REAL-TIME AND DISTRIBUTED AV CONTENT ANALYSIS SYSTEM FOR CONSUMER ELECTRONICS NETWORKS

Jan Nesvadba[1], Pedro Fonseca [1], Alexander Sinitsyn[1], Fons de Lange[1], Martijn Thijssen[1], Patrick van Kaam[1], Hong Liu[2], Rien van Leeuwen[2], Johan Lukkien[3], Andrei Korostelev[1], Jan Ypma[1], Bart Kroon[1], Hasan Celik[1], Alan Hanjalic[4], Umut Naci[4], Jenny Benois-Pineau [5], Peter de With[3], Jungong Han[3]

1 Philips Research, Eindhoven, The Netherlands 2 Philips Applied Technologies, Eindhoven, The Netherlands
3 TU Eindhoven, The Netherlands 4 TU Delft, The Netherlands 5 Labri, Bordeaux, France

ABSTRACT

The ever-increasing complexity of generic *Multimedia-Content-Analysis*-based (MCA) solutions, their processing power demanding nature and the need to prototype and assess solutions in a fast and cost-saving manner motivated the development of the *Cassandra Framework*. The combination of state-of-the-art network and grid-computing solutions and recently standardized interfaces facilitated the set-up of this framework, forming the basis for multiple cross-domain and cross-organizational collaborations [1]. It enables distributed computing scenario simulations for e.g. *Distributed Content Analysis* (DCA) across *Consumer Electronics* (CE) In-Home networks, but also the rapid development and assessment of complex multi-MCA-algorithm-based applications and system solutions. Furthermore, the framework's modular nature - logical MCA units are wrapped into so-called *Service Units* (SU) - ease the split between system-architecture- and algorithmic-related work and additionally facilitate reusability, extensibility and upgradeability of those SUs.

1. INTRODUCTION

Nowadays, terabytes of storage capacity on CE In-Home networks no longer belong to the realm of fiction. Consequently, users of such networks are confronted with a content management problem, e.g. how to retrieve desired AV content stored within networks. Using content descriptors (so-called metadata), either acquired from content-distribution-related services or generated by receiver-MCA-based algorithms, can alleviate this problem. While MCA has meanwhile reached semantically meaningful levels, the MCA systems and related software solutions face such a level of complexity, that modularization of the components into *Service Units* (SUs), is not only desired but also required for complexity-management and reusability [2][3]. Furthermore, modularization allows smart network management systems to balance the processing load across the available resources (also known as grid computing) in applicable networks, such as in-home-, inter-home-, in-vehicle-, but also on-chip networks. Such elaborated DCA systems can be seen as basis for *Ambient Intelligence* (AmI) applicable in various domains, such as CE, medical IT, car infotainment and personal healthcare. This paper describes a prototype realization of a network system for DCA, as well as a number of SUs that are currently integrated in this system, based on advanced MCA algorithms.

While section 2 describes the general system architecture for DCA, SUs and an example scenario, sections 3 to 8 describe the SUs for a media database, feature extraction, AV segmentation, face / object detection, automatic speech recognition and sport genre analysis, respectively. Section 9 concludes the paper.

2. SYSTEM ARCHITECTURE & SERVICE UNITS

The *Cassandra Framework* choices are motivated in three ways.

- First, the trend towards distributed computing and connected (legacy) devices opens up scenarios in which available computational resources are shared to realize complex advanced applications. It is in fact a requirement of the AmI concept that each device "blends in with the crowd", and it is through such cooperation that the value of the network as a whole is higher than the simple sum of its parts.
- Second, the need to quickly prototype and test new applications. MCA algorithms are developed at an increasing rate at many different locations, and are becoming available as black boxes (the SUs in Figure 1/ Figure 2) where the only details regarding their operation concern their input and their output.

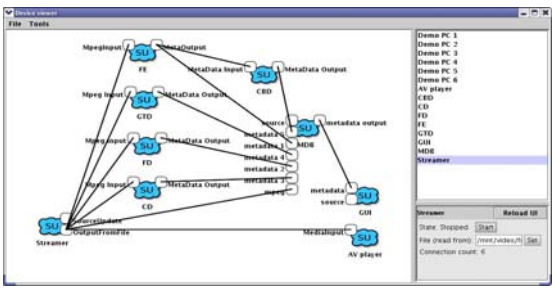

Figure 1. Management GUI

0-7803-9332-5/05/$20.00 ©2005 IEEE

- Third, CE-related devices and their actual operating context steadily increase in complexity, which requires new approaches, like this modularization into SUs.

