variability subspaces. [2]. Each subspace are then "trained on a database containing a large number of speakers each with several independently recorded sessions." [2].

A similar method is called "i-vector speaker modeling" [3] and also this is heavily researched and used at SAIVT. I-vector speaker modeling differs from JFA modeling in that it "represent the GMM super-vector by a single total-variability space" [4]. This total variability space contains both the speaker and the channel variability but requires "additional intersession compensation approaches" [4]. There have been some existing approaches that have shown effective for this task including WCCN, LDA and NAP [4].

2.2. Face Recognition

Face recognition has been a hot topic, especially in the field of Human Computer interactions [5]. Face recognition was defined as "visual pattern recognition problem where the face, represented as a three-dimensional object that is subject to varying illumination, pose, expression and other factors, needs to be identified based on acquired images" [6].

The Viola-Jones object detection framework has drastically increased the number of facial processing applications [5]. All these applications have provided very reliable approaches to the problem of face recognition.

The basic idea behind the Viola-Jones framework, more specifically the Haar-Cascade technique, is based on the idea of "boosted cascade of weak classifiers" [5]. It is the process of sequentially testing the image against numerous classifiers and rejecting those that fail. These classifiers include multiple "Haar-like features" which is an encapsulation of rectangular patterns [7]. There are three types of patterns, edge features, line features and centre-surround features [7]. Figure 1 illustrates a few of the patterns used in the Haar-Cascade face detection algorithm.

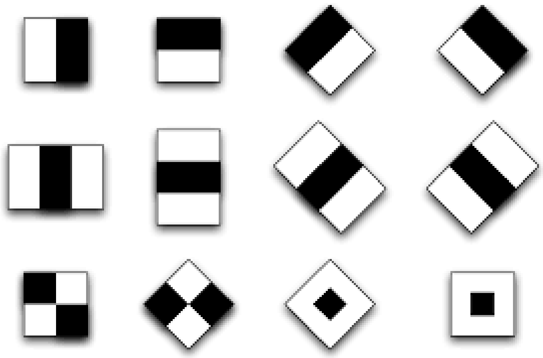


Figure 1: Haar-Cascade classifiers

2.3. Speech Clustering

Speech Clustering is the process of linking the entities found in an audio file based on identity [8]. This is achieved "using speaker diarization and speaker linking" [8].

The most extensively utilized method for speaker clustering in speaker diarization is agglomerative clustering with retraining (ACR) [3]. Even though ACR is the most widely used technique, it was found to be very inefficient. For this reason, when conducting speaker linking in large dataset, another method is preferably used [3].

This more efficient method is called complete-linkage clustering (CLC). This is achieved through merging based on a highest similarity, or lowest distance, score. [8]. Moreover, another technique for speech clustering is named cross likelihood ratio (CLR) [3]. This method has shown to be robust and effective especially when "incorporated into the JFA modeling framework" [3].

2.4. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is the process of converting typewritten or printed images to computer readable text [9]. This technique is mostly used in data entry, where it is necessary to import vast information stored in scanned images that needs to be computer readable. However, this process can be also very useful in entity extraction. [9].

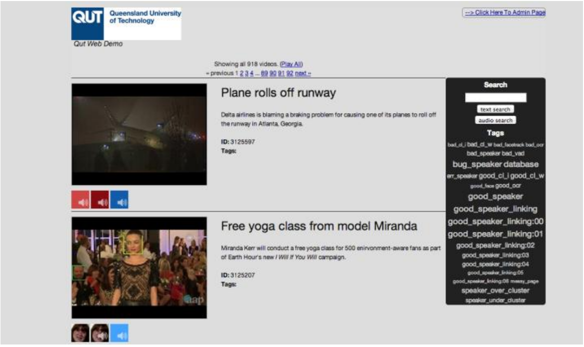
The most widely used engine for OCR is Tesseract, not only because it is open source, but also because it is very efficient [10]. Tesseract performs OCR by running various processes including: Line and Word Finding and Word Recognition [10]. Line and Word Finding are algorithms which essentially tell the computer where the lines of words are in the image. The Word Recognition process ultimately tries to understand exactly what computer readable word is a particular imaged word. [10].

3. The WebDemo and its Enhancements

As previously mentioned, the webdemo showcases a lot of the work that is done by SAIVT. This includes speaker recognition, face detection, OCR and much more. The webdemo's structure is one similar to that of YouTube in that it contains an index page which lists out all the videos contained in the database and it also has the video play page which allows a particular video to be played along with all its content. In the case of the webdemo, the content of the videos not only includes the title, summary, uploaded date, tags, but also includes all the entities that have been extracted from the videos. Entities being speakers and faces that have been detected through the various processes implemented by SAIVT.

