M/ORIS: A MEDICAL/OPERATING ROOM INTERACTION SYSTEM

Sébastien Grange
Laboratoire de Production Robotique
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
+41 (21) 693-7811
sebastien.grange@epfl.ch

Terry Fong
Laboratoire de Production Robotique
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
+41 (21) 693-5850
terrence.fong@epfl.ch

Charles Baur
Laboratoire de Production Robotique
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
+41 (21) 693-2569
charles.baur@epfl.ch

ABSTRACT

We propose an architecture for a real-time multimodal system, which provides non-contact, adaptive user interfacing for Computer-Assisted Surgery (CAS). The system, called M/ORIS (for Medical/Operating Room Interaction System) combines gesture interpretation as an explicit interaction modality with continuous, real-time monitoring of the surgical activity in order to automatically address the surgeon's needs. Such a system will help reduce a surgeon's workload and operation time. This paper focuses on the proposed activity monitoring aspect of M/ORIS. We analyze the issues of Human-Computer Interaction in an OR based on real-world case studies. We then describe how we intend to address these issues by combining a surgical procedure description with parameters gathered from vision-based surgeon tracking and other OR sensors (e.g. tool trackers). We called this approach Scenario-based Procedure and Activity Monitoring (SPAM). We finally present preliminary results, including a non-contact mouse interface for surgical navigation systems.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *Human information processing.*

General Terms

Algorithms, Design, Experimentation, Human Factors, Theory.

Keywords

Medical User Interfaces, CAS, HCI, Multimodal Interaction.

1. INTRODUCTION

Since 2001, a Swiss national research program has been investigating the potential that information technology offers for improving medical procedures and treatment. As part of this effort, we are developing user interface technologies to facilitate the use of computer equipment in the OR. Our long-term goal is to provide automated support services (equipment control, procedure monitoring, etc.) throughout the entire surgical process [7].

Computer Aided Surgery (CAS) can contribute to the general cost cutting trend in health care by making it possible to have fewer staff perform the same surgery in less time than with traditional methods. In particular, it is likely that in the near future, a single surgeon will have to control several computer-based processes during a surgical intervention. The design of an efficient User Interface (UI) that matches the constraints of surgical environments and that helps reduce the surgeon's workload will, in a large part, determine the success of CAS.

To address the need of a UI for CAS, we propose a multimodal framework that lowers the cognitive load of the surgeon when dealing with OR computers, while giving him more direct control over such equipment. Our current research focuses on an important part of this framework, namely a real-time computer-vision approach that combines surgeon detection and tracking, gesture recognition and activity monitoring. We believe that visual gesture recognition is well-suited to the OR for several reasons. First, the OR presents a controlled, well-defined environment. Consequently, variations in illumination (color and intensity) is not a significant problem. Second, a vision-based interface does not require physical contact, which makes it usable even on top of a sterile surgical field. Third, modern CMOS cameras are small, lightweight, and easily movable. Thus, a vision system can be easily integrated into an OR. Finally, visual gesture recognition does not require the surgeon to wear additional hardware (e.g., electromagnetic trackers).

Our design anticipates a scenario in which all computer-assisted tool controls and status are centralized on a console that provides the surgeon with compact information (in the spirit of [5]). The state of the console will change throughout the surgery in order to display relevant information to the surgeon at the appropriate time. The user interface that we are developing will provide two modalities: (1) it will allow the surgeon to explicitly interact with the GUI via gestures; (2) it will monitor the surgeon's activity to infer context information and, when appropriate, automatically adapt the computer-assister equipment to the progress of the procedure. Our system is intended for use with minimally invasive surgery (MIS) because: (1) such procedures typically require computer support (imaging, navigation, etc.) and thus will benefit from improved HCI; (2) there are well-defined periods when the surgeon interacts only with the computer (e.g., equipment or software setup and configuration) and is not using his hands to operate; and (3) there is always at least one OR location (e.g., on top of computer displays) with a clear, unobstructed view of the surgeon.

This paper discusses the relevance of the activity monitoring approach, as well as the requirements it must meet. The following section describes existing work in activity monitoring and medical interfaces. We then discuss the problem at hand and the

goals that must be achieved, before presenting possible ways to reach these goals that we are exploring. Finally, we present some of our results to date, including a computer vision system that enables surgeons to perform standard mouse functions with hand gestures.

2. RELATED RESEARCH

This section describes existing work in the field of gesture-based activity monitoring and surgical UI for CAS.

2.1 Activity Monitoring

It is worth mentioning that the term "activity monitoring" in the literature is somewhat broad, ranging from identifying single gestures (as in [1]) to qualifying the displacement pattern of an entire crowd of people [6] or more complex behaviors spread over time [36]. In vision-based activity monitoring, the object of the monitoring is first segmented and tracked over time. Parameters are then extracted from the movement of the tracked object, and these parameters used to identify the activity of the object from a set of known activities.

