combine to make a host of information available to anyone, at anytime.

Given these trends, the challenge is to organize, understand, and search multimodal information in a robust, efficient and intelligent manner, and to create dependable systems that support natural and intuitive multimodal interaction.

2 Vision and Goals

Multimodality is an inherent property of human cognition and communication. People perceive with all their senses — vision, hearing, smell, touch, and taste — and express themselves naturally by voice, gesture, gaze, facial expression, body posture, and motion. Information technology is increasingly multimodal: computers acquire, store, process, and display a variety of digital information including text, speech, images, video, 3D graphics, as well as geometry and other high-dimensional data arising from science and engineering. However, humans' ability to handle multimodal data exceeds the ability of existing computer systems. Computers are very efficient at processing large, well-structured data sets, but reach their limits when confronted with certain tasks that are intuitive and easy for humans, such as the production and understanding of natural language and the interpretation of visual and auditory stimuli have been notoriously difficult for computers. Furthermore, people handle multimodality in a deeply integrated way. In face-to-face communication, people accompany their utterances by facial expressions and gestures, and listeners simultaneously use verbal and non-verbal cues like gaze and facial expression for comprehension.

We aim to enable natural multimodal interaction with information systems anytime and anywhere, exploiting the wealth of modalities present in everyday human-to-human interaction. Such systems should be naturally accessible for casual users and, for experts, provide novel ways of exploring information. The systems must be aware of each user's environment and situation, must react to speech, text, and gestures and respond with speech, text, video, virtual 3D environments and virtual characters.

Multimodal interaction has at its counterpart multimodal computing that enhances the abilities of computer systems to acquire, process, and present different modes of data in an efficient and robust way. We aim to create systems that can analyze and interpret multimodal information even when it is large, distributed, noisy, and possibly incomplete; that can organize the obtained knowledge to enable powerful querying; and that can produce 3D virtual environments to visualize complex information in real-time.

The envisioned multimodal systems must be available anytime and anywhere, and must be reliable and secure. Therefore, we aim to develop principles for the design, implementation, and operation of dependable autonomous networked systems. In our vision, these systems will be self-organizing, operate autonomously, and respect users' legitimate privacy concerns while simultaneously holding them accountable for their actions.

combine to make a host of information available to anyone, at anytime.

Given these trends, the challenge is to organize, understand, and search multimodal information in a robust, efficient and intelligent manner, and to create dependable systems that support natural and intuitive multimodal interaction.

2 Vision and Goals

Multimodality is an inherent property of human cognition and communication. People perceive with all their senses — vision, hearing, smell, touch, and taste — and express themselves naturally by voice, gesture, gaze, facial expression, body posture, and motion. Information technology is increasingly multimodal: computers acquire, store, process, and display a variety of digital information including text, speech, images, video, 3D graphics, as well as geometry and other high-dimensional data arising from science and engineering. However, humans' ability to handle multimodal data exceeds the ability of existing computer systems. Computers are very efficient at processing large, well-structured data sets, but reach their limits when confronted with certain tasks that are intuitive and easy for humans, such as the production and understanding of natural language and the interpretation of visual and auditory stimuli have been notoriously difficult for computers. Furthermore, people handle multimodality in a deeply integrated way. In face-to-face communication, people accompany their utterances by facial expressions and gestures, and listeners simultaneously use verbal and non-verbal cues like gaze and facial expression for comprehension.

We aim to enable natural multimodal interaction with information systems anytime and anywhere, exploiting the wealth of modalities present in everyday human-to-human interaction. Such systems should be naturally accessible for casual users and, for experts, provide novel ways of exploring information. The systems must be aware of each user's environment and situation, must react to speech, text, and gestures and respond with speech, text, video, virtual 3D environments and virtual characters.

Multimodal interaction has at its counterpart multimodal computing that enhances the abilities of computer systems to acquire, process, and present different modes of data in an efficient and robust way. We aim to create systems that can analyze and interpret multimodal information even when it is large, distributed, noisy, and possibly incomplete; that can organize the obtained knowledge to enable powerful querying; and that can produce 3D virtual environments to visualize complex information in real-time.

The envisioned multimodal systems must be available anytime and anywhere, and must be reliable and secure. Therefore, we aim to develop principles for the design, implementation, and operation of dependable autonomous networked systems. In our vision, these systems will be self-organizing, operate autonomously, and respect users' legitimate privacy concerns while simultaneously holding them accountable for their actions.

Figure 1: Multimodal Computing and Interaction

3 Research Challenges

To make meaningful progress towards the stated goals the excellence cluster plans to specifically address the following issues during the funding period:

3.1 Multimodal Knowledge acquisition, Representation and Retrieval

We will derive the means to acquire, organize and extract information in improved and novel ways. Today, scientific publications available on the Internet are an important information source for scholars. Online encyclopedias like Wikipedia are a major source for students and the public. Digital libraries and thematic portals combine multiple literature and data collections, but provide little integration among sources. In addition, information searching is limited to keywords and simple metadata. Media like video, images, or speech are currently only searchable through manually created annotations.