3.1. Index Page

Figure 2 illustrates the old index page of the SAIVT webdemo. As shown the page is not extremely modern in terms of the look and feel of the page, the buttons are the default HTML buttons, the pagination links are not centered and have no style applied to them. Furthermore, the HTML of this page was not structured correctly and needed a lot of work.



QUT Queensland University of Technology
--> Click Here To Admin Page
Qut Web Demo
Search
text search
audio search
Tags
bad_cl_i bad_cl_w bad_facetrack bad_ocr bad_speaker bad_vad bug_speaker database err_speaker good_cl_i good_cl_w good_face good_ocr good_speaker good_speaker_linking good_speaker_linking:00 good_speaker_linking:01 good_speaker_linking:02 good_speaker_linking:03 good_speaker_linking:04 good_speaker_linking:05 good_speaker_linking:06 messy_page speaker_over_cluster speaker_under_cluster

Showing all 918 videos. (Play All)

« previous 1 2 3 4 ... 89 90 91 92 next »

Plane rolls off runway

Delta airlines is blaming a braking problem for causing one of its planes to roll off the runway in Atlanta, Georgia.

ID: 3125597
Tags:

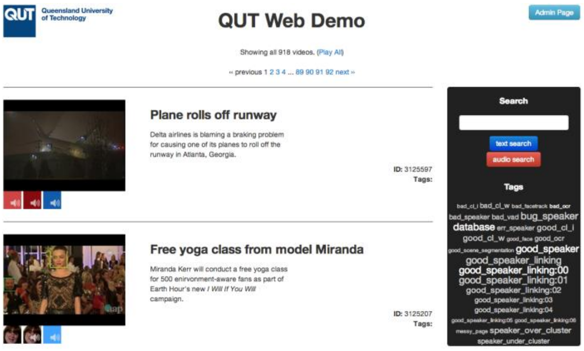
Free yoga class from model Miranda

Miranda Kerr will conduct a free yoga class for 500 enirvonment-aware fans as part of Earth Hour's new *I Will If You Will* campaign.

ID: 3125207
Tags:

Figure 2: Old Index Page

Figure 3 below illustrated the new and current index page. This page has the same content as Figure 2 but the page looks more alive and more modern. The buttons are no longer default, the pagination links are centred and look more modern, but most importantly, the HTML is structured much more professionally now.



QUT Queensland University of Technology
QUT Web Demo
Admin Page
Search
text search
audio search
Tags
bad_cl_i bad_cl_w bad_facetrack bad_ocr bad_speaker bad_vad bug_speaker database err_speaker good_cl_i good_cl_w good_face good_ocr good_scene_segmentation good_speaker good_speaker_linking good_speaker_linking:00 good_speaker_linking:01 good_speaker_linking:02 good_speaker_linking:03 good_speaker_linking:04 good_speaker_linking:05 good_speaker_linking:06 messy_page speaker_over_cluster speaker_under_cluster

Showing all 918 videos. (Play All)

« previous 1 2 3 4 ... 89 90 91 92 next »

Plane rolls off runway

Delta airlines is blaming a braking problem for causing one of its planes to roll off the runway in Atlanta, Georgia.

ID: 3125597
Tags:

Free yoga class from model Miranda

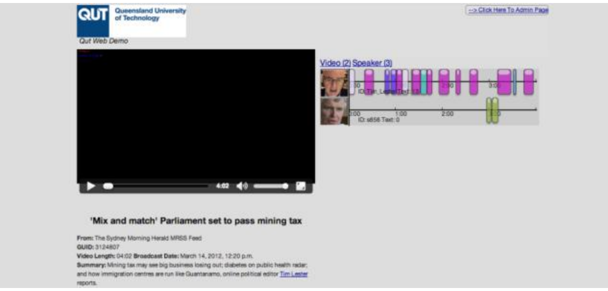
Miranda Kerr will conduct a free yoga class for 500 enirvonment-aware fans as part of Earth Hour's new *I Will If You Will* campaign.

