Adaptive Multimodal Fusion

Pedro Feiteira and Carlos Duarte

LaSIGE and Informatics Department,
Faculty of Sciences of the University of Lisbon,
Campo Grande, 1749-016 Lisboa, Portugal
{pfeiteira,cad}@di.fc.ul.pt

Abstract. Multimodal interfaces offer its users the possibility of interacting with computers, in a transparent, natural way, by means of various modalities. Fusion engines are key components in multimodal systems, responsible for combining information from different sources and extract a semantic meaning from them. This fusion process allows many modalities to be effectively used at once and therefore allowing a natural communication between user and machine. Elderly users, whom can possess several accessibility issues, can benefit greatly from this kind of interaction. By developing fusion engines that are capable of adapting, taking into account the characteristics of these users, it is possible to make multimodal systems cope with the needs of impaired users.

Keywords: Multimodal Interfaces, Adaptive multimodal fusion, Fusion engines, Evaluation.

1 Introduction

Interactive multimodal systems are starting to become widespread, covering many application domains and trying to support a variety of users in the completion of theirs tasks or needs. In the past, multimodal interfaces were centered on modalities such as speech or deictic gestures, but now they are striving to include more and more input options, like full-body gestures, eye gaze, touch interaction with tablets and many others.

By allowing multiple modalities (input and output) to be used, multimodal systems try to replicate a kind of "human-human" interaction with their users, establishing a more natural way of communication with them. This "flexibility" on modality choice, is beneficial for enhancing accessibility, especially for users with physical or cognitive impairments (e.g. visual and motor impairments).

A critical step to be taken in the functioning of a multimodal interactive system is multimodal fusion, which is a process responsible for scanning all the input sources, merging all the information together and making an interpretation out of that. Fusion engines, key components in multimodal systems, executing fusion algorithms and techniques, can also be distinguished by their adaptability capability. When dealing with end-users such as elderly citizens, with so many particularities in their limitations and capabilities, it is crucial to have adaptable user interfaces meeting user requirements.

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2011, LNCS 6765, pp. 373–380, 2011.

Many criteria can be considered when trying to implement an adaptive fusion engine (e.g. noise/error ratio, quality of the recognizers). However, in the scope of project GUIDE, the principle that defines how the fusion mechanism should act, are users and their characteristics.

2 Context

The European project GUIDE (Gentle User Interfaces for Elderly Citizens) intends to deliver a developer oriented toolbox of adaptive, multimodal user interfaces that target the accessibility requirements of elderly users in their home environment, making use of TV set-top boxes as a processing and connectivity platform.

The target users of GUIDE project are elderly people that possess accessibility issues, namely mild to moderate sensory, cognitive and physical impairments (loss or abnormality of psychological, physiological or anatomical structure or function) resulting from ageing or disability (restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being).

Different modalities for input and output will be available in GUIDE, such as speech, gestures, audio and animated avatars. This diversity will allow elderly people to have several options at their disposal to interact with applications. When dealing with this type of users, allowing a "natural" way of communication with the system proves useful, not only because they may be not used to technology, but also because they usually possess certain physical or cognitive impairments that prevents some paths of interaction. Through user models, which contain information about user's impairments or disabilities, the framework is able to adapt to each user and adjust the system behavior in an optimal way to provide the best interaction experience possible.

3 Multimodal Systems

In our everyday lives we are constantly communicating with each other, by means of modalities like vision or gestures, making our life truly multimodal. This kind of dialog is also desirable in HCI, because it would make the interaction with computer feel much more smooth and natural. This desire gave birth to what is called, multimodal systems, which differ considerably from traditional GUI interfaces.

Table 1shows the main differences between these two types of interfaces [1].

Table 1. Main differences between GUIs and MUIs according to Oviatt et al [1]

GUI	MUI
Single input stream	Multiple input streams
Atomic, deterministic	Continuous, probabilistic
Sequential processing	Parallel processing
Centralized architecture	Distributed & time-sensitive architecture

One of the reasons why people use multimodal systems is because they like it. Devices like the mouse, joystick or keyboard limit the ease with which a user can interact in today's computing environments, including, for example, immersive virtual environments. Providing interaction alternatives not only boosts user satisfaction but it also makes systems more robust, due to information combination which makes the weaknesses of a modality (in a certain physical or user related context) be complemented with the strengths of another. Multimodal interfaces have also shown to be more reliable and therefore reduce user's errors by 36% [2].

