MediaHub:
An Intelligent MultiMedia Distributed Platform Hub

Glenn G. Campbell, Tom Lunney, Paul Mc Kevitt

School of Computing and Intelligent Systems
Faculty of Engineering
University of Ulster, Magee Campus, Northland Road, Derry/Londonderry, BT48 7JL, N. Ireland
{Campbell-g8, TF.Lunney, P.McKevitt}@ulster.ac.uk

***Abstract*- Preliminary research on the development of an intelligent multimedia distributed platform hub (MediaHub) for the fusion and synchronisation of language and vision data is presented. Related research is reviewed and a potential new approach to decision-making within MediaHub based on Bayesian Networks is proposed. A system architecture, including a Dialogue Manager, Semantic Representation Database and Decision-Making Module, is outlined. Bayesian Networks will be employed in the decision-making process within the Decision-Making Module. Initial findings suggest that this will be a promising approach for MediaHub.**

Keywords: intelligent multimedia, distributed systems, multimodal synchronisation, multimodal fusion, multimodal semantic representation, decision-making, Bayesian Networks.

I. INTRODUCTION

The area of intelligent multimedia has seen considerable research into creating user interfaces that can accept multimodal input. This has led to the development of intelligent interfaces that can learn to meet the needs of the user, in contrast to traditional systems where the onus was on the user to learn to use the interface. A more natural form of human-machine interaction has resulted from the development of systems that allow multimodal input such as natural language, eye and head tracking and 3D gestures [1] [2]. Considerable work has also been completed in the area of knowledge representation within multimodal systems, with the development of several semantic mark-up languages [3]. Efforts have also been made to integrate natural language and vision processing, and the main approaches in this field are described in [2].

The area of distributed computing has been exploited to create intelligent multimedia systems that are human-centred and directly address the needs of the user. DACS (Distributed Applications Communication System) [4] is a powerful tool for system integration that provides numerous features for the development and maintenance of distributed systems. Communication within DACS is based on simple asynchronous message passing, with additional extensions to deal with dynamic system reconfiguration during run-time. Other more advanced features include both synchronous and asynchronous remote procedure calls and demand streams.

A. Objectives of MediaHub

The principle aim of the research discussed here is to develop a distributed platform hub (MediaHub) for the fusion and synchronisation of multimodal information, specifically language and vision data. The primary objectives of MediaHub are to:

- Interpret/generate semantic representations of multimodal input/output.
- Perform fusion and synchronisation of multimodal data (decision-making).
- Implement and evaluate MediaHub, a multimodal platform hub with a potential new approach to decision-making.

In pursuing these three objectives, several research questions need to be answered. For example:

- Will MediaHub use frames for semantic representation, or will it use XML or one of its derivatives?
- How will MediaHub communicate with various elements of a multimodal platform?
- Will MediaHub constitute a blackboard or non-blackboard model for semantic storage?
- What mechanism will be implemented for decision-making within MediaHub?

MediaHub will be tested as a plug-in within an existing multimodal platform such as CONFUCIUS [5] using multimodal input/output data.

Next, in section 2, we will look at research related to the development of MediaHub. Then, in section 3, we will focus on multimodal semantic representation. Section 4 discusses decision-making within MediaHub. Section 5 presents the proposed system architecture of MediaHub, while section 6 discusses potential tools and future development of MediaHub.

II. Related Research

This section gives a review of related research that is relevant to the design and implementation of MediaHub. Section 2.1 provides a review of the area of distributed processing, whilst section 2.2 looks at existing multimodal distributed platforms.

A. Distributed Processing

Recent advances in the area of distributed systems have seen the development of several software tools for distributed processing. These tools are utilised in the creation of a range of distributed platforms.

The Open Agent Architecture (OAA) [6] is a general-purpose infrastructure for creating systems that contain multiple software agents. OAA allows such agents to be developed in different programming languages and run on different platforms. All agents interact using the InterAgent Communication Language (ICL). ICL is a logic-based declarative language used to express high-level, complex tasks and natural language expressions.

