Multimedia Documentation Lab

Gerhard Backfried*, Dorothea Aniola*, Klaus Mak, H.C. Pilles**,**
Gerald Quirchmayr*, Werner Winiwarter***, Peter M. Roth******
* Sail Labs Technology AG, Vienna, Austria,
** Documentation Center of the National Defence Academy (NDA), Vienna, Austria,
***University of Vienna, Faculty of Computer Science, Vienna, Austria,
****Inst. f. Computer Graphics and Vision, Graz University of Technology, Graz, Austria

Abstract: In this paper we describe the Multimedia Documentation Lab (MDL[1]), a system which is capable of processing vast amounts of data typically gathered from open sources in unstructured form and in diverse formats. A sequence of processing steps analyzing the audio, video and textual content of the input is carried out. The resulting output is made available for search and retrieval, analysis and visualization on a next generation media server. The system can serve as a search platform across open, closed or secured networks. MDL can be used as a tool for situational awareness, information sharing or risk assessment, allowing the integration of multimedia content into the analysis process of security relevant affairs.

Keywords: Information Systems, Multimedia Computing, Speech Processing, Situational Awareness, Ontologies, Open Source Intelligence
Categories: H.3, H.5.1

1 Introduction

An ever-increasing amount of information is being produced by the second, put on the internet and broadcast by TV- and radio stations. News is produced around the clock and in a multitude of languages. The online content stored on web-pages grows massively and at a constantly accelerating rate and is estimated to already exceed 1.6×10^{20} bytes [1]. An increasingly large portion of this immense pool of data is multimedia content. To tap into this constant flow of information and make the multimedia contents searchable and manageable on a large scale, the MDL project aims to develop a framework and demonstrator which allows the flexible combination of a variety of components for the analysis of the different kinds of data involved. Information and clues extracted from audio- as well as video-tracks of multimedia documents are gathered and stored for further analysis. This is complemented by information extracted from textual documents of diverse qualities and formats. The resulting information is made available on a multi-media server for visualization and analysis.

[1] The MDL-project is part of KIRAS: Österreichisches Sicherheitsforschungs-Förderprogramm - eine Initiative des Bundesministeriums für Verkehr, Innovation und Technologie (BMVIT)

2 Project Scope and Aim

Whereas in the past textual content was the primary source for the extraction and gathering of intelligence in the area of situational awareness, the analysis of multi-media content has been receiving increased attention over the last few years. Information presented on national as well as international multimedia sources complement and extend the information from traditional media; together they form a broad basis for long-term trend analyses and situational awareness, e.g., for conflict and crisis situations. The scope of the MDL system (and project) thus is to allow for the integration of such multi-media content into the existing infrastructure. Content is gathered in a variety of languages and from sources spanning the globe in real-time to allow for short response intervals in crisis situations. Analysis of audio- as well as video-data complements the already existing analysis of textual data. In the process, ontologies serve as the central hub to streamline concepts used by the processing of the different modalities. The goal of the MDL project is to create a demonstrator which will comprise all the mentioned technologies and components and also provide interfaces to the existing infrastructure, allowing for integration into existing workflows.

3 System Description

MDL consists of a set of technologies packaged into components and models, combined into a single system for end-to-end deployment. Together with the components, a number of toolkits are delivered to allow end-users to update, extend and refine models and be able to respond flexibly to a changing environment.

Data enters the system via so-called *Feeders* and then runs through a series of processing steps. Multimedia data is split into an audio- and video-processing track. For audio-data, the processing steps include audio-segmentation, speaker-identification (SID), language-identification (LID), large vocabulary, automatic speech-recognition (ASR) as well as named-entity-detection (NED) and topic-detection (TD). For video-data, the processing comprises scene-detection, key-frame extraction as well as the detection and identification of faces and maps. Textual data is processed by normalization steps before undergoing NED and TD processing.

The resulting documents (in a proprietary XML format or MPEG7) of the individual tracks are fused at the end of processing (*late-fusion*). The XML-files are uploaded, together with a compressed version of the original media files, onto the Media Mining Server (MMS), where they are made available for search and retrieval.

The overall architecture of the MDL System is a server-client one and allows for deployment of the different components on multiple computers and platforms (not all components are multi-platform). Several Feeders, Indexers and Servers (also called Media Mining Feeder, -Indexer and -Server resp.) may be combined to form a complete system. Fig. 1 provides an overview of the components of the MDL System and their interaction.