3.1 Architecture

In Fig. 1 we can observe the general architecture of a multimodal system, along with its key components. This figure demonstrates on a software vision, the message flow in multimodal systems, from the user to system, even including the developer applications. As we can see, input modalities are first perceived through various recognizers, which output their results to the fusion engine, in charge of forming a common interpretation of the inputs. When the fusion engine comes to an interpretation, it communicates it to the dialog manager, in charge of identifying the dialog state, the transition to perform, the action to communicate to a given

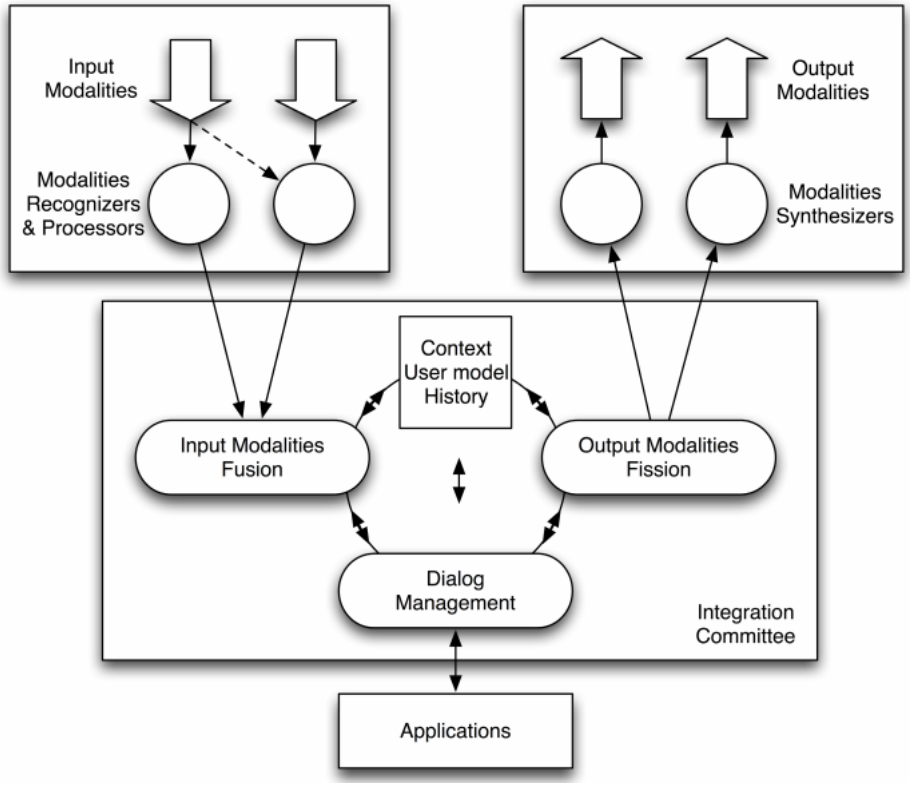

Fig. 1. General architecture of multimodal systems according to Dumas et al [1]

application, and/or the message to return through the fission component. Finally, the fission engine is responsible for returning a message to the user through the most adequate modality or combination of modalities, depending on the user profile and context of use. For this reason, the context manager, in charge of tracking the location, context and user profile, closely communicates any changes in the environment to the three other components, so that they can adapt their interpretations.

4 Multimodal Fusion

In multimodal interactive systems, multimodal fusion is a crucial step in combining and interpreting the various input modalities, and it's one of the distinguishing features that separate multimodal interfaces from unimodal interfaces [1]. The aim of sensor fusion is to analyze many measurements simultaneous, and try to construct semantic meaning from them, which would be harder if only individual measurements were taken into account. Table 2 shows how modalities can be used to interact with multimodal interfaces. The "Use of modalities" columns, expresses the temporal availability of modalities, while the lines represent the fact that information obtained from several modalities can be either combined or treated in an independent fashion. While sequential use of modalities forces the user to use them one at a time, the support for "parallel" use of modalities, allows the user to employ multiple modalities at once, increasing the rate of information transmission between user and system. If this information is further combined, it becomes a synergistic form of interaction.