JATLite [7] incorporates a set of Java packages that enable multi-agent systems to be constructed using Java. JATLite provides a Java agent platform that uses the KQML (Knowledge Query and Manipulation Language) Agent Communication Language (ACL) [8] for inter-agent communication. KQML is a message format and message-handling protocol used to support knowledge sharing among agents.

.NET [9] is the Microsoft Web services strategy that allows applications to share data across different operating systems and hardware platforms. The web services provide a universal data format that enables applications and computers to communicate with one another. Based on XML, the web services allow communication across platforms and operating systems, irrespective of what programming language is used to write the applications.

CORBA [10] is a specification released by the Object Management Group (OMG). A major component of CORBA is the Object Request Broker (ORB), which delivers requests to objects and returns results back to the client. The operation of the ORB is completely transparent to the client, i.e. the client doesn't need to know where the objects are, how they communicate, how they are implemented, stored or executed. CORBA uses the Interface Description Language (IDL), with syntax similar to C++, to describe object interfaces.

B. Multimodal Platforms

Numerous intelligent multimedia distributed platforms currently exist. With respect to these platforms, of particular interest to the design of MediaHub are their methods of semantic representation, storage and decision-making (fusion and synchronisation).

Ymir [11] is a computational model for creating autonomous creatures capable of human-like communication with real users. Ymir represents a distributed, modular approach that bridges between multimodal perception, decision and action in a coherent framework. The modules within Ymir are divided into four process collections. The Reactive Layer operates on relatively simple data. The Process Control Layer controls the global aspects of the dialogue and manages the communicative behaviour of the agent. The Content Layer hosts the processes that interpret the content of the multimodal input and generate suitable responses. The Action Scheduler within Ymir is used to coordinate appropriate actions. There are three main blackboards implemented in Ymir, and communication is achieved via message passing. The first blackboard, called the Functional Sketchboard, is primarily used for information exchange between the Reactive Layer and the Process Control Layer. The second blackboard is called the Content Blackboard. This deals with communication between the Process Control Layer and the Content Layer. The messages that are posted on the Content Blackboard are less time-critical than those on the Functional Sketchboard. The third blackboard is called the Motor Feedback Blackboard and is used to keep track of which part of a stream of actions is currently being planned or carried out by the Action Scheduler. Within the Ymir architecture, a prototype interactive agent called Gandalf has been created. Gandalf is capable of fluid turn-taking and dynamic sequencing.

CHAMELEON [12] is a platform for developing intelligent multimedia applications that makes use of DACS for process synchronisation and intercommunication. The hub of CHAMELEON consists of a dialogue manager and a blackboard. The role of the blackboard is to keep track of interactions over time, using frames for semantic representation. The architecture of CHAMELEON is shown in Fig. 1. CHAMELEON consists of ten modules, mostly programmed in C and C++, which are glued together by the DACS communications system. The blackboard and dialogue manager form the kernel of CHAMELEON. The blackboard stores the semantic representations produced by the other modules, keeping a history of all interactions. Communication between modules is achieved by exchanging semantic representations between themselves or the blackboard.

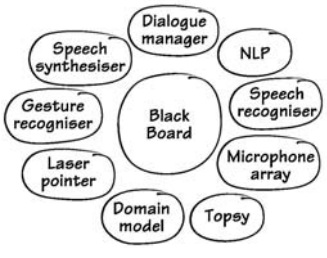

Fig.1. Architecture of CHAMELEON [12]

SmartKom [13] is a multimodal dialogue system that is being developed to help overcome the problems of interaction between people and machines. SmartKom focuses on developing multimodal interfaces for applications in the home, public and mobile domains. The system uses a combination of speech, gestures and facial expressions to facilitate a more natural form of human-computer interaction, allowing face-to-face interaction with its conversational agent Smartakus. For example, in the public domain, the user can allocate the task of finding a library to Smartakus.

MIAMM [14] is an abbreviation for Multidimensional Information Access using Multiple Modalities. The aim of the MIAMM project is to develop new concepts and techniques that will facilitate fast and natural access to multimedia databases using multimodal dialogues.