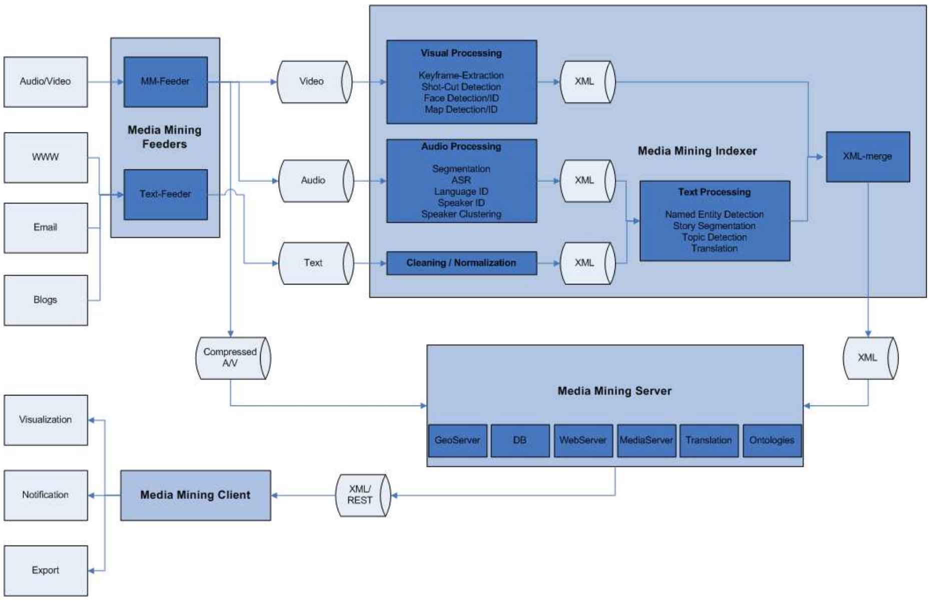

Fig 1. Architecture of the MDL System

3.1 Feeders

The feeders represent the input interface of the MDL System to the outside world. For audio or mixed audio/video input a variety of formats can be imported from external sources and processed by subsequent components. To handle textual input, such as data coming from Web-Pages, e-mails or blogs, separate feeders exist which extract the data from these sources and pass it on to the text processing components.

3.2 Multimedia Feeder

This feeder is based on the Microsoft DirectShow framework and handles a variety of different input sources and formats. Re-encoding is performed to provide Windows Media output files which are uploaded to the Media Mining Server (MMS) for storage and retrieval. The audio channel is passed on to the Media Mining Indexer (MMI) for processing. The video-channel is processed by a series of visual processing components.

3.3 Text Feeder

Text Feeders are specialized feeders which extract textual information from specific sources and pass their results on to the Media Mining Indexer. Two examples of such feeders are the Web-Collector and the e-mail-Collector which are used to gather and process data from internet sources such as web-pages, news-feeds or e-mail-accounts, respectively. The resulting text is cleaned and tokenized before being passed on to the text-processing components.

3.4 Media Mining Indexer (MMI)

The MMI forms the core for the processing of audio and text within the MDL system. It consists of a set of technologies, packaged as components, which perform a variety of analyses on the audio and textual content. Processing results are combined by enriching structures in an XML document. Facilities for processing a number of natural languages exist for the components of the MMI, e.g., ASR is available for more than a dozen languages already. A new set of models for Mandarin Chinese is being developed within the MDL project.

3.4.1 Segmentation

After having been converted to the appropriate format by the feeder, the audio signal is processed and segmented for further analysis. Normalization and conversion techniques are applied to the audio stream which is partitioned into homogeneous segments. Segmentation uses models based on general sounds of language as well as non-language-sounds to determine the most appropriate segmentation point [4]. The content of a segment is analyzed with regard to the proportion of speech contained, and only segments classified as containing a sufficient amount of speech are passed on to the ASR component.

3.4.2 Speaker Identification (SID)

SID is applied to the segments produced by the segmentation step using a set of predefined target models. These models typically comprise a set of persons of public interest. In case a speaker’s identity cannot be determined, the SID system tries to identify the speaker’s gender. Data of the same speaker is clustered and labelled with a unique identifier [3].

3.4.3 Automatic Speech Recognition (ASR)

The Sail Labs speech recognition engine is designed for large-vocabulary, speaker-independent, multi-lingual, real-time decoding of continuous speech. Recognition is performed in a multi-pass manner, each phase employing more elaborate and finer-grained models refining intermediate results, until the final recognition result is produced [5]. Subsequently, text-normalization as well as language-dependent processing (e.g., handling compound-words for German [2]) is applied to yield the final decoding result in a proprietary XML format. The recognizer employs a time-synchronous, multi-stage search using Gaussian tied mixture-models, context dependent models of phonemes and word- as well as (sub-) word based n-gram models. The engine per se is language independent and can be run with a variety of models created for different choices of language and bandwidth.

3.4.4 Language Identification (LID)

Language identification on audio data is used to determine the language of an audio-document in order to allow processing of data using a particular set of speech-recognition models. Textual analysis of language is used to classify text before passing it on to the text-pre-processing components or to the LMT.