2.1.1 Motion and Gesture Tracking

There has been a great deal of work in human feature detection and tracking, as well as gesture segmentation. In [12], Gavrila presents CV techniques for whole-body and hand motion recognition, as well as promising applications. In [23], Moeslund reviews CV-based human motion capture techniques categorized by initialization, tracking, pose estimation, and recognition. In [20], Marcel reviews the taxonomy of the human gesture used in communication contexts, describes recent work in gesture modeling, analysis and recognition for HCI, and discusses their applications. Finally, in [30], Porta presents basic concepts in image processing and user interfaces, before providing a global view of vision-based interfaces (VBIs), with a focus on office and home PC-based use in ordinary computing environments.

In [34], Turk points out that traditional GUIs do not offer the flexibility required to access complex computer features in modern ubiquitous computing environments, and describes how Perceptual User Interfaces (PUI) can potentially provide rich, natural interaction between men and machines. This approach is particularly well suited for the OR, where the sterile, busy context makes it difficult for the surgeon to access the computer's features.

2.1.2 Motion and Gesture Interpretation

Several projects perform activity recognition on specific gestures that are relevant to a particular activity. In [3], Bobick uses Motion Energy Images and Motion History Images to identify whole-body movement in low-resolution video sequences. In [17], Jojic identifies pointing gestures in disparity images, allowing users to interact with a virtual blackboard. In [9], Davis proposes a probabilistic technique to discriminate between standing, walking and running human silhouettes on a single image and multiple viewpoints.

2.1.3 Activity Identification

Many projects address activity classification over a larger scale, both in space and in time. In [4], Bodor uses the velocity and position parameters of a moving object to detect situations in which people may be in peril, as well as suspicious motion or activities at (or near) critical transportation assets. The VSAM project [8] combined multiple smart sensors on a large area to identify moving objects. Human activity is identified using motion analysis on a skeletonized representation of the object and is represented in a centralized, logged fashion. In [11], Fawcett proposes a method to monitor a large amount of data from any sensor to detect changes in activity patterns and find the optimum point in time to trigger alarms. In [16], Hill builds an activity pattern representation of people working in distributed locations to develop awareness of remote co-workers' work rhythm. Chellappa [6] uses the link between object shape and shape deformation as an activity pattern definition to detect abnormal behaviors. In [29], Peixoto uses multiple cameras to autonomously detect, track and log human intrusions in man-made environments.

Among the projects that track advances of complex activities to interact with users, Rickel [32] uses gesture input from a cyberglove to provide skill training and task assistance using a virtual agent through a head-mounted display. Oliver in [28] combines a bottom-up with a top-down approach to identify complex interaction scenarios between humans from visual information, such as following, meeting and splitting, or meeting and walking together. The system, however, can still only discriminate within a known, trained set of activities. In [22], Mikic uses heuristic detection of high-level activity based on a person's ID, location in the room, and speech localization information to classify people according to three activity scenarios (audience, speaker, blackboard presentation.)

2.2 User Interfaces for CAS

CAS techniques, whether they enhance traditional methods (e.g. image visualization [35]) or provide new tools such as augmented displays [10], share a common need for an OR-compatible UI. In [7], Cleary reports on a large scale survey conducted at a workshop on the future of spinal surgery. UI issues are systematically brought up in connection with new computer-assisted techniques, and poor UI design is cited as a significant limiting factor for many operations. In particular, surgeons criticize the lack of user-centered design, the difficulty to operate computer-assisted equipment during surgery, and the failure to convey information without otherwise constraining the surgeon.

To address these issues, several authors have developed guidelines for medical UI design. [33] and [37] stress the importance of a surgeon-centered design for both efficiency and safety. [24] proposes a framework for evaluating the benefits of new UI paradigms in CAS system design.

An interesting simulated application is described by Billinghurst et al. in [2]. They implement an Expert Surgical Assistant capable of interpreting multimodal commands based on voice commands, mechanical gesture tracking and interactions with virtual organs.

ICMI'04, October 13–15, 2004, State Collage, PA, USA.

Natural Language Processing techniques are used to interpret the multimodal input and identify the particular action in a graph representing the simulated procedure. The system is used with a virtual sinus surgery simulator to provide navigation information display and training assistance.

Nishikawa addresses the problem of laparoscopic camera alignment by using the FaceMouse [26], a face-tracking system that controls the camera position according to a face movement grammar. The goal is to replace the human assistant dedicated to this function with a surgeon-controlled system, saving time and frustration. The system proved useful, yet a user survey on a dedicated virtual testbed showed that the FaceMouse can have a negative influence on the surgeon's ability to perform precise actions. In [27], the system is augmented with a tool tracking feature to automatically align the camera with the surgical tool. The goal is to make solo surgery possible. The automated feature is only marginally used during solo surgery, but proves to be very useful.