The hardware part of the introduced framework therefore consists of a collection of nodes, e.g. CE devices and PCs, connected via a network. Consequentially, the software framework is based on the concept that functionality is encapsulated in software components, here SUs, and that those units expose their capabilities as services onto the network. In our prototyping framework, each SU is an independent process that communicates with other SUs using TCP/IP for data streaming and *Universal Plug and Play* (UPnP) for SU control. Every SU is a UPnP device providing a standardized middleware *Application Programmer's Interface* (API) accessible by a UPnP service. In this way, UPnP control points enable system management and monitoring. For example our management *Graphical User Interface* (GUI) application (Figure 1) allows a developer to visualize discovered SUs, to control SU-related parameter, to create connections between SUs and to monitor existing SU-inter-connections. Furthermore, due to UPnP the location and implementation of SUs, which is Operating System and language independent, is transparent to the GUI application.

Major challenges in this architecture correspond to the non-functional aspects, such as performance and robustness, mainly caused by the open, and distributed nature of the networked architecture. This represents additional potential points of failure and of performance loss when compared to single, centralized solutions. With respect to robustness SUs are extended with support for fault detection and correction. The fault model consists of i) communication channel failure, ii) SU failure or iii) node failure. Communication is extended to admit monitoring the presence of connections. In addition, services can be monitored for their availability. In this way a separate *Health Monitoring and Fault Recovery* (HMFR) application, acting as UPnP control point, can take proper action in case a failure is detected e.g. starting a new SU to replace a failing one or changing the service composition.

With respect to performance, automated workload distribution and balancing is under investigation. Also, algorithms that are able to autonomously set up a streaming graph and graph fitting for a specified purpose are being developed.

In our research environment, a prototyping framework for DCA has been set-up, running in real time on a set of networked PCs. This facilitates algorithm developer by several means:

- A straightforward integration of MCA algorithms into the framework. Furthermore, the provision of a standardized connect&control API by the framework and the generation of UPnP-specific code, providing network transparency.
- A SU specific control API can be added easily.
- A set of generic SUs is available: Transcoder, Decoder, AV Recorder etc.
- A real-time visualisation application displaying the resulting metadata of MCA-related SUs with a configurable view.
- For offline (non-real-time) application development in-process streaming can be used. This allows streaming modularity within SUs, without imposing the performance overhead of TCP/IP.

The main target platform of the framework is Linux on x86 and the language used is C++. However, abstractions have been used, which allow easy porting of the framework to comparable architectures.

Figure 2 shows an example of a real-time DCA framework. An incoming AV stream is trans-coded, stored in the *Media DataBase* (MDB) and streamed to the different content-analyzing SUs. The resulting metadata is streamed to the real-time visualization GUI application and to SU MDB for storage. The management GUI application and HMFR application take care of SU control. End-user applications may use the generated metadata to enhance their functionality e.g. AV archiving or AV management.

In the following sections an overview will be given of multiple SUs, which have been integrated in the framework.

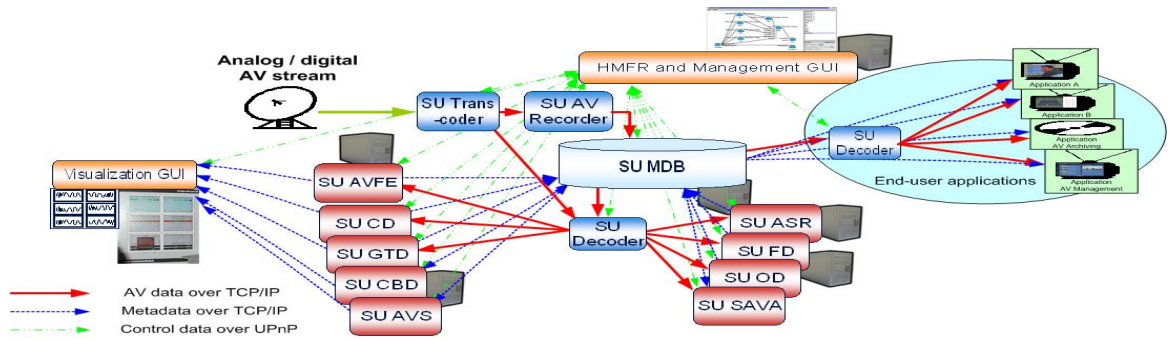

Figure 2. Real-time and distributed multimedia content analysis framework.