Table 2. Ways to interact with multimodal interfaces

		USE OF MODALITIES	
		Sequential	**Parallel**
FUSION	**Combined**	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT

Based on the type of information available, requirements, number and type of modalities, different levels of fusion may be considered for a multimodal system. Some of those options will be further discussed in section 4.1.

4.1 Levels of Fusion

Recent literatures consider three levels of fusion: sensor-data, feature and decision-level fusion [1], while others make a distinction between decision-level and opinion-level fusion [3]. According to Sanderson & Paliwal [3], in the former the output of each classifier forms a hard decision, where in the latter a classifier is viewed as an expert which provides an opinion on each possible decision. Information can either be combined before the use of classifiers or experts (pre-mapping) or after a translation from data/feature space to decision/opinion space has taken place (post-mapping).

In the next subsections several types of fusion are considered and explained, such as sensor-data fusion and feature level fusion (pre-mapping). Special focus is given to decision and opinion-level fusion (post-mapping) because they are the most common approaches and most adequate for a system which uses loosely-coupled modalities [1] such as GUIDE.

Sensor-data Level. Data-level fusion, also called sensor-level fusion, deals with raw data coming from recognizers, representing the richest form of information possible (quantitatively speaking). Because the signal is directly processed, no information loss occurs. It is normally used when dealing with multiple signals of the same type, involving a single modality.

Feature Level. Feature-level fusion is oriented for closely-coupled or time synchronized modalities, such as, for example, speech and lips movement. In this type of fusion, features are extracted from data collected by several sensors and are combined later. Unlike data-level fusion, it can suffer from data loss, but manages noise interference better.

Decision Level. One of the most common and widely accepted forms of fusion is . decision-level fusion, and that is because it allows multimodal systems to make effective use of loosely- coupled modalities, like speech and pen interaction. Because the information received by the fusion engines has already been processed, noise and failure are no longer issues to deal with. This means, that fusion has to rely on preprocessed information in order to construct semantic meaning from combining partial semantic information coming from each input mode. That preprocessed information constitutes a hard decision that was produced by one or more classifiers. Those decisions can then be combined using different techniques (e.g. majority voting, ranked lists, AND fusion, OR fusion) [3].

Opinion Level. Opinion-level fusion (also called score-level fusion) is very similar to decision-level fusion because both of them operate after the mapping of data/feature-level space into an opinion/decision space. However, in the case of opinion-level fusion, a group of classifiers, viewed as experts, provides opinions instead of hard decisions, and for that reason Sanderson and Paliwal [3] found more adequate to make a distinction between the two types.

Opinions combination can be achieved, for example, through weighted summation or weighted product approaches [3], before using a classification criterion (e.g. MAX operator) in order to reach a final decision. The main advantage of this approach over decision-level fusion is that opinions from each expert can be weighted, which allows to imprint adaptive features into a system, by setting the reliability and discrimination of experts through time according to the state of the environment/signal quality, users or application logic.

4.2 Adaptive Fusion

Fusion classifiers can be distinguished not only by the type of fusion or architecture they possess, but also by whether they are adaptive or non-adaptive [4]. The basic

concept around adaptive fusion (also called quality fusion) is to assign different weight values associated with a modality. As stated in section 4.1.4 Sanderson and Paliwal [3] pointed out two examples of how such weighting can be used in performing an adaptive opinion fusion; weighted summation fusion and weighted product fusion.

Poh et al [4] state that Adaptivity work as a function of the signal quality measured on one modality. The idea is, the higher quality a signal has, more weight will be set for it. One use of this kind of adaptation is for instance, a person's recognition in a biometric system. Because the light conditions can change and influence the system input (in this case, the face recognition), this visual modality may get a lower weight value whilst speech input would get a higher value, and thus considered more trustworthy in the recognition process. According to Poh & Kittler [4], signal quality can be measured through quality measures. These measures are a set of criteria used to assess the incoming signal quality of a modality. Such measures could be for example, lighting or reflections in face detection and SNR (speech noise ratio) for sound. An ideal quality measure should correlate, to some extent, with the performance of the classifier processing the modality [4]. This means that some characteristics of a classifier prone to affect its performance should make ideal quality measures (e.g. if head pose in face recognition affects the recognition process then head pose would serve as an optimal quality measure).