The Expert Surgical Assistant presented in [2] suggests that multimodal interfaces are necessary for surgical applications. Existing input modalities in the OR suitable for multimodal interaction include:

- mechanical or optical trackers [21], used for tool tracking or object registration. They include traditional trackers (Polhemus, Atracsys) and force-feedback devices (FCS, Force Dimension).
- voice recognition; Grasso [15] explains why the use of promising voice recognition technology is still not as prominent as one might think. The main reason is that clear benefits from using voice commands can only be seen in very specific contexts.
- dedicated tools such as sterilizable keyboards and joysticks, foot and knee pedals, etc.

3. PROBLEM DESCRIPTION

3.1 Problem Formalization

In existing CAS procedures, a human assistant is usually dedicated to controlling the computer. The surgeon gives verbal indications to the assistant on the task to perform or what button to click. This mode of control is remarkably suboptimal, and often yields misunderstandings and frustration for both the surgeon and the assistant. We were given the opportunity to observe such interaction during a CAS procedure similar to [5], in a setup as shown in Figure 1.

3.1.1 Delegated Control Error

During the surgical procedure, we witnessed a misunderstanding between the surgeon and the assistant regarding which button to press. The mistake required the consecutive intervention of 3 other assistants, trying to perform a recovery procedure that the surgeon was giving verbally. Eventually, the surgeon had to desterilize, perform the task himself, then resterilize and resume surgery. In total, 8 minutes and 5 different people were required to perform a single mouse click. While the interface design itself contributed to this misunderstanding, the inefficiency of the assistant-in-the-middle (delegated control) approach was largely responsible for both the error and the difficulty to recover from it.

The same issue related to the frustration of "using" a human assistant to give surgeon access to the computer is also reported in [7] and in [26].

Figure 1. actual CAS setup; the surgeon gives spoken directives to the assistant (to the right of the monitor) to control the computer UI.

3.1.2 Requirements

Our solution to this problem consists in giving the surgeon direct control over the computer UI, removing the assistant from the loop. The non-contact mouse described in Section 4 is a promising first step towards a generic point and click UI in the OR.

However, requiring explicit input from the surgeon to access the computer is not always desirable. As pointed in [27], automated tools can be highly beneficial to the surgical procedure improvement by helping to reduce the surgeon's workload and preventing loss of situational awareness. With the right level of automation, the surgeon can focus on the surgery, not on how to get the computer to perform the right operation.

3.1.3 Characteristics of CAS Procedures

Surgical interventions in general follow detailed protocols, as shown in Figure 2. Computers are typically used in parallel with traditional techniques to provide advanced visualization or navigation information. In practice, each UI goes through various states, each providing different information relevant to the action to perform.

Borrowing from the field of activity theory [18] and its application to HCI [25], it is possible to model a surgical procedure from a UI perspective as a succession of activities. Each activity corresponds to a particular state of the UI, while a

set of relevant actions made of atomic operations can be performed by the surgeon in each state.

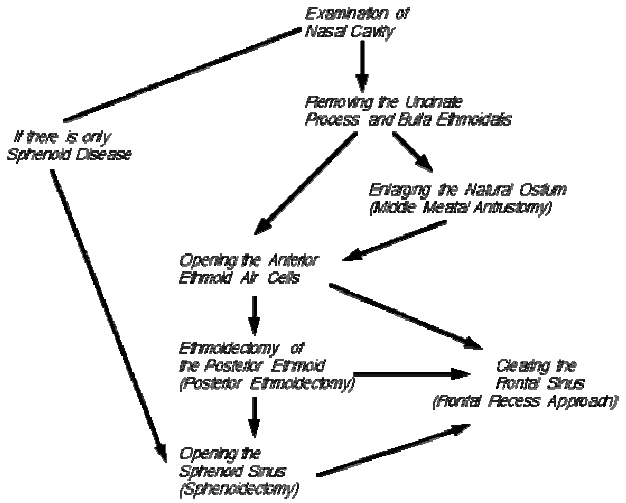

Figure 2 - operative steps in one phase of paranasal sinus surgery, from [31]. The number of steps that need to be completed varies with each patient.

