eNTERFACE'05
July 18- August 12, 2005 – Faculté Polytechnique de Mons, Belgium

A Multimodal (Gesture+Speech) Interface for 3D Model Search and Retrieval Integrated in a Virtual Assembly Application

Project Title: A Multimodal (Gesture+Speech) Interface for 3D Model Search and Retrieval Integrated in a Virtual Assembly Application

Principal investigator:
Dr. Dimitrios Tzovaras (ITI-CERTH)

Candidates:
Konstantinos Moustakas

Date:
10/3/2005

Abstract:

The goal of the project is the development of a multimodal interface for content-based search of 3D objects based on sketches. This user interface will integrate graphical, gesture and speech modalities to aid the user in sketching the outline of the 3D object he/she wants to search from a large database. Finally, the system will be integrated in a virtual assembly application, where the user will be able to assemble a machine from its spare parts using only speech and specific gestures.

Project objective:

Search and retrieval of 3D objects is a very challenging issue with application branches in numerous areas like recognition in computer vision and mechanical engineering, content-based search in e-commerce and edutainment applications etc. These application fields will expand in the near future, since the 3D model databases grow rapidly due to the improved scanning hardware and modeling software that have been recently developed.

The difficulties of expressing multimedia and especially three dimensional content via text-based descriptors reduces the performance of the text-based search engines to retrieve the desired multimedia content efficiently and effectively. To resolve this problem, 3D content-based search and retrieval (S&R) has drawn a lot of attention in the recent years. A typical S&R system evaluates the similarities between query and target objects according to low-level geometric features. However, the requirement of a query model for searching by example often reduces the applicability of an S&R platform, since in many cases the user knows what kind of object he wants to retrieve but he does not have a 3D model to use as query.

Imagine the following use case: The user of a virtual assembly application is trying to assemble an engine of its spare parts. He inserts some rigid parts into the virtual scene and places them in the correct position. At one point he needs to find a piston and assemble it to the engine. In this case, he has to manually search in the database to find the piston. It would be faster and much more easier if the user had the capability of sketching the outline of the piston using specific gestures combined with speech in order to perform the search.

In the context of this project the integration of speech and gestures in the S&R platform will be addressed. Speech commands are going to be used for selecting categories of objects to be searched and for inquiring the automatic sketching of simple geometrical objects. The system will also use gesture information for deforming the geometrical shapes, for sketching new lines and curves and also for connecting them in order to be able to design complex object outlines.

In addition to the research part of the project, it has also an application part since the final test bed of the system will be a virtual assembly application.

In particular, the objectives of the project during the workshop will be to research and develop fusion strategies and to develop applications in the following areas:

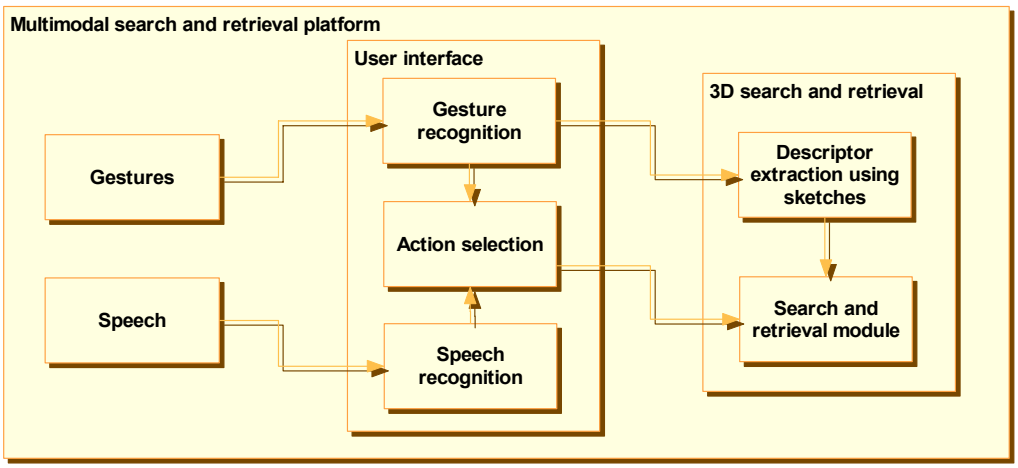
- Gesture and speech recognition
- 3D search and retrieval
- Descriptor extraction using sketches
- Multimodal user interface for 3D S&R
- Speech driven manipulation of the S&R interface
- Prototype implementation of the speech driven 3D S&R platform using sketches.
- Implementation of the virtual assembly application environment

Detailed technical description:

A. Technical description

This project aims mainly to develop a 3D search and retrieval platform, which will not use for query a 3D model, but speech and sketches generated from gestures. The usability of the platform will be tested in a virtual assembly application. It is clear that the project is very ambitious for the short period of 4 weeks. However, it is expect to contribute significantly in the following technical points:

1. **3D S&R platform:** Low-level geometrical characteristics of the objects/parts in the database will be extracted using the spherical trace transform. Variations of the method, which will help to the development of the final multimodal S&R platform, will be studied.
2. **Gesture recognition:** Specific gestures will be used to guide the multimodal S&R platform. In addition, the system will be able to track the motions of the hands so as to extract the sketches of the objects.
3. **Speech recognition:** Specific speech commands will guide the S&R interface as well as the virtual assembly application.
4. **Integration of gesture & speech:** Integration of gesture and speech modalities in a unified multimodal interface.
5. **Definition of "3D sketch descriptors":** The 3D sketches obtained using gesture recognition and tracking, will be processed so as to create their descriptor, which will be used to evaluate the similarity of the sketch with existing 3D models. This is the most challenging research part of the project, since there is no direct link between a 3D sketch (i.e. lines in the 3D space) and a 3D virtual object. Specific gestures, which will correspond to deforming, combining, etc, actions are expected to be useful for the descriptor extraction.
6. **Multimodal 3D S&R platform:** The 3D S&R platform will be manipulated by the multimodal (gesture & speech) interface. The following Figure illustrates a block diagram of the architecture of the platform.



7. **Virtual assembly application:** The multimodal 3D search and retrieval platform will be tested in the context of a virtual assembly application. The testing scenario will be the assembly of a 3D puzzle using sketches.

B. Resources needed

Equipment:

- A personal computer will be needed for each participant.
- At least one of them should be equipped with at least 512 MB of memory and a GeForce 5700 or higher graphics card.
- Cameras for gesture capturing.
- Microphones for voice recording.

Software:

- Speech recognition software (to be provided by the participators)
- Gesture analysis and recognition software (to be provided by the participators)

Staff:

- Experienced researcher(s) in the area of speech recognition
- Experienced researcher(s) in the area of gesture analysis and recognition

C. Project management

The first task will be to split the participants in three groups:

1. 3D modeling, search and retrieval group
2. Speech recognition group
3. Gesture recognition group

During the workshop these groups will have close cooperation, because their tasks will dependent to each other. At least two times per week meetings will be organized to discuss in detail the progress of the work and to plan the work of the following days.

Workplan and implementation schedule:

Due to the restricted duration of the workshop, preliminary work, i.e. literature survey, team discussions, etc., will be done before the starting date of the workshop. During the workshop and for each week the following schedule should be approximately followed.

1st Week:
The first week will be introductory for the participants. Initially, they will be split in 3 groups. Next, they will have to set up their computers according to the tasks they will have to follow during the workshop. The most important tasks during the first week of the project are:

- Discussion – exchange of expertise
- System architecture design
- The two last days of the week will be used to implement simple programs in order to get used to the software, libraries, etc. and to check the feasibility of the proposed system architecture.

2nd Week:
During the second week each group will work on independent tasks, but all groups will be in close cooperation due to the interrelation of their tasks. The tasks of each group during the 2nd week of the project are:

Group 1 (3D modeling, search and retrieval group):

- Development of the S&R graphical interface.
- Development of the multimodal virtual assembly interface
- Development of the sketching system in cooperation with Group 3

Group 2 (Speech recognition group):

- Word vocabulary definition
- Development of the speech recognition system.

Group 3 (Gesture recognition group):

- Gesture vocabulary definition
- Development of gesture recognition
- Development of the sketching system in cooperation with Group 1

3rd Week:
The 3rd week will be dedicated to the integration of the developments of each group in a single platform. Despite the close cooperation of the groups at the 2nd week, during the integration various problems (e.g. communication, synchronization, etc.) are expected to arise. The work done in this week will also address these problems by extending the work of the 2nd week and by designing interfaces for the communication of the different modalities. In particular, the work will address:

- Integration of the gesture recognition in the 3D S&R system.
- Integration of the speech recognition in the 3D S&R user interface.
- Integration of the developed sketching system to the S&R interface.
- 3D object descriptor extraction using sketches.
- Evaluation of the result – Corrections.

4th Week:
During the 4th week the virtual assembly application will be developed, which will make use of the multimodal S&R platform implemented during the 3rd week of the project. Finally, progress, results and evaluation reports will be written and future work and cooperation will be discussed. Summarizing the work of the 4th week:

- Final implementation of the multimodal virtual assembly application.
- Evaluation – Results – Demonstration
- Progress, evaluation and result reports writing.
- Discussion about future work and cooperation.

Expected outcomes:

At the end of the workshop a multimodal 3D model search and retrieval platform should have been produced. In addition a virtual assembly application using the platform should have been implemented. Finally, reports will be also produced describing in detail the work conducted during the workshop, the methodologies used and the solution to specific problems that have been arisen.