In the future, knowledge will be automatically acquired and continuously maintained by a suite of methods for natural-language processing, video-content recognition and analysis, information extraction, relation inference, and semantic disambiguation. The resulting knowledge base will be organized with explicit relations and predicate-argument structures and will thus foster flexible and highly accurate searches. Our Open Science Web Demonstrator will show the practical viability of our methods and tools in different sce-

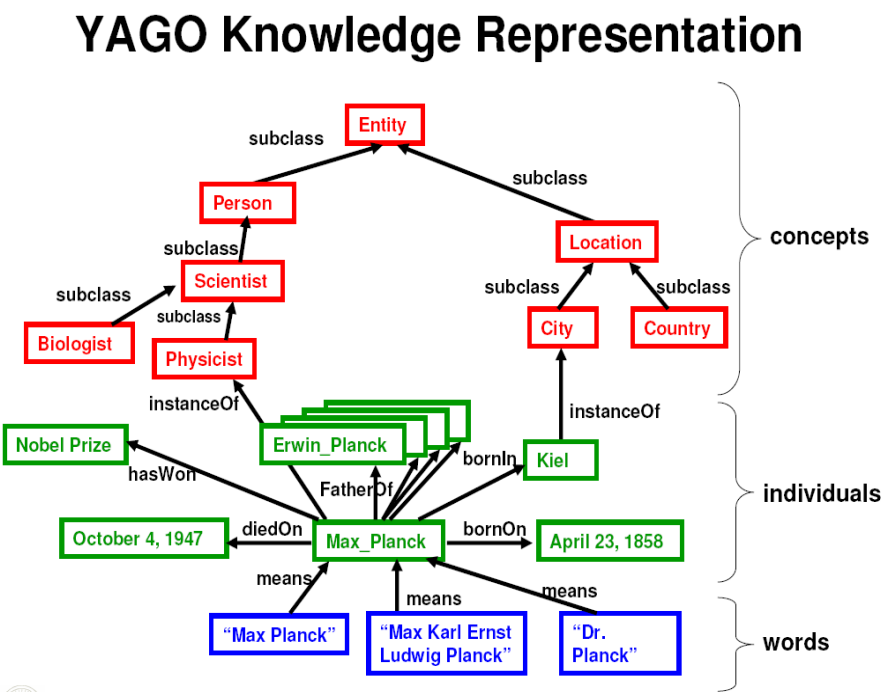

Figure 2: YAGO Knowledge Representation

narios through the creation of an enriched Deep Wikipedia knowledge base with explicitly identified semantic relations and automatically generated connections to relevant sources on the open Web.

3.2 Virtual Environments

We will create convincing virtual environments for enhanced presentation of multimodal data. The visual aspect of these will be realized by realistic and/or abstract synthesis of technical objects like cars or airplanes; by sophisticated postprocessing of existing footage like images, video, or 3D scans; or by a combination of these two basic approaches. We will develop the necessary techniques and software to realize large-scale, integrated, physically accurate and visually rich virtual environments. A somewhat orthogonal but closely related aspect is the creation of human-like synthetic virtual characters. They should look and speak realistically, show convincing emotions, and mimic the behavior of real people in an individualistic and characteristic way. Virtual characters provide a powerful system interface, but can also be used to populate virtual or mixed reality environments.

Figure 3: Virtual Environments

3.3 Symmetric Multimodal Dialog Systems

We aim at the foundation for a new generation of symmetric multimodal dialog systems. These systems will create a natural experience for the user akin to daily human-to-human communication by allowing both the user and the system to combine the same spectrum of modalities for input and for output. As a test scenario that is rich yet controllable, we will use drivers and passengers travelling in the car of the future. We will develop a series of Multimodal In-Car Dialog Demonstrators that enable passengers to interact not only with advanced car services, but also with the environment while in transit. The interactive control of all these services is a rich test bed for multimodal human-to-technology communication. We aim for technologies that comprehend user actions based on computational models of the current task, context, domain, and the user's state and cognitive load to provide appropriate multimodal responses.

3.4 Autonomous, Dependable Computing and Communication Infrastructure:

We will develop principles for the design, implementation and operation of dependable, autonomous networked systems that meet the demands of a pervasive multimodal computing and communications infrastructure. Key assets of these systems will be self-organization as well as infrastructure independence, thus providing computing and communication at

Figure 4: Adaptive Multimodal and Multilingual Human-Computer Dialog Systems

all times in all places. We aim for systems that operate autonomously, with manual administration limited to the installation and replacement of hardware components. These systems must be capable of delivering personalized, relevant and timely information and communication. Moreover, the systems must be dependable and respect users' legitimate privacy concerns, while simultaneously holding them accountable for their actions. Such systems are a necessary platform for the three previously stated objectives.