This is best understood with an example. The endo-nasal procedure described in [5] essentially consists of removing polyps from the upper nasal area. It is a delicate intervention, as the surgical tools must be maneuvered close to sensitive areas within the skull. A computer-assisted navigation system has therefore been designed. By identifying the surgical tool and tracking it in real-time, the system shows the position of the tool within a virtual representation of the patient's skull. From the UI perspective, there are 4 different modes of operation, each corresponding to a different activity. Table 1 lists the UI mode:

Table 1 – UI modes for [5]

1. setup	the surgeon loads a file containing the patient's data, including the virtual skull model.
2. calibration	the surgeon registers several predefined points on the patient's skull to align the virtual model
3. testing	a test procedure assesses the accuracy of the calibration; step 2 is repeated if necessary.
4. navigation	the system displays real tooltip position and an endo-camera view on a single display; surgeon can switch between 2D views using a foot pedal

Interaction with the computer occurs mostly to transition from one mode to the next. We believe such transitions could be automated based on visual monitoring of the surgeon, combined with other sensors available in the OR. Combining tool location with surgeon feature detection and tracking provides a rich set of parameters to characterize the surgical activity.

3.2 Limitations of traditional approaches

Traditional activity monitoring techniques, as described in Section 2, are designed primarily to label discrete, separable activities (running, walking, trespassing, pointing, talking, etc.) While they vary in the complexity of the task they can identify, ranging from single gestures to complex behaviors, these approaches are not designed (in general) to track the progress of a known activity. Moreover, most techniques rely on statistical classifiers that are only appropriate for precisely defined situations, and cannot be easily applied to complex scenarios involving branching and looping. Finally, trying to identify atomic gestures and sequences in a surgical procedure with such probabilistic approaches is bound to fail, due to the large variations in gesture between patients, surgeons and setups.

Therefore, we need to define a new way to use sensor data available in a typical CAS setup to infer in real-time the evolution of the surgery with respect to the pre-operative procedure planning. To perform such a task, we introduce the concept of *scenario-based procedure and activity monitoring*.

3.3 Scenario-based activity monitoring

With Scenario-based Procedure and Activity Monitoring (SPAM), our objective is to track the progress of a known scenario with respect to a description of such a scenario. This requires dynamic analysis of both the signals being monitored and of the scenario itself, as the description of the activity can contain branches and inner-loops.

Our approach to perform SPAM in a surgical context is to consider the UI of the computer-based surgical equipment as a state-machine, each state corresponding to a particular task/information the UI must provide/display with respect to the surgeon's need at that stage of the procedure. From the step-by-step representation of our example procedure (Table 1), we can build the state machine shown in Figure 3.

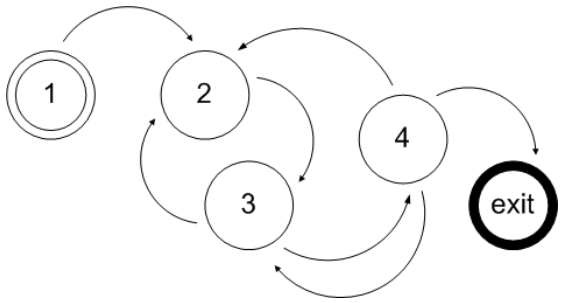

Figure 3. Activity state-machine for the procedure in [5] (corresponding states are described in Table 1).

Each transition from one state to the next can be triggered in numerous ways. The most obvious way will be by explicit command from the surgeon, using a gesture grammar or the non-contact mouse to act directly on the UI. Other UI state changes will be triggered automatically. These automated triggers, or rules, will meet several criterias: (1) they must be human readable to facilitate system design, (2) they must be safe (i.e. unambiguous) to interpret and (3) they must be predictable to the surgeon. Rules include feature location and events, such as:

"If surgeon takes Tool 1 goto State 3"

"If Tool 2 goes from left to right hand, switch view"

"If Tool 2 hasn't moved in 10 seconds, goto State 2"

A different set of rules can be assigned to each state: defined by the surgeon during planning, based on personal preferences or customized for a particular pathology in a given patient.

While setting up rules from the tool tracking is reasonably straightforward, a semantic interpretation of the output of the feature tracker will be required to meet the criterions mentioned earlier. Table 2 gives a list of the typical parameters we anticipate will be available from the visual feature tracking:

Table 2 – visually tracked feature parameters

3D Head position	surgeon head location
3D Torso orientation	with respect to image plane
3D Left/Right Hand	position velocity shape-change

As noted in [19], there is an exhaustive set of possible surgical gestures, meaning that the same gestures are often repeated multiple times. Thus, multi-resolution integration over time of the tracked features parameters (*a priori* parameters) will allow us to: (1) detect changes in activity and (2) qualify and, in some cases, quantify the current activity with parameters that are human-readable. We believe that this approach, which we have already tested in an office environment in [14], will make it possible to extract a semantic description of the gestures such as "the left hand has moved to the right and is performing a different gesture pattern".

Another "benefit" of having a semantic description of the gestures performed is the possibility to generate a very detailed log of the procedure execution. However, the use of such a sensitive tool brings up legitimate concerns in the medical community, and evaluating these human factors is also addressed by our research.

