

Multimodal Data Collection in the AMASS++project

Scott Martens¹, Jan Hendrik Becker², Tinne Tuytelaars², Marie-Francine Moens³

¹ Centrum voor Computerlinguïstiek

² IBBT-PSI - Center for Processing Speech and Images

³ Department of Computer Science

K.U.Leuven

Leuven, Belgium

E-mail: Scott.Martens@ccl.kuleuven.be, JanHendrik.Becker@esat.kuleuven.be,

Tinne.Tuytelaars@esat.kuleuven.be, Marie-Francine.Moens@cs.kuleuven.be

Abstract

The AMASS++ project is a project sponsored by Flemish public interests aimed at increasing the usefulness and usability of multimedia archives - notably combined text, audio and visual data - through the application of natural language and image processing technologies. To this end, we are collecting digital print media news reports, and television news programming. This data will be thematically organized and includes annotated television news programming and text media in both English and Dutch. The project's purpose is to develop and implement technologies to perform cross-language and cross-media search, summarization and user-friendly, productive presentation of results.

1. Introduction

Digital multimedia archives are now the first place many people turn to for information about current and relatively recent events, not just for ordinary media consumers but also professionals in journalism, business, academia and government. Improving the accessibility and usefulness of this class of resource is the objective of a number of public and private initiatives.

The AMASS++ project (Advanced Multimedia Alignment and Structural Summarization) is a project sponsored by Flemish public interests¹ aimed at increasing the usefulness and usability of multimedia archives – notably combined text, audio and visual data – through the application of natural language and image processing technologies. AMASS++ touches not only on the problem of finding materials relevant to queries, but also on the importance of presenting them in the most productive manner, rather than simply as a Google-style list of ranked pointers.²

The goals of this project are the development of:

- Technologies to align comparable content across media, e.g. text and video news reporting of the same events.
- Technologies for providing a structured, cross-media and cross-language summary of information about topics encompassing results from different sources, e.g. both text summaries and images.

¹ AMASS++ is funded by IWT (Institute for Innovation in Science and Technology) project No. 060051, and funding for some of the video research is provided by a fellowship from the FWO (Fund for Scientific Research Flanders).

² See <http://www.cs.kuleuven.be/~liir/projects/amass/> for further details about AMASS++.

Media firms of various kinds have shown their interest in the outcome of the project through their participation in the AMASS++ user committee, and some of them are providing us with text and video data. Additional resources are acquired through Internet crawling and recording broadcasted material.

Proof-of-concept and evaluation are performed using news materials – text news articles and televised news reports – because they are the kinds of materials whose producers can most immediately profit from the results of this project, because they are readily classifiable topically, and because they are widely available.

The scope of this project has been limited to natural language texts and image processing in order to simplify the problem and given the broad availability of subtitling and reliable transcripts. We are not considering the classification or alignment of audio or processing of the output of a speech recognition system.

2. Data contents and collection

The test data consists of text news reports (including accompanying still images in many cases), video capture data, subtitling acquired along with the video capture data, and video transcripts (where available). Because this project involves the grouping of comparable materials, data capture methods are in part oriented towards the collection of news media concerning specific events, such as the American presidential elections or the Olympic games, although for some sources, more longitudinal collection processes are also at work. Sources include news reports - in text format and televised video - in both English and Dutch originating from British, Dutch and Flemish sources.

In total, we are aiming for approximately 200 hours of video with subtitles, and an as yet indeterminate amount

of text and web derived multimedia data.

2.1 Text data

The text data collected in this project comes from a number of sources: newspaper reports from the Internet, transcripts of TV programming, captured subtitles and autocues.

We have crawled the Google News website for URLs to articles in Dutch that Google has classified in political categories including foreign politics. Those articles are then downloaded as raw HTML and accompanying images. The HTML is filtered to separate essential content from advertisements, navigation menus, and other peripheral materials. This is challenging because articles are retrieved from a variety of news providers, each of whom structures their website somewhat differently. We use tools developed in-house for retrieving and filtering web pages. The crawler follows a breadth-first rule. The HTML filter integrates heuristic rules that balance the generality and accuracy of the filtering procedure. To date we have collected roughly 1GB of processed text data from the web.

Preprocessing. The language of the text is first identified, using procedures that select not only which language the material is in, but also assess if it is a language other than those we are prepared to process. It is then tokenized and tagged using the TnT statistical tagger (Brants, 2001), trained for Dutch using the *Corpus Gesproken Nederlands* (Oostdijk et al., 2002) and for English using the British National Corpus (Aston & Burnard, 1998); and chunked using the ShaRPa2.1 chunker (Vandeghinste, 2008). Dutch texts are also processed using an in-house decompounder (Vandeghinste 2008). The Dutch POS tagger has been trained to use a subset of the CGN/D-COI tagset (Van Eynde, 2005) and the English tagger uses the C5 tagset deployed in the British National Corpus (Aston & Burnard, 1998).