Profile of team:

A. Leader (with a brief CV)

Dr. Dimitrios Tzovaras
Dr. Dimitrios Tzovaras is a Senior Researcher Grade C (Assistant Professor) at the Informatics and Telematics Institute. He received the Diploma in Electrical Engineering and the Ph.D. in 2D and 3D Image Compression from the Aristotle University of Thessaloniki, Greece in 1992 and 1997, respectively. Prior to his current position, he was a senior researcher on the Information Processing Laboratory at the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki. His main research interests include virtual reality, haptics, computer graphics, 3D data processing, multimedia image communication, image compression and 3D content-based search. His involvement with those research areas has led to the co-authoring of over thirty articles in refereed journals and more than eighty papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Since 1992, Dr Tzovaras has been involved in more than 20 projects, funded by the EC and the Greek Ministry of Research and Technology. Dr. Tzovaras is an Associate Editor of the Journal on Applied Signal Processing

B. Staff proposed by the leader (with brief CVs)

Konstantinos Moustakas
Konstantinos Moustakas is a PhD student at the Aristotle University of Thessaloniki, Greece. He received the Diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2003. His main research interests include image and video processing, 3D content based search and retrieval, virtual reality, computer graphics and computer vision. Currently he is working on rigid and deformable object modeling and 3D search and retrieval.

C. Other researchers needed (required expertise for each)

In general, researchers with strong background in all areas of the project will be needed. Priority will be given to:

- **Speech recognition:** PhD student(s) or researcher(s) with experience in speech recognition will be needed.
- **Gesture recognition:** PhD student(s) or researcher(s) with experience in gesture recognition and/or real-time hand tracking will be needed.

References:

3D content based search and retrieval:

- P.Daras, D.Zarpalas, D.Tzovaras and M.G.Strintzis: "*3D Model Search and Retrieval Based on the Spherical Trace Transform*", IEEE International Workshop on Multimedia Signal Processing (MMSP 2004), Siena, Italy, October 2004.
- D.Zarpalas, P.Daras, D.Tzovaras and M.G.Strintzis: "*3D Model Search and Retrieval Based on the 3D Radon Transform*", International Conference on Communications (ICC 2004), Multimedia Technologies and Services Symposium (MMTS), Paris, France, June 2004.
- D.V. Vranic and D. Saupe: "*Description of 3d-shape using a complex function on the sphere*". In Proceedings IEEE International Conference on Multimedia and Expo, pp. 177–180, 2002.
- P.Daras, D.Zarpalas, D.Tzovaras and M.G.Strintzis: "*Shape Matching Using the 3D Radon Transform*", 3D Data Processing, Visualization & Transmission (3DPVT 2004), Thessaloniki, September 2004.
- P.Daras, D.Zarpalas, D.Tzovaras and M.G.Strintzis: "*Generalized Radon Transform Based 3D Feature Extraction for 3D Object C classification*", 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2004), Lisboa, Portugal, April 2004.

Speech recognition:

- Kosarev Yu., Ronzhin A., Lee I., Karpov A. "Continuous Speech Recognition without Use of High-Level Information", Proceedings of 15-th International Congress of Phonetic Sciences", Barcelona, Spain, pp. 1373-1376, 2003.
- Duchnow ki P.,Lum D.S.,Krause J.C.,Sexton M.G.,Bratakos M.S., Braida L.D., "Development of speechreading supplements based on automatic speech recognition", IEEE Transactions on Biomedical Engineering,April 2000,47(4):487-496.
- S. Oviat, P. Cohen, L. Wu, J. Vergo, L Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro "Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions", Human Computer Interaction, 15(4):263--322, August 2000.
- Rabiner, L. R., and Schafer R. W., Digital Processing of Speech Signals, Prentice Hall, 1978.
- Alexey A. Karpov, Andrey L. Ronzhin, Alexander I. Nechaev, Svetlana E. Chernakova. Assistive multimodal system based on speech recognition and head tracking, Proc. of the 9-th International Conference SPECOM'2004, St. Petersburg: "Anatoliya", 2004, pp. 521-530
- Kosarev Yu., Lee I., Ronzhin A., Karpov A., Savage J., Haritatos F. "Robust Speech Understanding for Voice control system". Proceedings of Workshop SPECOM'2002, St. Petersburg, pp. 13-18, 2002.

Gesture recognition:

- H. Lee and J.H Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 961-973, 1999.
- C. Keskin, A.N. Erkan, L. Akarun, Real time hand tracking and 3D gesture recognition for interactive interfaces using HMM, Proceedings ICANN/ICONIP, Istanbul, 2003.
- A. Caplier, L Bonnaud, S. Malassiotis, M. Strintzis - Comparison of 2D and 3D Analysis For Automated Cued Speech Gesture Recognition – SPECOM 2004 Proceedings, St Petersburg, Russia, September 2004, pp. 35-41.
- R. Grzeszcuk, G. Bradski, M. H. Chu, and J. Y. Bouguet, "Stereo based gesture recognition invariant to 3D pose and lighting," In Proceedings IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 826-833, 2000
- V. I. Pavlovic, R. Sharma, and T. S. Huang. "Visual interpretation of hand gestures for human-computer interaction: A review" IEEE Trans. Pattern Anal. and Mach. Intell., 19(7): 677-695, July 1997.
- Shimada, N., Y. Shirai, Y. Kuno, and J. Miura, "Hand Gesture Estimation and Model Refinement Using Monocular Camera - Ambiguity Limitation by Inequality Constraints", Proceedings of 3rd Conference of Face and Gesture recognition, pp. 268-273, 1998.