ID: 3125207
Tags:

Figure 3: New Index Page

3.2. Video Play Page

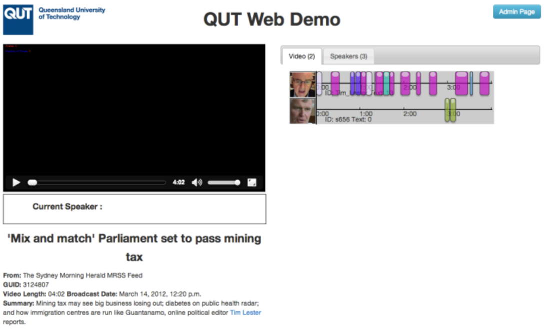
Figure 4 depicts the old video play page whilst Figure 5 illustrates the new video play page. It is clear that Figure 5 shows a much more modern look and feel than Figure 4. One of the many enhancements done on this page is the functionality to switch between viewing the faces and viewing the speakers on the right hand side. This functionality is now done with tabs instead of having links above that contain no style.



'Mix and match' Parliament set to pass mining tax

From: The Sydney Morning Herald MRSS Feed
GUID: 3124807
Video Length: 04:02 **Broadcast Date:** March 14, 2012, 12:20 p.m.
Summary: Mining tax may see big business losing out; diabetes on public health radar; and how immigration centres are run like Guantanamo, online political editor Tim Lester reports.

Figure 4: Old Video Play Page



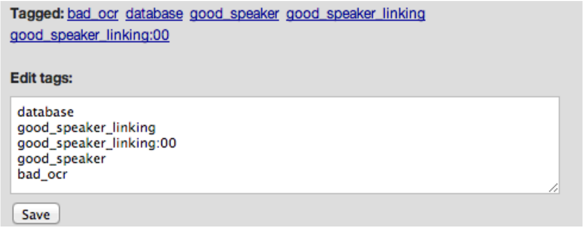
'Mix and match' Parliament set to pass mining tax

From: The Sydney Morning Herald MRSS Feed
GUID: 3124807
Video Length: 04:02 **Broadcast Date:** March 14, 2012, 12:20 p.m.
Summary: Mining tax may see big business losing out; diabetes on public health radar; and how immigration centres are run like Guantanamo, online political editor Tim Lester reports.

Figure 5: Old Video Play Page

3.3. Tags

Adding tags to video is a very common practice and is present in most websites which contain a database of videos. Previously, editing the tags was done through the use of a textarea in which, following a particular syntax of having each tag on a different line, would edit the tags of the video, as shown in Figure 6.



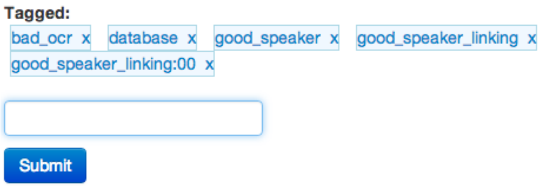
Tagged: bad_ocr database good_speaker good_speaker_linking good_speaker_linking:00

Edit tags:

```
database
good_speaker_linking
good_speaker_linking:00
good_speaker
bad_ocr
```

Figure 6: Old Tagging System

This is very unprofessional and not very standard. This system has been replaced by what is shown in Figure 7. The tags are listed as shown and next to each is a cross which when clicked will remove it from the list of tags. Furthemore, a small form is present with a textbox and a submit button that allows for new tags to be added to the list. The form contains validation for spaces and duplicates. All these calls to the database are done asynchronously through AJAX calls.



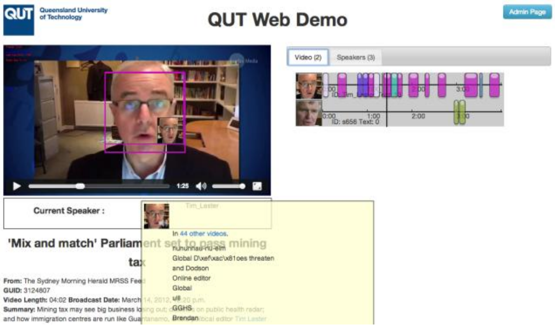
Tagged:
bad_ocr x database x good_speaker x good_speaker_linking x good_speaker_linking:00 x

Figure 7: New Tagging System

3.4. Entity Display Enhancements

The sole purpose of the webdemo is to showcase the work done in SAIVT through the extracting entities from videos. The main entities include faces and speakers. A major part of my work for the final project was to enhance the displays of these entities on the video play page. Currently, the entities are only being listed out on the right hand side and the video has been encoded with a square to encompass the faces found.

First of all, a feature was needed to display the current speaker underneath the video as illustrated in Figure 8. This way the user would not need to constantly check the speakers list on the right hand side to find out the current speaker.