2.2 Video data

Video data is captured using a Hauppauge WinTV PVR-350 card from analog broadcasts and stored as MPEG2 at the standard 768x576 PAL resolution, although letterboxing reduces the actually used area to 750x430. Subtitles are extracted by capturing text from the subtitle teletext page. This produces HTML output for the subtitles, which includes text color information that often designates changes of speaker. Timing information for subtitles is also preserved so that realignment with the video is possible (although some manual adjustment is currently still needed). Although we have chosen not to focus on audio processing within this project, the audio is stored in 48kHz stereo format, generally encoded at roughly 200kbit/sec in MPEG2-Layer3 (a.k.a. MP3) format.

At present, we are capturing a daily news broadcast from

Flemish public broadcaster VRT, and from Flemish commercial broadcaster VTM, as well as one daily news program from the BBC. We also receive higher quality video and autocue data directly from VTM.

To date, we have collected more than 50 hours of English language news broadcasts and 50 hours of Dutch news broadcasts (mostly from VTM), covering February and March 2008. Our intention is to expand our video capture procedures to cover more broadcasts oriented towards specific events in order to obtain more parallel coverage.

Preprocessing. This data is subjected to shot-cut detection and automatic keyframe extraction (Osian & Van Gool, 2004). This is necessary for the later stages of processing, in which computationally more intensive processes can be restricted to keyframes only. Additionally, we detect frontal faces in the keyframes using the method proposed by Viola and Jones (Viola & Jones, 2004) and fit a 3D morphable face model to the data (De Smet et al., 2006). We also extract local features (Bay, Tuytelaars & Van Gool, 2006; Matas, Koubaroulis & Kittler, 2002), and plan to track these over time in the near future. These local features serve as basic image representation for further high-level recognition tasks.

Ongoing work includes cleaning up the teletext output (removing repetitions), as well as an automatic tool for aligning the VTM autocues (without precise timing information) with the subtitles extracted from the teletext (aligned with the actual video data).

```
- <newsitem id="20080301-BBC 1-0400-17" begin="26519" end="29384" subject="Kosovo Independence"
  subjectdetail="Police Presence" subjecttype="fullreport" country="Serbia">
  <anchor begin="26519" end="26696"/>
  <graphics begin="26697" end="26938" type="map"/>
  <transition begin="26939" end="26945" type="crossblend"/>
  - <reportage begin="26946" end="29384">
  <report begin="26942" end="27632"/>
  <interview begin="27633" end="27978" type="1p+onscene"/>
  <report begin="27979" end="28192"/>
  <speech begin="28193" end="28516" type="1p+outdoor"/>
  <report begin="28517" end="28935"/>
  <interview begin="28936" end="29113" type="1p+onscene"/>
  <report begin="29114" end="29384"/>
  </reportage>
  </newsitem>
- <newsitem id="20080301-BBC 1-0400-18" begin="29385" end="29761" subject="Politics"
  subjectdetail="peaceful demonstration" subjecttype="briefreport" country="Spain" city="Madrid">
  <anchor begin="29385" end="29576"/>
  <reportage begin="29577" end="29761" type="brief"/>
  </newsitem>
<transition begin="29761" end="29773" type="crossblend"/>
```

Figure 1: Example ground truth segmentation from a BBC news broadcast.

For a subset of the video material (30-50 news broadcasts), we are generating detailed ground truth information about story segmentation, story classification, and scene classification. See Figure 1 for an example. For this same subset, we also plan to generate ground truth connecting people's names with their faces.

3. Applications

This data is collected and processed as a test bed for the development of new algorithms for multimodal search and retrieval using advanced video and natural language processing technologies. It is, therefore, oriented towards the collection of data that provides useful tests of such technology.

For example, we intend to collect extensive coverage of the US Presidential elections in November 2008. The vast number of stories this event will no doubt generate, in all media and in many languages, makes it an excellent example of the type of event we expect to be able to use. Events will be explicitly located in time and space, in both print and visual media, and routinely described in relation to the present – the time of the news broadcast or article's publication. The elections also offer good conditions for testing algorithms to identify proper nouns and connect them to recognizable faces. Various media of different types will, inevitably, report on the same events, yielding many different reports on the same things and creating a large corpus of comparable texts and video reports. Furthermore, events will be reported on before they happen, and referred to afterwards, offering a straightforward test of text understanding algorithms designed to fix events in time.

We also intend to use coverage of the 2008 Summer Olympics to construct a corpus with narrower focus, as the coverage of the Olympics refers to fewer events outside of the Olympics themselves, and touches less on larger news stories. Furthermore, because the coverage of the Olympics is driven by specific names and times of events, named sportsmen, and explicit data about results (e.g. heights of high jumps, the times of athletes in races), it offers a chance to test algorithms designed to extract structured data from real world information sources.

4. Issues in multimodal data collection

There are a number of issues in constructing multimodal corpora that this project will have to address.