III. Multimodal Semantic Representation

One of the central questions in the development of intelligent multimedia or multimodal systems is what form of semantic representation should be used. The term 'semantic representation' refers to the method employed to represent the meaning of media representation [3]. This semantic representation must support interpretation and generation, multimodal input and output and a variety of semantic theories. The representation may contain architectural, environmental and interactional information. Architectural information comprises the producer/consumer of the information, information confidence and input/output devices. Environmental representation contains timestamps and spatial information, whilst interactional information includes the speaker/user's state. The majority of the work in multimodal systems employs either frames or XML as the method of semantic representation. A discussion will follow on both of these approaches.

A. Frames

A frame is a collection of attributes with associated values that represent some real world entity. Minsky [15] first introduced frames as a method of semantically representing situations in order to facilitate decision-making and reasoning. The idea of frames is based on human memory and the psychological view that, when faced with a new problem, humans select an existing frame (remembered framework) and adapt it to fit the new situation by changing appropriate details. Although frames have limited capabilities on their own, a frame system provides a powerful mechanism for encoding information to support reasoning and decision-making. Frames can be used to represent concepts, including real world objects, for example "the village of Dromore". The frames used to represent each concept have slots which represents the attributes of the concept. Frame-based methods of semantic representation are implemented in Ymir [11] and CHAMELEON [12].

```
[SPEECH-RECOGNISER
UTTERANCE:(Point to Hanne's office)
INTENTION: instruction!
TIME: timestamp]

[GESTURE
GESTURE: coordinates (3, 2)
INTENTION: pointing
TIME: timestamp]
```

Fig. 2. Example frame from CHAMELEON [12]

Fig. 2 shows an example of the frame semantic representation that is utilised in CHAMELEON. The example frame in Fig. 2 illustrates how speech and gesture input are represented using input frames in the CHAMELEON platform. Note that although the syntax and structure of frames will vary from system to system, the basic idea of knowledge representation will remain the same.

B. XML

Besides frames, the other most popular method of semantic representation in multimodal systems is XML (eXtensible Mark-up Language). XML, created by W3C (World Wide Web Consortium) [16], is a derivative of SGML (Standard Generalised Mark-up Language). XML was originally designed for use in large-scale electronic publishing but is now used extensively in the exchange of data via the web. XML documents contain both parsed and unparsed data, with the former being either mark-up or character data (data between a pair of start and end mark-ups). The mark-up encodes a description of the storage layout and logical structure of the document. A mechanism is provided within XML that allows constraints to be imposed on the storage layout and logical structure. The main purpose of XML is to provide a mechanism that can be used in the mark-up and structuring of documents. XML is different to HTML in that tags are only used within XML to delimit pieces of data. The interpretation of the data is left completely to the application that reads it. Another advantage of using XML is that it is possible to easily create new XML tags.

With respect to semantic representation, SmartKom [13] and MIAMM [14] both use an XML-based method of semantic representation. It is common that a derivative of XML is used for semantic representation. For example, SmartKom uses an XML-based mark-up language, M3L (MultiModal Markup Language), to semantically represent information passed between the various components of the platform. An example of M3L is shown in Fig. 3. The M3L code in Fig. 3 is used to present a list of TV broadcasts to the user in response to a user-request. The exchange of information within MIAMM is also facilitated through a derivative of XML called MMIL (Multi-Modal Interface Language). Any programming language can manipulate data

```
<presentationTask> <presentationGoal>
    <inform> <informFocus> <RealizationType>list </RealizationType>
</informFocus> </inform>
    <abstractPresentationContent>
<discourseTopic> <goal>epg_browse</goal> </discourseTopic>
<informationSearch id="dim24"><tvProgram id="dim23">
  <broadcast><timeDeictic id="dim16">now</timeDeictic>
            <between>2003-03-20T19:42:32 2003-03-
20T22:00:00</between>
      <channel><channel id="dim13"/> </channel>
  </broadcast></tvProgram>
</informationSearch>
       <result> <event>
<pieceOfInformation>
  <tvProgram id="ap_3">
<broadcast> <beginTime>2003-03-20T19:50:00</beginTime>
      <endTime>2003-03-20T19:55:00</endTime>
      <avMedium> <title>Today's Stock News</title></avMedium>
      <channel>ARD</channel>
</broadcast>……..
       </event>  </result>
</presentationGoal> </presentationTask>
```

Fig. 3. Example M3L code [13]

in XML and a range of middleware technology exists for managing data in XML format.