Automation in a surgical environment must be designed very carefully for obvious safety reasons. Thus, any automated UI will have to be *conservative* and *overridable* by the surgeon at any time. Traditional safeguards already used in medical UI, such as validation of sensitive commands, must be considered. Voice recognition validation [15] might be a good candidate for such a task.

4. PRELIMINARY RESULTS

4.1 Multimodal Architecture

We designed M/ORIS, a Medical/Operating Room Interaction System (M/ORIS) that provides a multimodal framework to perform both *explicit* command interpretation and *automated* support during CAS. A conceptual overview of the computer-vision module within M/ORIS is showed in figure 4. SPAM is one of the modules of M/ORIS.

In order to be reliable, SPAM must gather sufficient information from the OR and make its actions known to the surgeon. Several modalities of input and output channels are available to perform both tasks.

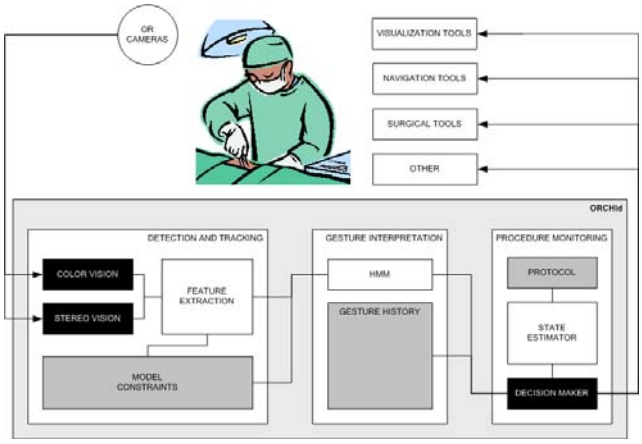

Figure 4. Overview of the M/ORIS computer-vision module.

Among the different input modalities available, SPAM will heavily rely on computer-vision to perform continuous activity monitoring. Other modalities however will be equally important in providing discrete "landmark" events (such as validation commands, start/stop triggers for a particular task, etc.) Speech recognition is an ideal modality for such commands, alongside with more dedicated interaction tools for task-specific control (foot pedals, etc.) M/ORIS will integrate all modalities into a single modal UI.

On top of the visual feedback that is the prime motivation behind M/ORIS, different output modalities must also be used to translate different importance in events. Normal advancement of the procedure (such as state changes) can be briefly reported by a short beep with a particular sound pattern, as sound is a modality that is widely accepted in an OR. Similarly, more critical decisions taken by SPAM will use different sound patterns, or require validation from the surgeon, either via gesture or speech.

Computer Vision is the primary modality used in M/ORIS along with tool tracking, as it allows continuous, non-invasive monitoring of the surgical activity. M/ORIS fuses data from color and stereoscopic vision to perform surgeon features detection and tracking. Detection and tracking is described in section 4.2. The result of the tracking is used to control a non-contact mouse interface, as described in section 4. 3.

4.2 Detection and Tracking

Currently, the system is able to perform reliable, continuous detection and tracking of surgeon features in a simulated surgical environment. We have investigated various methods to fuse color and stereo data, which are appropriate for the OR. Several hypotheses related to the lighting conditions, the size, movement and position of the target features, as well as the specific geometry of a typical surgical setup led to the design of algorithms for reliable detection of a surgeon's face and hands. We placed particular care to design and implement multiple cross-checking mechanisms in order to detect false-positives and compensate for temporary obstructions. These include a labeling algorithm that uses absolute metrics of the human silhouette in the disparity space, combined with color segmentation that takes advantage of the particular lighting conditions found in an OR. The result is a continuous "surgeon detection" algorithm that

achieves real-time (i.e. > 15 Hz), 3D tracking of the surgeon's face and hands. Experiments have been made to track other features such as torso orientation and/or head orientation with satisfying results. Figure 5 shows some results of the detection and tracking algorithm.

Figure 5. Real-time feature detection and tracking (head and hands are robustly detected at 15 Hz).

4.3 Non-contact mouse

Currently the most mature interaction modality in M/ORIS is the non-contact mouse. By combining feature detection with a hand-to-mouse conversion algorithm, we have developed a robust visual gesture mouse. This non-contact system consists of a dedicated hand detection algorithm, a 3D hand tracking algorithm and a 3D hand coordinate to mouse coordinate conversion algorithm that addresses the difference in resolution, dynamics and stability of a human hand with respect to a mouse cursor. Interaction occurs within a fixed workspace, defined by the user, and two clicking modalities are implemented, namely (1) "push-to-click", using a pressing movement of the hand as a trigger, and (2) "wait-to-click", where clicking occurs if no movement occurs during a certain length of time. A detailed description of the system can be found in [13]. Figure 6 shows the output of the hand-tracking algorithm for the non-contact mouse interface from a real surgical setup [5].

