KUNGL
TEKNISKA
HÖGSKOLAN

CID-209 • ISSN 1403-0721 • Department of Numerical Analysis and Computer Science • KTH

Using Marking Menus to Develop Command Sets for Computer Vision Based Hand Gesture Interfaces

Sören Lenman, Lars Bretzner, Björn Thuresson
Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 239-242. Aarhus, Denmark 2002

CID, CENTRE FOR USER ORIENTED IT DESIGN

KUNGL
TEKNISKA
HÖGSKOLAN

CID-209 • ISSN 1403-0721 • Department of Numerical Analysis and Computer Science • KTH

Using Marking Menus to Develop Command Sets for Computer Vision Based Hand Gesture Interfaces

Sören Lenman, Lars Bretzner, Björn Thuresson
Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 239-242. Aarhus, Denmark 2002

CID, CENTRE FOR USER ORIENTED IT DESIGN

Sören Lenman, Lars Bretzner, Björn Thuresson

Using Marking Menus to Develop Command Sets for Computer Vision Based Hand Gesture Interfaces
Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 239-242
Report number: CID-209
ISSN number: ISSN 1403 - 0721 (print) 1403 - 073 X (Web/PDF)
Publication date: October 2002
E-mail of author: lenman@nada.kth.se

Reports can be ordered from:

CID, Centre for User Oriented IT Design
NADA, Deptartment of Numerical Analysis and Computer Science
KTH (Royal Institute of Technology)
SE- 100 44 Stockhom, Sweden
Telephone: + 46 (0)8 790 91 00
Fax: + 46 (0)8 790 90 99
E-mail: cid@nada.kth.se
URL: http://cid.nada.kth.se

Using Marking Menus to Develop Command Sets for Computer Vision Based Hand Gesture Interfaces

Sören Lenman
Centre for User-Oriented IT-Design
Royal Institute of Technology
100 44 Stockholm, Sweden
lenman@nada.kth.se

Lars Bretzner
Comp. Vision and Active Perc. Lab
Royal Institute of Technology
100 44 Stockholm, Sweden
bretzner@nada.kth.se

Björn Thuresson
Centre for User-Oriented IT-Design
Royal Institute of Technology
100 44 Stockholm, Sweden
thure@nada.kth.se

ABSTRACT
This paper presents the first stages of a project that studies the use of *hand gestures* for interaction, in an approach based on computer vision. A first prototype for exploring the use of marking menus for interaction has been built. The purpose is not menu-based interaction per se, but to study if marking menus, with practice, could support the development of autonomous command sets for gestural interaction. Some early observations are reported, mainly concerning problems with user fatigue and precision of gestures. Future work is discussed, such as introducing flow menus for reducing fatigue, and control menus for continuous control functions. The computer vision algorithms will also have to be developed further.

Keywords
Hand gesture, computer vision, HCI, gesture command, marking menu.

INTRODUCTION
This paper presents the first stages of a project that studies the use of *hand gestures* for interaction, in an approach based on computer vision. Remote control of electronic appliances in a home environment, such as TV sets and DVD players, has been chosen as a starting point. This is an existing, common interaction situation, familiar to most. Normally it requires the use of a number of devices, and there are clear benefits to an appliance-free approach. So far we have only implemented a first prototype for exploring pie- and marking menus [4], [8] for gesture-based interaction. The purpose is not menu-based interaction per se, but to study if marking menus, with practice, could support the development of autonomous command sets for gestural interaction, a generally overlooked area in gesture-based interaction.

Perceptive and Multimodal User Interfaces
Two main scenarios for gestural interfaces can be distinguished. One aims at developing *Perceptive User Interfaces* (PUI), as described by Turk [15], striving for automatic recognition of natural, human gestures integrated with other human expressions, such as body movements, gaze, facial expression, and speech. The aim is to develop conversational interfaces, based on what is considered to be natural human-to-human dialog. For example, Bolt [2] suggested that in order to realize conversational computer interfaces, gesture recognition will have to pick up on unintended gestures, and interpret fidgeting and other body language signs.

However, our current work falls within the second approach to gestural interfaces, *Multimodal User Interfaces,* where hand poses and specific gestures are used as commands in a command language. Here, the gestures need not be natural gestures but could be developed for the situation, or based on a standard sign language. In this approach, gestures are either a replacement for other interaction tools, such as remote controls and mice, or a complement, e.g., gestures used with speech and gaze input in a multimodal interface. Oviatt et al. [10] noted that there is a growing interest in designing multimodal interfaces that incorporate vision-based technologies. They also contrast the passive mode of PUI with the active input mode, addressed here, and claim that although passive modes may be less obtrusive, active modes generally are more reliable indicators of user intent, and not as prone to error.