3. SERVICE UNIT MEDIA DATABASE

SU MDB, a sort of central memory of the framework, serves as persistent storage for both content data and acquired and generated metadata. Furthermore, SU MDB maintains links between content and content-related metadata, generated by other SUs. Consequentially, the presence of this SU enables delayed, sequential or off-line metadata processing. Moreover, it facilitates AV-management-related applications such as search and retrieval. The current implementation includes a general purpose *Database Management System* (DBMS) with a scheme-dependent metadata abstraction layer. In addition, it includes an extendable set of object-oriented interfaces accepting and expressing metadata in different formats, i.e. UPnP DIDLite and MPEG7. This abstraction makes it possible to use any DBMS underneath, without the need to change the interface. In the near future the central single MDB that is currently used will be distributed into clusters that will act and respond as one single virtual transparent database to the network.

4. SERVICE UNIT AV FEATURE EXTRACTION

SU *AV Feature Extraction* (AVFE), a basic component that will serve various other SUs, represents incoming AV data with appropriate AV low-/mid-level descriptors. This representation serves both the purpose of reducing the amount of input data and of representing the actual data in such a way that higher-level SUs may perform their AV analysis. SU AVFE generates statistical colour-, motion- and texture-based low-level video descriptors, but also various mid-level metadata. Furthermore, it extracts low-level audio descriptors, such as modulation frequencies or audio signal power, and mid-level audio metadata such as audio class categorization. The output of this SU serves as other SUs input, further described in upcoming sections.

5. SERVICE UNIT AV SEGMENTATION

The first SU integrated in our framework with the purpose of extracting semantically meaningful information is SU *AV Segmentation* (AVS). In this SU the AV content items are divided into consistent sections [4], first into shots, by means of Shot Boundary Detectors - such as *Cut Detectors* (CD) [5] and *Gradual Transition Detectors* (GTD) [6] - and then into semantic segments. The latter are derived from video classifiers, e.g. *Commercial Block Detector* (CBD), and an AV segmentation algorithm, which is based on Bayesian decision rules[4].

In more detail, SU CD [5] receives metadata from SU AVFE and automatically indexes abrupt shot transitions [5]. Complementary, SU GTD detects the gradual transitions based on the extraction of relevant gradient-based features from spatio-temporal image blocks after modeling them to detect various transition types including dissolves, fades, and wipes [6]. The gradual transitions are then indexed based on their start and end time stamps. Both the abrupt and the gradual shot transitions are described using MPEG-7 compliant descriptors.

In parallel, SU CBD determines the location of commercial blocks in AV content items by means of a classification algorithm based on the temporal behavior of metadata provided by SU AVFE. This classifier does not aim for the detection of specific features (e.g., the concurrency of silences and video cuts), which can be easily overridden or skipped by broadcasters. Instead, it tries to learn the behavior of what a commercial 'looks and sounds' like and is therefore both robust and future-proof.

As soon as metadata of SU AVFE, CD, GTD and CBD are available, online or offline through the MDB, SU AVS chapters AV content into meaningful segments for, e.g., non-linear browsing [4].

6. SERVICE UNITS FACE / OBJECT DETECTION

Besides temporal AV segmentation, the framework also includes two SUs, one for face and one for object localization, which generate valuable high-level metadata.

The SU *Face Detector* (FD) (and its embedded realization [2]) is originally based on a modified version of the Viola-Jones face detector [8]. This particular face detection algorithm consists of a boosted cascade of simple classifiers and achieves good performance detecting frontal faces. In order to reduce the search window and thus, to improve the efficiency of the algorithm, color-based skin segmentation (on the YCbCr colour space) and tracking were included. The SU FD is currently being extended towards an efficient multi-view face detector [9]. With this extension, the face detection procedure will start by estimating the pose of the face using simple Haar wavelet-like features on successive scales of the image. Hereupon the pose is estimated, and subsequently a face detector is triggered specifically trained for that pose.

The other SU, SU *Object Detector* (OD), is based on the scale-space-based algorithm proposed by Lowe [7], which exhibits robustness to variations in illumination, orientation, scale and viewpoint. Adjustments were made in both the feature- and object detection parts in order to make it possible to use integer operations in most of the processing stages and thus, to allow for its implementation on advanced Philips camera platforms. Currently, the whole object detection algorithm is embedded on such platforms and performs with a typical detection speed of a few hundred milliseconds.

7. SERVICE UNIT SPEECH RECOGNITION

Another semantic-metadata-generating SU, is the SU *Automatic Speech Recognition* (ASR), which analyzes provided audio input data and extracts keywords from identified speech sequences. These keywords are then used to provide the user with a rough semantic categorization with which access to parts of the audio content is provided. Despite on-going progress in the field of large-vocabulary continuous speech recognition, a system as such [10] is still not able to attain "human-like" performance and needs a considerable amount of training data for the intended application areas. This holds equally for the phoneme-based pattern matching part (signal-processing layer) and the language-model-based part of this unit.