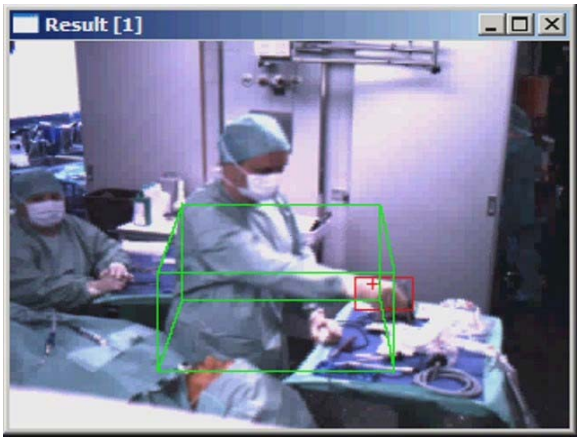

Figure 6. Non-contact mouse in the OR based on real-time hand tracking using color and stereo vision.

4.3.1 System Performance

Because we use both color and depth processing, hand detection works quite well. In particular, having depth information provides two key benefits: (1) it enables us to restrict the search to a pre-defined 3D volume (workspace); and (2) it allows us to match hands using real-world size. As a result, the rate of false-positive and false-negative errors is low.

A hand may appear radically different from image to image, even if the posture seems identical from the human point of view. This is especially true when the hand is moving laterally (with respect to the camera), rotating out of the image plane, or changing form (e.g., switching from open palm to closed fist). Additionally, our current hand detection scheme sometimes identifies only portions of the hand, such as a finger or two. Thus, our hand tracker is designed to recognize changes in hand shape, size, and orientation and to adapt (re-initialize) tracking accordingly.

As with all vision systems, lighting conditions can greatly influence performance. Although we use normalized color to reduce the impact of lighting, our current system has difficulty in situations dominated by saturation effects (e.g., full sunlight) and dynamic changes in intensity. However, since our system is designed for use in OR's, which have controlled lighting and generally do not have exterior windows, this is not a significant problem. In a series of tests, we evaluated our system in a range of ambient lighting conditions. We found that between 200 lux (dim, fluorescent indoor) and 1200 lux (bright, indirect sunlight), both hand detection and tracking worked well.

4.3.2 Usability Tests

To evaluate the usability of the non-contact mouse, we developed a mock-up medical interface. This user interface tests a variety of interaction modalities: menu navigation, button presses, and analog scale setting (2D and 3D). To provide visual feedback, the cursor appearance changes to indicate when mouse control is acquired and when a click is about to be triggered.

In a first set of tests, 16 subjects (including 2 medical students and a perceptual user interface expert) with varied background and computer experience were asked to explore the interface and then to perform various tasks, some of which were timed. At the end of each test session, each subject completed a questionnaire and was asked questions about their experience.

Overall, we found the usability of the system to be good. All subjects were able to rapidly learn how to use the system. We found that navigation and button clicking were the fastest tasks: average time to click anywhere on the full-screen display was less than 5 sec. Setting an analog scale took more time, since cursor positioning needs to be precise. On average, setting a scale to within 1% of the target value required 12 sec. We observed that all subjects initially had difficulty working inside the 3D workspace. At first, users would lose control of the mouse because their hand inadvertently passed out of the workspace. With experience, however, users learned to use rapid hand motion to access all points on the display while keeping their hand in the workspace. In general, subjects preferred the "push-to-click" mode because it provides some (minimal) level of kinesthetic feedback.

We found that there are three primary weaknesses with the current system: (1) the mouse pointer jitters too much under dim lighting conditions; (2) the system has difficulty following rapid

gestures; and (3) user confusion due to perceived differences between hand position and mouse pointer position.

4.3.3 Initial OR Testing

To assess strengths and weaknesses, we installed our system in an OR (Inselspital, September 2003) and collected image data during a computer assisted endoscopic operation. We observed the following:

- There are numerous objects located in the workspace throughout the operation.
- The ambient lighting is generally very dim, in order to provide an acceptable endoscope camera image to the surgeon, but lighting is ideally brighter when interaction is required.
- The endoscope display provides an ideal location for the stereo camera: 1.5 to 2 m from the surgeon with a completely unobstructed view throughout the operation.

After the operation was complete, we conducted a cognitive walkthrough test with the surgeon. This testing revealed the following:

- The surgeon preferred the "wait to click" paradigm because he felt it was easier to use (i.e., requires less hand motion) while offering higher accuracy.
- Adding static hand posture recognition was not felt to be a necessary, nor beneficial, change. In fact, the surgeon argued that static hand gestures would require training and additional concentration, both of which are undesirable given the surgeon's already heavy workload.
- A fixed workspace, defined by surgeon, is compatible with the plan-structured nature of surgery.
- The short delay required for initial hand detection and mouse pointer acquisition was not a problem. In fact, avoiding unintentional cursor control (by explicitly having to engage the system) is considered to be an important design feature.