4.3 Benchmark and Evaluation

Evaluation of multimodal systems has mainly focused so far on user interaction and user experience evaluation [5]. Although this may give valuable insight about multimodal systems, it is only one possible source of errors. Two other important sources that should be considered and studied are modalities recognizers and fusion engines. When a query to the application doesn't produce the expected results any of these three sources is a possible culprit: the user didn't make the proper action to his intent; a recognizer issue came up or delays in system communication made the fusion process fail.

Several frameworks are available today to quickly develop multimodal interfaces with easy plug-in of fusion engines, such as OpenInterface [6] or HephaisTK [7].

Dumas et al. [7] proposed a "divide-and-conquer" approach to evaluate multimodal interfaces, in which each component of a system (e.g. fusion engine, fission engine, recognizers output) is to first be tested on its own, and only later a test of the whole system occurs, with real user-data. The testbed developed by Dumas et al. [7] to use in HephaisTK, enables to focus on fusion algorithms and rules, since the output generated by modalities recognizers are simulated in order to discard all the error they may create and to test specific scenarios. Knowing exactly what information is passed to the fusion engine, it's possible to establish a "ground-truth" set of expected results, which will be compared with the interpretation given by the fusion process. In this manner performance of fusion engines can be evaluated. Some of the metrics defined to measure the quality of engines based on their performance were response time (time between the instant the fusion engine receive its input and the instant it returns

an interpretation), efficiency (comparison of ground-truth data and produced output of the engine), adaptability (to context and user) and extensibility (the capability of supporting new or different input sources).

5 Progress in GUIDE

As stated before, because GUIDE intends to put many loosely-coupled modalities at user's disposal, the choice for a fusion approach becomes narrower, being the possible candidates, decision or opinion-level fusion. Even though these two are very similar, opinion-level fusion by setting weights for different modalities at specific points in time, allows the system to cope with the very particular needs of each user and adapt to it. User profiles are stored and constantly updated by the dialog manager of the system, which allows the fusion engine to access that data and adjust modalities weights accordingly, based on user characteristics and context. All of the input data provided by users shall also have the purpose of analyzing interaction patterns so that profiles will be updated if needed. Since adaptation based on user disabilities and limitations is not yet a much explored area of research, some strategies have to be developed in order to correctly map that data from user profiles into proper modalities weights.

6 Conclusion

In this paper we presented some of the state-of-the-art aspects of multimodal systems and their respective fusion engines. Different approaches to performing fusion are now available, as well as frameworks to quickly develop and integrate them into multimodal interfaces. Many of the ideas and concepts presented will have a significant impact on the design of the GUIDE project multimodal adaptive framework and its respective fusion engine. In this project special focus is given to elderly users whom can possess very particular needs and limitations. Taking that into account adaptability is obviously a critical concern in a system such as GUIDE. The future work will go through studying some of those frameworks available in order to implement and evaluate our initial adaptive fusion algorithms and techniques, which will be greatly influenced by user models.

References

1. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal interfaces: A survey of principles, models and frameworks. Human Machine Interaction 5440(2), 3–26 (2009)
2. Oviatt, S.L.: Multimodal interactive maps: Designing for human performance. Human-Computer Interaction 12(1), 93–129 (1997)
3. Sanderson, C., Paliwal, K.K.: Information fusion and person verification using speech & face information. Research Paper IDIAP-RR 02-33 1(33) (2002)

4. Poh, N., Bourlai, T., Kittler, J.: Multimodal information fusion. In: Multimodal Signal Processing Theory and Applications for Human Computer Interaction, p. 153. Academic Press, London (2010)
5. Lalanne, D., Nigay, L., Palanque, P., Robinson, P., Vanderdonckt, J.: Fusion engines for multimodal input: a survey. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, pp. 153–160. ACM, New York (2009)
6. Lawson, L., Al-Akkad, A.-A., Vanderdonckt, J., Macq, B.: An open source workbench for prototyping multimodal interactions based on off-the-shelf hetero-geneous components. In: Proceedings of the 1st ACM SIGCHI Symposium on Engineering Interactive Computing Systems EICS 2009, p. 245. ACM Press, New York (2009)
7. Dumas, B., Ingold, R., Lalanne, D.: Benchmarking fusion engines of multimodal interactive systems. In: Proceedings of the 2009 International Conference on Multi-modal Interfaces ICMIMLMI 2009, pp. 169–176. ACM Press, New York (2009)