3.4.5 Text-based Technologies

The text-based technologies perform their processing either on the output of the ASR-component or on data provided by the text-normalization components. Textual normalization includes the pre-processing, cleaning, normalization and tokenization steps. Language-specific processing of text (e.g., special handling of numbers, compound-words, abbreviations, acronyms) textual segmentation and normalization of spellings are all carried out by these components.

Named entity detection (NED) of persons, organizations or locations as well as numbers is performed on the output of the ASR component, or, alternatively, on text provided by the text normalization components. The NED system is based on patterns as well as statistical models defined over words and word-features and is run in multiple stages [6]. The topic-detection component (TD) first classifies sections of text according to a specific hierarchy of topics. Coherent stories are found by grouping together similar sections. Subsequently, the already classified sections are compared to each other and similar, adjacent sections are merged. The models used for TD and story segmentation are based on support vector machines (SVM) with linear kernels [7].

3.4.6 Toolkits

Currently two toolkits are available to allow user intervention: the Language Model Toolkit (LMT), which allows users to adjust or extend the speech recognition models, and a Speaker ID Toolkit (SIT) allowing users to train SID models for new speakers.

3.4.7 Visual Processing

In order to complement the information extracted from the audio stream of input data, MDL also provides facilities to extract information from the visual signal. In particular, faces of persons and maps are detected and identified. These tasks are typically performed on single images; however, for use by MDL the required data has to be extracted from video streams. Several limitations such as resolution or compression artefacts have to be handled and, due to the large amount of data, only computationally efficient methods can be applied. First, coherent data packages within the visual input referred to as shot boundary detection are identified [8]. To cope with the large amount of data, methods providing a high accuracy such as [9, 10] are applied. These methods either use efficient multi-stage classifiers [9] or directly exploit information provided by the MPEG container [10]. Once the shot boundaries are identified, the recognition steps can be run on the obtained image sequences. For face detection, faces are localized [15] and a recognition step is performed [16].The available temporal information is employed by applying a probabilistic voting over the single image results [12] or by running a tracker and performing the recognition on the identified image locations [13]. More sophisticated approaches inherently using the temporal information in a combined detection/ recognition process [14] will be investigated.

Since neither the position nor the appearance of maps typically change over time, a simple detection/recognition approach is applied. First, images are classified as maps or non-maps [11] and then the positively classified images are identified using a shape-based recognition procedure.

All processing is carried out in separate components, which can readily be plugged-in into the overall system. The results of the individual visual components are collected and output in XML format. This XML is merged (based on time-tags) with the XML output produced by audio- and textual-processing in a subsequent step. The combined XML is then uploaded to the server and made available for search.

3.5 Media Mining Server (MMS)

The MMS comprises the actual server, used for storage of XML and media files, as well as a set of tools and interfaces used to update and query the contents of the database. All user-interaction takes place through the Media Mining Client.

3.5.1 Media Server

The actual server provides the storage for the XML index files, the audio and the video content. It uses Oracle 11g which provides all search and retrieval functionalities. The Semantic Technologies provided by Oracle form the basis for all ontology-related operations within MDL.

3.5.2 Ontologies

An ontology model within the MDL system consists of a set of concepts, instances, and relationships [18]. Ontologies form a central hub within the MDL system and serve to link information originating from different sub-systems together [19]. Translations are associated with concepts in the ontology to allow for concept-based translation. During search, ontologies can be used to widen or narrow the search focus by allowing the user to navigate through the network of concepts. Related concepts can be displayed and examined in a graph according to structural information provided by the ontology. As a first model, a geographic ontology had been created. More advanced usages of topic-related ontologies are currently under development, such as for the field of natural desasters, in particular flood warnings. The goal is to provide an interface in the final version that allows users to flexibly import ontologies which are created with standard ontology tools such as Protégé [20].

3.5.3 Translation

Different types of and interfaces to translation facilities are offered by the MMS [17]. Parallel translations can be created for a transcript as it is uploaded to the server. This is achieved via the integration of 3rd-party machine translation engines. Furthermore keyword translation and human translation, via an e-mail based interface, are supported. Translations are taken into account for queries and visualization.

3.6 Media Mining Client

The Media Mining Client provides a set of features to let users query, interact with, visualize and update the contents of the database. Through a web-browser, users can perform queries, download content, request translations or add annotations to the information stored. Queries can be performed using free text or logical combinations of terms; they can be tailored to address only specific portions of the data stored on the server. Queries can be stored and later used for automatic notification of users to allow for rapid alerting when new documents matching a particular profile appear on the server.

The results of queries are presented according to the structure produced by the MMI. Additional information, such as the names of speakers, named-entities, or detected maps is displayed along with a transcript of the associated audio. Playback of audio and video content can be triggered on a per-segment basis or for the complete document.