Overall, the surgeon showed strong interest in the system and was confident that visual gesturing could be useful inside OR's. He emphasized, however, that it is important for the system not to impose additional cognitive load, nor interfere with the way surgical gestures are normally performed.

5. FUTURE WORK

We are currently implementing SPAM on a test procedure (i.e. the complex process of making orange juice) and plan to conduct experiments during the next few months. Evaluation will be performed on a range of applications conceptually similar to surgery, such as following the progress of a kitchen recipe, and on a database of 2 hours of multimodal surgical data. Other tools are currently being investigated to complement or reinforce M/ORIS, such as a HMM-based pose and gesture recognition module.

ACKNOWLEDGEMENTS

We would like to thank Dr. M. Caversaccio (Inselspital) and Dr. P.-Y. Zambelli (Orthopedic Hospital of Suisse Romande) for providing valuable comments and their medical insight. This work was supported by a grant from the Swiss National Science Foundation Computer Aided and Image Guided Medical Interventions (NCCR CO-ME) project.

REFERENCES

[1] M. Bichsel and A. P. Pentland, "Automatic Interpretation of Human Head Movements", *JICAI Workshop on Looking At People*, Chambery France, August 30, 1993.

[2] M. Billinghurst, J. Savage, P. Oppenheimer, C. Edmond, "The Expert Surgical Assistant: An Intelligent Virtual Environment with Multimodal Input", *Medicine Meets Virtual Reality IV*, Amsterdam, 1996.

[3] A. F. Bobick and J. W. Davis, "The Recognition of Human movement Using Temporal Templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, March 2001.

[4] R. Bodor, B. Jackson and N. Papanikolopoulos, "Vision-Based Human Tracking and Activity Recognition," *Proc. of the 11th Mediterranean Conf. on Control and Automation*, June 18-20, 2003.

[5] M. Caversaccio, R. Baechler, K. Laedrach, G. Schroth, L.-P. Nolte and R. Haeusler, "The Bernese Frameless Optical Computer Aided Surgery System", *Computer Aided Surgery*, no. 4, pp. 328-334, 1999.

[6] R. Chellappa, N. Vaswani, and A. K. R. Chowdhury, "Activity Modeling and Recognition using Shape Theory", *Behavior Representation in Modeling and Simulation*, 2003.

[7] K. Cleary, "Workshop Report", *Workshop on Technical Requirements for Image-Guided Spine Procedures*, April 17-20, Georgetown, USA, 1999.

[8] Collins, R. T., Lipton, A. J., Fujiyoshi, H., Kanade, T., "Algorithms for Cooperative Multisensor Surveillance", *Proc. of IEEE,* Vol.89, pp1456-1477, Oct. 2001.

[9] J. W. Davis and A. Tyagi, "A Reliable-Inference Framework for Recognition of Human Actions", *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 169-176, Miami, Florida, July 2003.

[10] J. Ellsmere, J. Stoll, D. Rattner, D. Brooks, R. Kane, W. Wells, R. Kikinis and K. Vosburgh, "A Navigation System for Augmenting Laparoscopic Ultrasound", *R.E. Ellis and T.M. Peters (Eds.) MICCAI 2003*, LNCS 2879, pp. 184–191, 2003.

[11] T. Fawcett and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behavior", *Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, San Diego, USA, 1999.

[12] D. M. Gavrila, "The Visual Analysis of Human Movement: A survey", *Computer Vision and Image Understanding*, Academic Press, vol. 73, nr. 1,, pp. 82-98, 1999.

[13] C. Graetzel, T. Fong, S. Grange, and C. Baur, "A Non-Contact Mouse for Surgeon-Computer Interaction", *Technology and Health Care*, IOS Press, (in press) 2004.

[14] S. Grange, E. Casanova, T. Fong, and C. Baur, "Vision-based Sensor Fusion for Human-Computer Interaction", *International Conference on Intelligent Robots and Systems*, IEEE/RSJ, Lausanne, Switzerland, October 2002.

[15] M. A. Grasso, “The Long-Term Adoption of Speech Recognition in Medical Applications”, *Proc. of the 16th IEEE Symposium on Computer-Based Medical Systems (CBMS 2003)*, pp: 257-262, 2003.

[16] R. Hill and J. Begole, "Activity Rhythm Detection and Modeling", *Proc. of CHI 2003*, Ft. Lauderdale, Florida, USA, April 2003.

[17] N. Jojic, B. Brummit, B. Meyers, S. Harris, T. Huang, “Detection and Estimation of Pointing Gestures in Dense Disparity Maps”, *Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, p.468, March 26-30, 2000.

[18] V. Kaptelinin and B. A. Nardi, "Activity Theory: Basic Concepts and Applications", *CHI 97 Electronic Publications, Tutorials*, 1997.