4 Structure and People

The cluster is structured into nine research areas and coordinated by twelve principal investigators (PIs). Four of these research areas (Text and Speech Processing, Visual Computing, Algorithmic Foundations, Secure Autonomous Networked Systems) are of a more basic character, while the remaining five research areas (Open Science Web, Information Processing in the Life Sciences, Large-Scale Virtual Environments, Synthetic Virtual Characters, Multimodal Dialog Systems) are of a more applied nature. The principal investigators are Michael Backes (UdS-CS), Matthew Crocker (UdS-CL), Peter Druschel (MPI-SWS), Thomas Lengauer (MPI-INF), Kurt Mehlhorn (MPI-INF), Manfred Pinkal (UdS-CL), Hans-Peter Seidel (MPI-INF), Raimund Seidel (UdS-CS), Philipp Slusallek

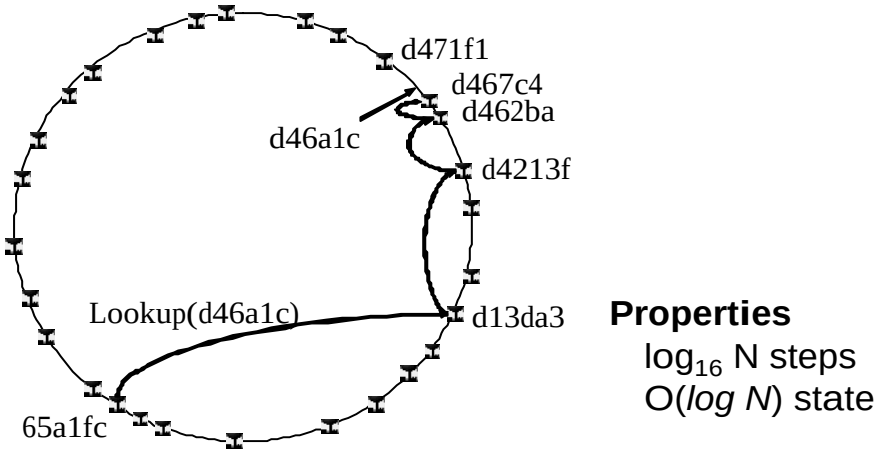

Figure 5: Secure Autonomous Networks – Self-Organizing Structured Overlay

(UdS-CS), Hans Uszkoreit (DFKI), Wolfgang Wahlster (DFKI), Joachim Weickert (UdS-CS), Gerhard Weikum (MPI-INF).

The research areas RA1 Text and Speech Processing (PIs: M. Pinkal, H. Uszkoreit) and RA2 Visual Computing (PIs: J. Weickert, H.-P. Seidel) form the basis of any advanced scientific work on multimodality, since they contribute core expertise in handling the most prominent multimodal data types: text, speech, images and video. In both research areas, we will advance processing methods for raw data to meet the requirements of the envisioned applications concerning efficiency, robustness, and reliability. RA1 and RA2 will jointly explore cross-modal computation techniques, for example, image processing that is linguistically informed or linguistic processing using visual context information. Text and speech, as well as visual processing, can build on rapid advances in these fields, to which Saarbrücken researchers have already substantially contributed. In particular, RA1 is based on a well-established and long-running interdisciplinary collaboration between computer science and computational linguistics.

The investigation of processing methods for linguistic and visual data will be done in close collaboration with RA3 Algorithmic Foundations (PIs: K. Mehlhorn, R. Seidel). More generally, RA3 provides efficient processing techniques for all other areas, in particular, algorithms for handling massive data sets, computational geometry techniques, efficient and effective indexing methods, and general methods for obtaining reliable implementations. Furthermore, computing and interaction is increasingly distributed and needs to

Figure 6: 3D Scanning and 3D Reconstruction - Digital Michelangelo and Minerva of Arezzo

satisfy stringent availability, security and privacy requirements. These requirements are common to most application domains, and are addressed in RA4 Secure Autonomous Networked Systems (PIs: P. Druschel, M. Backes). RA3 and RA4 will cooperate on efficient distributed infrastructures.

The application-oriented research areas RA5 to RA9 depend on and stimulate foundational research in RA1 to RA4. The methods and demonstrator systems stemming from RA5 to RA9 provide important building blocks for the envisioned future multimodal environment.