3.6.1 Visualization

Special emphasis was given to a series of visualization mechanisms which allow users to display data and its properties in various ways. Search results and summaries can be displayed in a variety of manners in the MDL system. This lets users view data and search-results from different angles, thus allowing them to focus on relevant aspects first and to iteratively circle-in on relevant issues. A globe-view geographically relates events to locations on the globe. A relationship-view relates entities to one another, while a trend-view, relates entities with their occurrences over time. Finally, a cluster-view relates entities and news-sources mentioning them. Further queries can be triggered from all types of visualization, allowing to e.g., summarize events by geography through plotting them on the globe, and subsequently launching further queries by clicking on specific locations on the globe. Ontologies can be used to modify and guide searches. Information derived from ontologies is used for queries (by expanding query terms to semantically related terms) and for the presentation of query results.

4 Status and Outlook

A number of components, such as the feeders, all components which are part of the MMI, and the majority of components which form part of the MMS, are already operational. Other components are currently under development or in the stage of research prototypes, such as the visual processing components or the ontology-related components of the MMS. The existing components have been installed at the end-user’s site in order to allow for rapid feedback during development and for evaluation purposes. New features and technologies are phased in as they become available.

The MDL system presents a state-of-the-art end-to-end Open-Source-Intelligence (OSINT) system. The final demonstrator will include all described audio- and video-processing components. Feeders will be used for monitoring news on a 24x7 level to provide a constant flow of input to the MMI, which continuously processes the incoming data stream and sends its output to the MMS. Likewise, text-feeders will be used to provide constant input from the Web. All processing is targeted to take place in real-time and with minimal latency. Results of all processing will be made available on the MMS and can be exported to the existing infrastructure. It is envisaged that the final version of the MDL System will serve as a core component for the existing Situation Awareness Center (SAC) of the Documentation Center/NDA.

References

[1] C.H.Best, Open Source Intelligence, JRC, European Commission, reference to IDC

[2] R. Hecht, J. Riedler, G. Backfried, Fitting German into N-Gram Language Models, TSD 2002

[3] D. Liu, F. Kubala, Online Speaker Clustering, ICASSP 2003

[4] D. Liu, F. Kubala, Fast Speaker Change Detection for Broadcast News Transcription and Indexing, Eurospeech 1999

[5] R. Schwartz, L. Nguyen, J. Makhoul, Multiple-Pass Search Strategies, Automatic Speech and Speaker Recognition, 1996

[6] D. Bikel, S. Miller, R. Schwartz, R. Weischedel, Nymble: High-Performance Learning Name-Finder, Conference on Applied Natural Language Processing, 1997

[7] T. Joachims, Text Categoriation with Support Vector Machines: Learning with many Relevant Features, in ECML, 1998.

[8] A.F. Smeaton, R. Taban, and P. Over, The TREC-2001 video track report, In Proc. Text Retrieval Conference, 2001.

[9] Y. Kawai, H. Sumiyoshi, and N. Yagi, Shot boundary detection at TRECVID 2007, In Proc. TREC Video Retrieval Evaluation Workshop, 2008.

[10] R. Jinchang, J. Jianmin, and C. Juan, Determination of shot boundary in MPEG videos for TRECVID 2007, In Proc. TREC Video Retrieval Evaluation Workshop, 2008.

[11] M. Michelson, A Goel, and C.A. Knoblock, Identifying maps on the world wide web, In Proc. Int'l Conf. on Geographic Information Science, 2008.

[12] S. Gong, S.J. McKenna, and A. Psarrou, Dynamic Vision: From Images to Face Recognition, Imperial College Press, 2000.

[13] G. Dedeoglu, T. Kanade, and S. Baker, The asymmetry of image registration and its application to face tracking, IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(5):807–823, 2007.

[14] R. Chellappa and S.K. Zhou, Face tracking and recognition from video, In Handbook of Face Recognition, pages 169–192, Springer, 2005.

[15] P. Viola and M.J. Jones, Robust Real-Time face detection, Int'l Journal of Computer Vision, 57(2):137–154, 2004.

[16] S. Kim, S. Chung, S. Jung, S. Jeon, J. Kim, and S. Cho, Robust face recognition using AAM and gabor features, In Proc. World Academy of Science, Engineering and Technology, 2007.

[17] D. Jurafsky, J. H. Martin. Speech and Language Processing, Pearson, 2009

[18] D. Allemang, J. Hendler. Semantic Web for the Working Ontologist. Morgan Kaufman, 2008.

[19] J. Davies, M. Grobelnik, D. Mladenic (eds). Semantic Knowledge Management. Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies, Springer-Verlag, 2009

[20] http://protege.stanford.edu/