First, our focus on video and text means that there is a rather large disparity in the quantities of data we collect of each type. An hour of video with transcripts is a great deal of image data and a quite small quantity of text. A text corpus of moderate size may easily include thousands of times as much text data as the transcripts of even the largest video corpora. In order to apply the most effective techniques of corpus linguistics, we will need far more text data than the video sources alone will produce. This will involve extending the text portion of the corpus with additional materials.

Second, video data requires considerably more storage space than most other kinds of media. One high quality news broadcast of 30 minutes requires about 10GB of storage, while a comparable length of CD quality stereo audio requires only some 300MB and the corresponding teletext takes on the order of 10KB of space with lossless

compression. Video data can be compressed, but generally not without some reduction in image quality. This project, therefore, has had to make a trade-off between required video quality and available storage.

Third, it is not clear what the basic unit of analysis in multimodal corpora should be. In broadcast video corpora, this is typically the *shot*. In textual media, there are a variety of possible units of analysis - documents, paragraphs, sentences, words or other units – none of which corresponds in an obvious way to shots. This poses a significant challenge both to corpus indexing and to processing.

Fourth, a sizable portion of daily news broadcasts is devoted to coverage of local events and events with little long term interest that are ill-suited to an archive of comparable materials. Using materials from different nations means that many stories that generate significant attention from one broadcaster will be completely ignored by the others. Automatically identifying these cases is another challenge to this project.

5. Related work

There have been a few other efforts to collect large multimodal datasets focused on news. TrecVid (Smeaton et al., 2006) organizes a competition focused on the analysis of video material on a yearly basis and makes a large dataset of American, Chinese and Arab news broadcasts (including some talk shows and commercials) available to its participants. However, since 2005, these have no longer included transcripts or subtitles.

The European IST project *Reveal-This*, has also made a large effort to collect multimodal data (Pastra, 2006). They have collected news programming as well as recordings of European parliament sessions and travel programs. The dataset contains materials in English and Greek.

Lastly, the Informedia project (Hauptman, 2005) has developed a fully automated process for daily content capture, information extraction and online storage. Their library contains more than 1,500 hours of daily news and documentaries produced for government agencies and public television over several years. Only a small part of their dataset has been made available through the OpenVideo Project.³

None of these efforts has focused on combining visual and textual material. Although this list is not intended to be exhaustive, as far as we know, there is no substantial corpus containing both video and a corresponding clean, accurate text, such as might be obtained from subtitles. Moreover, no effort to date includes Dutch language data.

6. Conclusion

We believe this to be a fairly novel class of multimodal resource and one of immediate value in the production of useful natural language and image processing systems

³ <http://www.open-video.org>

with immediate real world applications.

By constructing a multi-modal corpus of news reports indexed topically, we intend to concurrently construct tools for performing alignment across media and using materials in one medium to assist in the segmentation and discovery of searchable features in another. This has potentially broad application in information retrieval, as it lends itself to the production of multimedia summaries and the enhancement of search applications.

7. References

- Aston, G. and Burnard, L. (1998) *The BNC Handbook*. Edinburgh Univ. Press, Edinburgh, UK.
- Bay, H., Tuytelaars, T., & Van Gool, L. (2006) Surf: Speeded up robust features. In: *Proceedings European Conference on Computer Vision 2006*.
- Brants, T. (2001) *TnT – A Statistical part-of-Speech Tagger*. Published online at <http://www.coli.uni-sb.de/thorsten/tnt>.
- De Smet M., R. Fransens, L. Van Gool. (2006) A generalized EM approach for 3D model based face recognition under occlusions, In: *Proceedings IEEE computer society conference on computer vision and pattern recognition - CVPR2006*. vol.2, p.1423-1430.
- Hauptmann, A. (2005) Lessons for the Future from a Decade of Informedia Video Analysis Research. In: *International Conference on Image and Video Retrieval - CIVR'05*. LNCS 3568, pp. 1-10.
- Matas, J., Koubaroulis, D. & Kittler, J. (2002) Robust wide baseline stereo from maximally stable extremal regions. In: *Proceedings of the British Machine Vision Conference*. Vol. 1. pp. 384-389.
- Osian, M., Van Gool, L. (2004) Video shot characterization. *Machine Vision and Applications*, 15 (3), pp. 172-177.
- Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., Baayen, H. (2002) Experiences from the Spoken Dutch Corpus Project. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*. Vol.1, pp. 340-347
- Pastra K. (2006) Beyond multimedia integration: corpora and annotations for cross-media decision mechanisms, In: *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*
- Smeaton, A., Over, P., & Kraaij, W. (2006) Evaluation campaigns and TRECVID. In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval MIR '06*. pp. 321-330.
- Van Eynde, F. (2005) Part-of-Speech Tagging en Lemmatisering van het D-Coi corpus. Annotation Protocol. Centrum voor Computerlinguïstiek, KU Leuven.
- Vandeghinste, V. (2008) *A Hybrid Modular Machine Translation System – LoRe-MT: Low Resources Machine Translation*. LOT, Utrecht.
- Viola P., Jones M. (2004) Robust real-time face detection. *International Journal of Computer Vision (IJCV)* 57(2) 137-154.