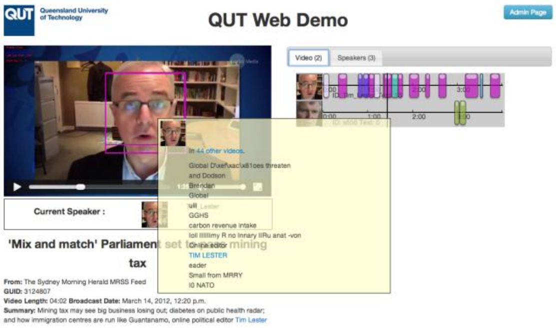
'Mix and match' Parliament set to pass mining tax

In 44 other videos.
nununnau-nu-elm
Global D\xef\xac\x81oes threaten
and Dodson
Online editor
Global
ull
GGHS
Brendan

From: The Sydney Morning Herald MRSS Feed
GUID: 3124807
Video Length: 04:02 **Broadcast Date:** March 14, 2012, 12:20 p.m.
Summary: Mining tax may see big business losing out; diabetes on public health radar; and how immigration centres are run like Guantanamo, online political editor Tim Lester

Figure 8: Current Speaker

Furthermore, another feature needed to be implemented show the faces on the video as interactive HTML elements overlayed on the videos. Once this was implemented, the users could have interactive HTML elements to display information on the current faces as depicted in Figure 9. Moreover, the encoded video could be removed be replaced by the original video as the squares will be overlayed on top of the video.



In 44 other videos.
Global D\xef\xac\x81oes threaten
and Dodson
Brendan
Global
ull
GGHS
carbon revenue intake
loll lllllimy R no lnnary llRu anat -von
Online editor
TIM LESTER
eader
Small from MRRY
I0 NATO

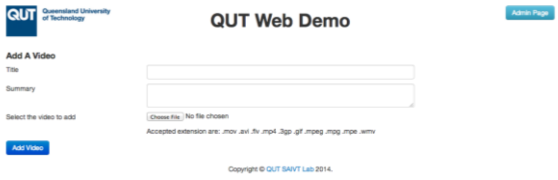
'Mix and match' Parliament set to pass mining tax

From: The Sydney Morning Herald MRSS Feed
GUID: 3124807
Video Length: 04:02 **Broadcast Date:** March 14, 2012, 12:20 p.m.
Summary: Mining tax may see big business losing out; diabetes on public health radar; and how immigration centres are run like Guantanamo, online political editor Tim Lester

Figure 9: Current Faces

3.5. Upload Video

The last major task of the project was to implement a feature to allow the user to add their own video to the webdemo that would then be processed to extract the entities and be displayed as a video on the webdemo. A form was necessary to allow the user to upload their video to the webdemo as this form is displayed in Figure 10.



Accepted extension are: .mov .avi .flv .mp4 .3gp .gif .mpeg .mpg .mpe .wmv

Copyright © QUT SAIVT Lab 2014.

Figure 10: Upload Video Form

Once uploaded, the video will be available for viewing but would show that it has yet to be processed. A cron job would then send the video to a High Performance Computer (HPC) to process. The processing needs to be initiated manually on the HPC. Upon completion of the processing, the video along with all its extracted entities would be pulled back to the webdemo server and the database would be rebuilt with the new information about all the entities found for the video uploaded. This is all done via the one cron job that runs every 30 minutes.

4. References

[1] Campbell Jr, J. P. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), 1437-1462.

[2] Vogt, R. J., Baker, B. J., & Sridharan, S. (2008). Factor analysis subspace estimation for speaker verification with short utterances.

[3] Ghaemmaghami, H. (2013). Robust automatic speaker linking and attribution.

[4] Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S., & Mason, M. W. (2011, August). I-vector based speaker recognition on short utterances. In Proceedings of the 12th Annual Conference of the International Speech Communication Association (pp. 2341-2344). International Speech Communication Association (ISCA).

[5] Castrillón-Santana, M., Déniz-Suárez, O., Antón-Canalís, L., & Lorenzo-Navarro, J. (2008). Face and facial feature detection evaluation performance evaluation of public domain haar detectors for face and facial feature detection.

[6] Jain A., Li, S. (2005). Handbook of Face Recognition.

[7] Kasinski, A., & Schmidt, A. (2010). The architecture and performance of the face and eyes detection system based on the Haar cascade classifiers. Pattern Analysis and Applications, 13(2), 197-211.

[8] Ghaemmaghami, H., Dean, D., & Sridharan, S. (2013, August). Speaker attribution of Australian broadcast news data. In Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM): CEUR Workshop Proceedings, Volume 1012 (pp. 72-77). Sun SITE Central Europe.

[9] Mori, S., Nishida, H., & Yamada, H. (1999). Optical character recognition. John Wiley & Sons, Inc..

[10] Smith, R. (2007, September). An Overview of the Tesseract OCR Engine. In ICDAR (Vol. 7, pp. 629-633).