IV. Decision-Making Within MediaHub

The aim of this research is to develop a multimodal platform hub (MediaHub) which will use a potential new approach to decision-making over language and vision data. We will now consider the types of decisions that MediaHub will be required to make. Essentially these can be divided into two main categories:

- Decisions relating to input
- Decisions relating to output

With regard to decisions concerning input, these can be further categorised into the following three areas:

- Determining the semantic content of the input.
- Fusing the semantics of the input (into frames). That is, fuse the semantics of the language input such as "Whose office is this?" with the visual input (i.e. the pointing information/data) [12].
- Resolving any ambiguity at the input.

An example of ambiguity at the input could be if the user points three times while saying "Show me the best possible route from this office to this office" [12]. Here, synchronisation (e.g. using timestamps) could be used to determine which two offices the user is referring to. Another example could be in an industrial environment where a control technician points at two computer consoles saying "Copy all files from the 'process control folder' of this computer to a new folder called 'check data' on that computer." In this example, synchronisation of the visual and audio input may be needed to determine which two computers the control technician is referring to. Resolving ambiguity at the input will be a key objective for the decision-making component of MediaHub.

In relation to decisions at the output, synchronisation issues could arise in order to match, for example, a laser movement with a speech output. As is the case in CHAMELEON [12], a statement of the form "This is the best route from Paul's office to Glenn's office" may need to be synchronised with the laser output tracing the route between the two offices. A decision may also need to be made on what is the best modality to use at the output (i.e. language or vision?). For example, the directions from one office to another may be best presented visually using a laser, while a response to a user's query may be better presented using natural language output. Another example could be when the driver of a car asks an in-car intelligent system for directions to the nearest petrol station. Here the system could respond by presenting a map to the driver or by dictating directions using speech output. The system response in this case would depend on whether or not the car was moving. That is, if the car is stopped in a lay-by, the response could be given to the user via the map. If however the car is moving (i.e. the drivers eyes are pre-occupied on the road), then the system would respond using speech output.

Of course, there are numerous other possible decisions that will be needed in relation to multimodal input and output in MediaHub. Ultimately, the decisions required in MediaHub will depend on its application. The ideal scenario for a multimodal platform hub is that it will be capable of making all possible decisions that could be required in a multimodal system.

V. System Architecture

MediaHub will be an intelligent multimedia distributed platform hub for the fusion and synchronisation of language and vision data. MediaHub's proposed architecture is shown in Fig. 4.

The key components of MediaHub are:

- Dialogue Manager
- Semantic Representation Database
- Decision-Making Module

The role of the Dialogue Manager is to facilitate the interactions between all components of the platform. It will act as a blackboard module, with all communication between components achieved via the Dialogue Manager. It will also

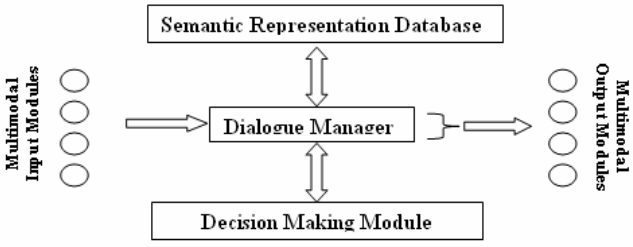

Fig. 4. Architecture of MediaHub

be responsible for the synchronisation of the multimodal input and output.

The Semantic Representation Database in MediaHub will use an XML-based method of semantic representation. XML has been chosen due to its widespread use in the area of knowledge and semantic representation in intelligent multimedia. XML's ease of use will allow it to be easily integrated into MediaHub.

The Decision-Making Module will employ an Artificial Intelligence (AI) technique to provide decision-making on language and vision data. Bayesian Networks and CPNs (Causal Probabilistic Networks) [17] are currently being investigated, to determine if they will be suitable for decision-making within MediaHub. It may also be possible to use other techniques such as Fuzzy Logic, Neural Networks, Genetic Algorithms or a combination of techniques to provide this functionality. With regard to multimodal input and output, existing input/output data structures will be assumed.

