Multimodal workbench for automatic surveillance applications

Dragoş Datcu
D.Datcu@tudelft.nl

Zhenke Yang
Z.Yang@tudelft.nl

Léon Rothkrantz
L.J.M.Rothkrantz@tudelft.nl

Man-Machine Interaction Group
Delft University of Technology
2628 CD, Delft,
The Netherlands

1. Introduction

Noticeable developments have lately been achieved on designing automated multimodal smart processes to increase security in every-day life of people. As these developments continue, proper infrastructures and methodologies for the aggregation of various demands that will inevitably arise, such as the huge amount of data and computation, become more important. In this research, we introduce a multimodal framework with support for an automatic surveillance application. The novelty of the attempt resides in the modalities to underpin data manipulation as a natural process but still keeping the overall performance at high levels. At the application level, the typical complexity behind the emerging distributed multimodal systems is reduced in a transparent manner through multimodal frameworks that handle data on different abstraction levels and efficiently accommodate constituent technologies. The proposed specifications includes the use of shared memory spaces (XML data Spaces) and smart document-centered content-based data querying mechanisms (XQuery formal language [1]). We also report on the use of this framework in an application on aggression detection in train compartments.

2. Overview of the work

The challenge to build reliable, robust and scalable automated surveillance systems has interested security people ever since the first human operated surveillance facilities came into operation. Moreover, since the bombings in London and Madrid in 2005, research in methods to detect potentially unsafe situations in public places has taken flight.

Given the size and complexity of the sensing environment surveillance systems have to cope with, including the unpredictable behavior of people interacting in this environment, current automated surveillance systems typically employ diverse algorithms (each focusing on specific features of the sensor data), many sensors (with overlapping sensing area and able to communicate with each other through a network), and different types of sensors (to take advantage of information only available in other modalities).

It is our belief that in modern surveillance applications, satisfactory performance will not be achieved by a single algorithm, but rather by a combination of interconnected algorithms. Our framework is centered on the shared memory paradigm, the use of which allows for loosely coupled asynchronous communication between multiple processing components. This decoupling is realized both in time and space.

The framework (figure 1) enhances the data handling by using a document centered approach to tuple spaces. The shared memory in the current design of the framework takes the form of XML data spaces. All the data is stored in XML documents and these are subsequently received by the data consumers following specific XML queries. This suggests a more human-modeled alternative to store, retrieve and process data. XML Schema [2] is used to validate each existing XML document prior to extracting the meaningful information. Furthermore, the binary data can be easily interchanged via XML documents after converting it using XML MIME protocol. For the implementation, we use Xerces [7] for parsing XML files using DOM/SAX parsers.

In addition, the framework also consists of a set of software tools to monitor the state of registered processing components, to log different type of events and to debug the flow of data given any running application context.

The framework specifications fully comply with the requirements of data manipulation in a multi data producer/consumer context where the availability of data is time-dependent and some connections might be temporarily interrupted.

Depending on the type of application to be built on top of the multimodal framework, a routing algorithm has been designed to manage the data transfers among existing shared memories on different physical networks. This

Acknowledgments. The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

capability is highly required commonly for multimodal applications that involve distinct wireless devices. Considering the study case of an automatic surveillance application, this capability allows wireless devices such as PDAs or mobile phones equipped with video camera to communicate with the system core and to send useful video data.

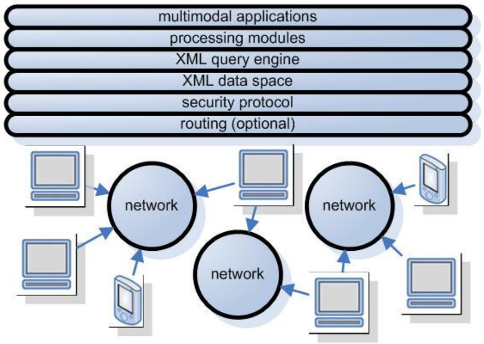

Figure 1: The multimodal framework diagram.

The framework specifications solely emphasize the presence and role of all its components through existing technologies and standards and not on the implementation details. Several proposed technologies present a certain degree of freedom in some functional aspects for the implementation phase. Although the multimodal framework has been designed by taking into consideration the further development of an automatic surveillance oriented application, it can be adopted as basis for any kind of complex multimodal system involving many components and heavy data exchange.

Each data processing component registered in the framework publishes its input and output XML formatted specifications using Web Services Description Language (WSDL) [2].

3. Results

The resulting surveillance application that was built on top of the framework consists of several interconnected detection modules made available by the framework. Each module is specialized in handling a specific task. For example, the sound event detection module detects whether there is someone shouting. The final module, called the aggression detection module, detects unusual behavior by fusing the results of different detection modules. In the application, we have made a distinction between primary and secondary modules. Primary modules require real time data from the sensors and are continuously active. Primary modules are typically feature extraction or object detection algorithms, processing data in the raw data XML and intermediate data XML spaces. The secondary modules operate in semantic data XML space and do not require real time attention and are typically triggered by the results of primary modules. For example, the object detection primary module fuses data from the cameras and detects a stationary object. This triggers the suspicious object secondary module that queries the objects history from the XML tuple space to determine the risk of the object.

For training and testing the algorithms, we conducted some experiments inside a Dutch international train. BeNeLux train compartments are already equipped with cameras and we have used these pre-installed cameras to capture video. The scenarios for the recordings involved a number of hired actors and train conductors playing specific (normal as well as unusual) scenarios. We used four cameras and four microphones in the compartments to capture these scenarios. The data was used as input to a surveillance application built on top of the framework.

The unusual scenarios we asked the actors to play fall in the category of the behaviors we want our system to detect, namely: fighting (including aggression towards the conductor and disturbance of peace), graffiti and vandalism, begging and sickness. The modules used from the framework to detect this behavior include face recognition component, gesture recognition component, face detection component, facial expression recognition component [4], and emotion recognition from speech component [5].

References

[1] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, J. Siméon, XQuery 1.0: An XML Query Language. Candidate Recommendation http://www.w3.org/TR/xquery/, World Wide Web Consortium, 2006.

[2] D. Booth, C. K. Liu, Web Services Description Language (WSDL). Candidate Recommendation http://www.w3.org/TR/wsdl20-primer/, World Wide Web Consortium, 2006.

[3] D. C. Fallside, XML Schema. Technical Report http://www.w3.org/TR/xmlschema-0/, World Wide Web Consortium, 2000.

[4] D. Datcu, L.J.M. Rothkrantz, Facial expression recognition with Relevance Vector Machines. IEEE International Conference on Multimedia & Expo (ICME '05), ISBN 0-7803-9332-5, Jul. 2005.

[5] D. Datcu, L.J.M. Rothkrantz, The recognition of emotions from speech using GentleBoost Classifier. CompSysTech'06, Jun. 2006.

[6] P. Thompson, Ruple: an XML Space Implementation. http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/04-05-03/04-05-03.html, 2002.

[7] Xerces Parser, http://xml.apache.org/xerces-c/pdf.html.