RA5 Open Science Web (PIs: G. Weikum, H. Uszkoreit) aims at facilitating meaningful answers to advanced queries by collecting, organizing, and semantically understanding information drawn from distributed sources. While RA5 will focus on raw data of comparably high quality, e.g., scientific publications, RA6 Information Processing in the Life Sciences (PIs: T. Lengauer, G. Weikum) concentrates on noisy raw data and its curation, and presentation and visualization in the context of computational biology. The Open Science Web demonstrator will be a joint effort between RA5 and RA6. RA7 Large-Scale Virtual Environments (PIs: P. Slusallek, J. Weickert) will develop the necessary techniques and software to realize visually rich virtual environments, while RA8 Synthetic Virtual Characters (PIs: H.-P. Seidel, W. Wahlster) focuses on the creation of realistic virtual characters with respect to appearance, speech, and behavior. RA9 Multimodal Dialog Systems (PIs: W. Wahlster, M. Pinkal, M. Crocker) will focus on symmetric multimodal dialog systems. It will make use of the realistic virtual characters of RA8 and will be essential for the human-system interface in RA5 and RA7.

While the work in each research area is ambitious in its own right, it is the combination of the different research streams that will allow us to achieve the intended breakthroughs.

Figure 7: International Scope and Promotion of Young Researchers – Graduate students of the International Max Planck Research School (IMPRS) during a project discussion

5 Interdisciplinary Collaboration, Partners, and Expertise

The cluster brings together researchers from Computer Science, Bioinformatics, Computational Linguistics, and Linguistics and Phonetics. Researchers in these fields already work closely and intensively together on the Saarbrücken campus; the cooperation so far can be called exemplary.

The cluster comprises the Computer Science (UdS-CS) and Computational Linguistics and Phonetics (UdS-CL) departments of the Universität des Saarlandes (UdS), the Center for Bioinformatics (UdS-CBI), the Max Planck Institute for Informatics (MPI-INF), the German Research Center for Artificial Intelligence (DFKI), and the newly established Max Planck Institute for Software Systems (MPI-SWS). A total of around 160 PhD-holding researchers work at these institutions; about three quarters of them work within the cluster's scope.

6 Long Term Structural Effects and Promotion of Young Researchers

While there is a long history of successful collaboration among several groups, this marks the first time that all the participating institutions have agreed on a common long-term agenda. This joint effort will promote and enable the progress of science and technology in the field of multimodal computing and interaction. Moreover, it will have long-term structural effects for all partners and have an impact on computer science, in general.

The strength and sustainability of the Saarbrücken cluster will be augmented through several top-level appointments. An integral goal of our cluster is the qualification and promotion of young researchers. It is a goal of paramount importance, given the rapidly growing demand in both academia and industry. Over the past seven years, more than 70 former members of research groups headed by the PIs have received tenured faculty appointments. We will commit the majority of the requested funding for the establishment of junior research groups.

Acknowledgement

The Cluster of Excellence on Multimodal Computing and Interaction was established by the German Research Foundation (DFG) within the framework of the German Excellence Initiative that aims to promote top-level research at German universities and research institutions. More information about the cluster can be found at http://www.mmci.uni-saarland.de.

The Cluster of Excellence on Multimodal Computing and Interaction – Robust, Efficient and Intelligent Processing of Text, Speech, Visual Data, and High Dimensional Representations

Prof. Dr. Hans-Peter Seidel

Max-Planck-Institut Informatik
Stuhlsatzenhausweg 85
66123 Saarbrcken
Germany

Abstract: The past three decades have brought dramatic changes in the way we live and work. This phenomenon is widely characterized as the advent of the Information Society. Ten years ago, most digital content was textual. Today, it has expanded to include audio, video, and graphical data. The challenge is now to organize, understand, and search this multimodal information in a robust, efficient and intelligent way, and to create dependable systems that allow natural and intuitive multimodal interaction.

The Cluster of Excellence on *Multimodal Computing and Interaction*, established by the German Research Foundation (DFG) within the framework of the German Excellence Initiative, addresses this challenge. The term multimodal describes the different kinds of information such as text, speech, images, video, graphics, and high-dimensional data, and the way it is perceived and communicated, particularly through vision, hearing, and human expression.

The cluster comprises the Computer Science and Computational Linguistics and Phonetics departments of Saarland University, the Max Planck Institute for Informatics, the German Research Center for Artificial Intelligence, and the newly established Max Planck Institute for Software Systems. An integral goal of our cluster is the promotion of young researchers, and as such, we will commit the majority of the requested funding to the establishment of junior research groups.

1 Introduction

The past three decades have brought dramatic changes in the way we live and work. This phenomenon is widely characterized as the advent of the Information Society. It is fueled by the power of information technology to acquire, store, process and transmit data compactly, inexpensively and at greater speeds than ever before. Ten years ago, most digital content was textual. Today, graphical and audiovisual I/O devices are in widespread use and modern personal computers have multimedia capabilities. As a result, current digital content additionally comprises speech, audio, video and graphics. Ubiquitous sensing devices will further increase the global volume of digital data. The availability of digital content in different modalities and the increasingly pervasive access to the Internet