Fig. 5 illustrates the flow of data through MediaHub, with the semantic representation, decision-making and synchronisation processes delineated within the dashed rectangle. The circles represent the main processes within the hub. The multimodal input data is first parsed by suitable processing tools and is then passed on to the dialogue manager. The information is then semantically represented using an XML-based semantic representation language. The dialogue manager has the option of using the decision making database, though the data may simply be passed on to the synchronisation process, as indicated in the diagram. It is anticipated that the data flow and the MediaHub architecture will be constantly refined as the development of MediaHub progresses.

VI. Potential Tools And Future Development

The development of MediaHub is in its infancy. Key decisions that lie ahead relate to the method of semantic representation, semantic storage, communication and decision-making. Several implementations of XML could be used by the Semantic Representation Database. Initially, XHTML + Voice may be a suitable choice, since it combines the vision capabilities of XHTML and the speech capabilities of VoiceXML. Other XML-based languages such as the Synchronised Multimedia Integration Language (SMIL) and EMMA (Extensible MultiModal Annotation mark-up language) will also be considered.

A major focus of the future development of MediaHub will be in the area of decision-making over multimodal data. The HUGIN software tool [17], a tool implementing Bayesian Networks as CPNs, will be investigated for its potential to provide MediaHub with decision-making capabilities. Hugin offers an API which is implemented in the form of a library written in the C, C++ and Java programming languages. The API can be used like any other library and can be linked to applications, allowing them to implement Bayesian decision-making. The Hugin API encloses a high performance inference engine that, when given descriptions of causal relationships, can perform fast and accurate reasoning. Whilst Hugin may be used for the development of MediaHub, Microsoft's MSBNx [18] is also a viable option – particularly if the .NET framework is to be used as a distributed processing tool within MediaHub. Other software tools for implementing Fuzzy Logic, Neural Networks and Genetic Algorithms will also be considered.

VII. Conclusion

The objectives of MediaHub, in providing a distributed platform hub for the fusion and synchronisation of language and vision data, have been defined. A review of various existing distributed systems and multimodal platforms has given an insight into the recent advances and achievements in the area of intelligent multimedia distributed computing. The various existing methods of multimodal semantic representation, storage and decision-making, which will be of critical importance in the development of MediaHub, were also considered. The area of Bayesian Networks has been considered with regard to the possibility of using Bayesian decision-making in MediaHub. This provides a potential new approach to decision-making over language and vision data. In conclusion, this paper presents a summary of the motivation for, and future direction of, the development of MediaHub.

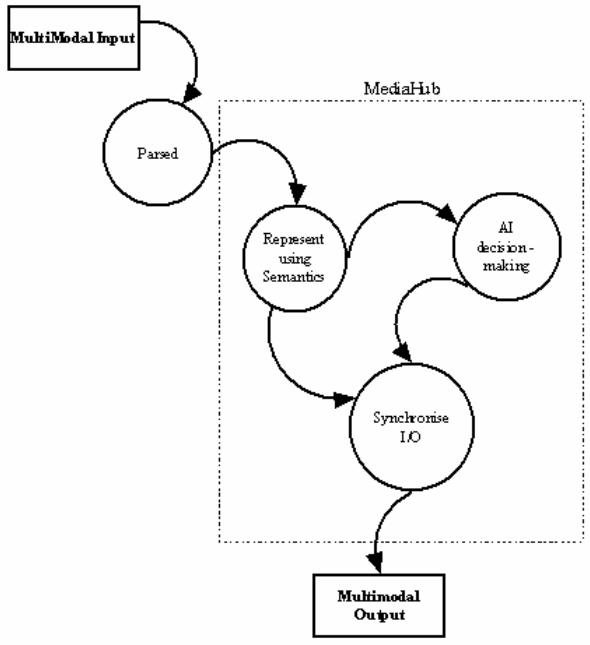

Fig. 5. Data flow in MediaHub

REFERENCES

[1] M.T. Maybury (Ed.), "Intelligent Multimedia Interfaces", Menlo Park: AAAI/MIT Press, 1993.