8. SERVICE UNIT SPORTS AV ANALYSIS

Finally, SU *Sports AV Analysis* (SAVA), a domain- and application-specific high-level SU, analyzes a sport scenario, in particular tennis matches, focusing on low-/mid-level features to extract high-level semantic meanings of sports scenes. The input of this SU corresponds to segmented AV scenes and face detection as determined by other SUs. The sports analyzer starts by detecting scenes with tennis matches and separates the playing scenes from other shots like breaks or audience views. Afterwards, it carries out a camera calibration procedure [11] to find the relation between image coordinates and the respective position in the real world. The location of the court and of each player is then tracked [12]. Finally, the semantic analysis module first determines the real positions of players in the standard tennis court by making use of a 3-D camera model; based on this, events such as services and highlights are automatically extracted. After detecting such events, an abstract of the game is provided.

9. CONCLUSIONS

The increasing complexity of advanced DCA-based system and application solutions required the introduction of a new research and development framework. After extensive usage of the proposed framework, numerous considerable advantages have been verified.

First of all, the framework enables fast and time-efficient integration, evaluation, verification and demonstration of DCA system- and application solutions, which are based on numerous heterogeneous MCA algorithms developed by various partners in multitudinous cross-collaborations [1]. Furthermore, minimal effort is required to apply the framework in a cross-disciplinary environment due to the clear separation between system and algorithmic development through the usage of SUs (black boxes for the system) and of standardized interfaces.

Finally, the generic nature of the framework allows its efficient usage for numerous application domains including medical IT, automotive, security and robotics.

Acknowledgements. We thank our other team members Robbert Eggermont, Dzevdet Burazerovic, Marc A. Peters, Georg Bauer, Harry Broers, Emile Aarts and Dirk Farin for their active contribution and our ITEA-Candela partner Bosch Security Systems Eindhoven for the collaboration.

10. REFERENCES

[1] Cassandra:www.research.philips.com/technologies/storage/cassandra/, MultimediaN:www.multimedian.nl/, Candela: www.hitech-projects.com/euprojects/candela/

[2] J. Nesvadba, P. Fonseca, et al., "Face Related Features in Consumer Electronic device environments", Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, pp 641-648, The Hague, The Netherlands, 2004.

[3] F. de Lange, J. Nesvadba, "A Networked Hardware/Software Framework for the Rapid Prototyping of Multimedia Analysis Systems", Proc. Int. Conf. on Web Information Systems and Technologies, Miami, USA, 2005.

[4] J. Nesvadba, N. Louis, J. Benois-Pineau, et al., "Low-level cross-media statistical approach for semantic partitioning of audio-visual content in a home multimedia environment", Proc. Int. Workshop on Systems, Signals and Image Processing, pp. 235-238, Poznan, Poland, 2004.

[5] J. Nesvadba, et.al., "Comparison of Shot Boundary Detectors", Proc. Int. Conf. for Multimedia and Expo, Amsterdam, The Netherlands, 2005.

[6] U. Naci, A. Hanjalic, "A Unified Framework for Fast and Effective Shot Transition Detection Based On Analysis Of Spatiotemporal Video Data Blocks", Proc. Int. Workshop on Content-Based Multimedia Indexing, Riga, Latvia, 2005.

[7] D.G. Lowe, "Distinctive image features from scale-invariant key-points", International Journal of Computer Vision, Vol.60, 2, pp.91-110, 2004.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", in Proc. Computer Soc. Computer Vision and Pattern Recognition, vol. 1, HI, pages 511-518, Dec 2001.

[9] J. Nesvadba, A. Hanjalic, et.al. , "Towards a real-time and distributed system for face detection, pose estimation and face-related features", Technical Report, http://www.hitech-projects.com/euprojects/candela/pr.htm (to appear in Proc. Int. Conf. on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 2005).

[10] P. Beyerlein, et.al. , "Automatic Transcription of English Broadcast News", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[11] Dirk Farin, et.al., "Robust camera calibration for sports videos using court models", SPIE Storage and Retrieval Methods and Applications for Multimedia, Vol. 5307, pages 80-91, 2004.

[12] Jungong Han, et.al., "Automatic tracking method for sports video analysis", Proc. Symposium on information theory in the Benelux, Brussels, Belgium, 2005.