[19] R. Kumar, G. D. Hager, A. Barnes, P. Jensen and R. H. Taylor, “An Augmentation System for Fine Manipulation”, *Proc. of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2000*, Pittsburgh, Pennsylvania, USA, October 11-14, 2000.

[20] S. Marcel, "Gestures for Multi-Modal Interfaces: A Review", *IDIAP Technical Report IDIAP-RR 02-34*, September 2002.

[21] S. Martelli, S. Bignozzi, M. Bontempi, S. Zaffagnini and L. Garcia, "Comparison of an Optical and a Mechanical Navigation System", *R.E. Ellis and T.M. Peters (Eds.): MICCAI 2003*, LNCS 2879, pp. 303–310, 2003.

[22] I. Mikic, K. Huang and M. Trivedi, “Activity Monitoring and Summarization for an Intelligent Meeting Room”, *IEEE Workshop on Humon Motion*, Austin, Texas, December 2000.

[23] T. B. Moeslund and E. Granum, “A Survey of Computer Vision-Based Human Motion Capture”, *Computer Vision and Image Understanding*, Volume 81 Issue 3, pp. 231 – 268, March 2001.

[24] A. B. Mor, J. E. Moody, D. Davidson, R. S. Labarca, B. Jaramaz, and A. M. Digioia III, "A Framework for Determining Component and Overall Accuracy for Computer Assisted Surgery Systems", *R.E. Ellis and T.M. Peters (Eds.), MICCAI 2003*, LNCS 2879, pp. 985−986, 2003.

[25] A. B. Nardi, *Context and Consciousness: Activity Theory and Human-Computer Interaction*, The MIT Press, Cambridge, Massachusetts, 1996.

[26] A. Nishikawa, S. Asano, R. Fujita, S. Yamaguchi, T. Yohda, F. Miyazaki, M. Sekimoto, M. Yasui, Y. Miyake, S. Takiguchi and M. Monden, "Selective Use of Face Gesture Interface and Instrument Tracking System for Control of a Robotic Laparoscope Positioner", *R.E. Ellis and T.M. Peters (Eds.) MICCAI 2003*, LNCS 2879, pp. 973–974, 2003.

[27] A. Nishikawa, T. Hosoi, K. Koara, D. Negoro, A. Hikita, S. Asano, H. Kakutani, F. Miyazaki, M. Sekimoto, M. Yasui, Y. Miyake, S. Takiguchi, and M. Monden, "FAce MOUSe: A Novel Human-Machine Interface for Controlling the Position of a Laparoscope", *IEEE Transactions on Robotics and Automation (Special Issue on Medical Robotics)*, 2003.

[28] N. Oliver, B. Rosario and A. Pentland, “A Bayesian Computer Vision System for Modeling Human Interactions”, *Proc. of Intl. Conference on Vision Systems ICVS99*. Gran Canaria, Spain, January 1999.

[29] P. Peixoto, J. Batista, and H. Araùjo, "Real-Time Human Activity Monitoring Exploring Multiple Vision Sensors", *Robotics and Autonomous Systems*, 2000.

[30] M. Porta, “Vision-based user interfaces: methods and applications”, *International Journal of Human-Computer Studies*, Vol. 57, pp. 27-73, Elsevier Science - Academic Press, 2002.

[31] D. H. Rice, S. D. Schaefer, "Endoscopic Paranasal Sinus Surgery", pp. 159-186, Raven Press, New York, 1993.

[32] J. Rickel and W. L. Johnson, “Animated Agents for Procedural Training in Virtual Reality, Perception, Cognition, and Motor Control”, *Applied Artificial Intelligence*, Volume 13, pp. 343-382, 1999.

[33] D. Sawyer, K. J. Aziz, C. L. Backinger, E. T. Beers, A. Lowery and S. M. Sykes, “An Introduction to Human Factors in Medical Devices”, US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health; 1996.

[34] M. Turk, “Perceptive Media: Machine Perception and Human Computer Interaction”, *Chinese Journal of Computers*, Vol. 23, No. 12, pp. 1235-1244, 2000.

[35] F. Vogt, S. Krüger, H. Niemann and C. Schick, "A System for Real-Time Endoscopic Image Enhancement", *R.E. Ellis and T.M. Peters (Eds.), MICCAI 2003*, LNCS 2879, pp. 356–363, 2003.

[36] E. Wahlstrom, O. Masoud, and N. Papanikolopoulos, "Vision-based Methods for Driver Monitoring", *Proc. IEEE 6th International Conference on Intelligent Transportation Systems*, pp. 903-908, Shanghai, China, Oct. 2003.

[37] M. E. Wicklund, “Making Medical Device Interfaces More User-Friendly”, *Medical Device and Diagnostic Industry*, 1998.