[2] P. Mc Kevitt (Ed.), "Integration of Natural Language and Vision Processing" (Vols I-IV), London, U.K.: Kluwer Academic Publishers, 1995.

[3] P. Mc Kevitt, "MultiModal semantic representation", In Proceedings of the SIGSEM Working Group on the Representation of MultiModal Semantic Information, First Working Meeting, Fifth International Workshop on Computational Semantics (IWCS-5), Harry Bunt, Kiyong Lee, Laurent Romary, and Emiel Krahmer (Eds.), Tilburg University, Tilburg, The Netherlands, January, 1-16, 2003.

[4] G.A. Fink, N. Jungclaus, F. Kummert, H. Ritter and G. Sagerer, "A Distributed System for Integrated Speech and Image Understanding", International Symposium on Artificial Intelligence, Cancun, Mexico, 117-126, 1996.

[5] M. Ma & P. Mc Kevitt, "Semantic representation of events in 3D animation", In Proc. of the Fifth International Workshop on Computational Semantics (IWCS-5), Harry Bunt, Ielka van der Sluis and Roser Morante (Eds.), 253-281. Tilburg University, Tilburg, The Netherlands, January, 2003.

[6] A. Cheyer, L. Julia, and J.C. Martin, "A Unified Framework for Constructing Multimodal Experiments and Applications", In *Proceedings of CMC '98*: Tilburg, The Netherlands, 63-69, 1998.

[7] T. Kristensen, "T Software Agents In A Collaborative Learning Environment", In International Conference on Engineering Education, Oslo, Norway, Session 8B1, 20-25, August, 2001.

[8] T. Finin, R. Fritzson, D. McKay & R. McEntire, "KQML as an Agent Communication Language", In Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM '94), Gaithersburg, MD, USA, 456-463, 1994.

[9] D.Q.M. Fay, "An architecture for distributed applications on the internet: Overview of microsoft's .net platform", In 17th International Parallel and Distributed Processing Symposium, 7-14, Nice, France, April, 2003.

[10] S. Vinoski, "Distributed object computing with CORBA", C++ Report, Vol. 5, No. 6, July/August, 32-38, 1993.

[11] K.R. Thórisson, "A Mind Model for Multimodal Communicative Creatures & Humanoids", In International Journal of Applied Artificial Intelligence, Vol. 13 (4-5), 449-486, 1999.

[12] T. Brøndsted, P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund and K.G. Olesen, "The IntelliMedia WorkBench - An Environment for Building Multimodal Systems", In Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98, Tilburg, The Netherlands, January 1998, Selected Papers, Harry Bunt and Robbert-Jan Beun (Eds.), 217-233. Lecture Notes in Artificial Intelligence (LNAI) series, LNAI 2155, Berlin, Germany: Springer Verlag, 2001.

[13] W. Wahlster, "SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell", In: Krahl, R., Günther, D. (Eds.), 47-62, Proceedings of the Human Computer Interaction Status Conference, June. Berlin, Germany: DLR, 2003.

[14] N. Reithinger, C. Lauer & L. Romary, "MIAMM: Multimodal Information Access using Multiple Modalities", In Proc. of the International CLASS workshop on Natural, Intelligent and Effective interaction in MultiModal Dialogue Systems, Copenhagen, Denmark, 28-29 June, 2002.

[15] M. Minsky, "A Framework for representing knowledge", In Readings in knowledge representation, R. Brachman and H. Levesque (Eds.), 245-262, Los Altos, CA: Morgan Kaufmann, 1975.

[16] World Wide Web Consortium http://www.w3.org May 2005.

[17] F. Jensen, "Bayesian belief network technology and the HUGIN system", In Proceedings of UNICOM seminar on Intelligent Data Management, Alex Gammerman (Ed.), 240-248. Chelsea Village, London, England, April, 1996.

[18] C.M. Kadie, D. Hovel & E. Horvitz, "MSBNx: A Component-Centric Toolkit for Modeling and Inference with Bayesian Networks", Microsoft Research Technical Report MSR-TR-2001-67, July, 2001.