An overview of computer vision based gesture recognition applications in HCI can be found in Lenman et al. [9]. Detailed descriptions and taxonomies concerning hand gestures from the point of view of computer vision can be found in, e.g., Pavlovic & Sharma [11].

Gestural Command Sets
With the exception of Baudel et al. [1], very little attention has been paid to the development of command sets for gesture-based interaction. The design space for such commands can be characterized along three dimensions: *Cognitive aspects, Articulatory aspects,* and *Technological aspects.*

Cognitive aspects refer to how easy commands are to learn and to remember. It is often claimed that gestural command sets should be natural and intuitive, e.g. [2] [16], mostly meaning that they should inherently make sense to the user. However, there might not exist any shared stereotypes to build on, except in very specific situations. If the aim is

Accepted for presentation at The Second Nordic Conference on Human-Computer Interaction, NordiCHI 2002, 19-23 October 2002, Aarhus, Denmark.

gestural control of devices, there is no cultural or other context for most functions.

Articulatory aspects refer to how easy gestures are to perform, and how tiring they are for the user. Gestures involving complicated hand or finger poses should be avoided, because they are difficult to articulate and might even be impossible to perform for a substantial part of the population. They are common in current computer based approaches, because they are easy to recognize by computer vision. Repetitive gestures that require the arm to be held up and moved without support are also unsuitable from an articulatory point of view because of fatigue.

Technological aspects refer to the fact that in order to be appropriate for practical use, and not only in visionary scenarios and controlled laboratory situations, a command set for gestural interaction based on computer vision must take into account the state-of-the art of technology, now and in the near future. For example, Sign Language recognition might be desirable for a number of reasons, not least for people who need to use Sign Language for communication. This is currently far from feasible. Still, much work can be done with reduced sets of Sign Language, e.g., Starner et al. [13], as a first step towards a long-term objective.

CURRENT WORK

The point of departure for our current work is *cognitive*, leaving articulatory aspects aside at the moment for reasons of technical feasibility. A command language based on a menu structure has the cognitive advantage that commands can be recognized rather than recalled. Traditional menu-based interaction is not attractive in a gesture-based scenario, however. Menu navigation is far from the directness that gestural interaction could provide. However, pie- and marking menus might provide a foundation for developing directness and autonomous gestural command sets.

Pie- and Marking Menus

Pie menus were first described by Callahan et al. [4]. They are pop-up menus with the alternatives arranged radially. Because the gesture to select an item is directional, users can learn to make selections without looking at the menu. In principle this could be learned also with linear menus, but it is much easier to move the hand without feedback in a given direction, as with a pie menu, than to a menu item at a given distance, as in a linear menu. This fact can support a smooth transition between novice and expert use. For an expert user, working at high speed, menus need not even be popped up. The direction of the gesture is sufficient to recognize the selection. If the user hesitates at some point in the interaction, the underlying menus can be popped up, always giving the opportunity to get feedback about the current selection. Hierarchic marking menus [8] is a development of pie menus that allow more complex choices by the use of sub-menus. The same principles apply: expert users could work without feedback. The shape of the gesture (mark) with its movements and turns can be recognized as a selection, instead of the sequence of distinct choices between alternatives.

Hierarchic Marking Menus for Gesture-Based Interaction: In the work presented here the assumption is that autonomous command sets for computer vision based gesture interfaces can be created from hierarchical marking menus. As to articulatory characteristics, a certain hand pose, e.g., holding the hand up with all fingers outstretched, could be used for initiating a gesture and activating the menu system. This would correspond to the pen-down event in a pen-based system. The gesture could then be tracked by the computer vision algorithms, as the hand traverses the menu hierarchy. Finally, a certain hand pose could be used to actually make the selection, e.g., the index finger and thumb outstretched, corresponding to a pen-up event in pen-based interface. Put differently, the gestures in the command set would consist of *a start pose*, a *trajectory*, defined by menu organization, for each possible selection, and, lastly, a *selection pose.* Gestures ending in any other way than with the selection pose would be discarded, because either they could mean that the user abandoned the gesture, or simply that tracking of the hand was lost.

For a novice user, this would amount to a traditional menu-selection task, where selections are made by navigating through an hierarchical menu structure. This, as such, could provide for unencumbered interaction in remote control situations but, as noted above, is the directness of a gesture-interface would be lost. The assumption here, however, is that over time users will learn the gesture corresponding to each selection and no longer need visual feedback. The interaction would develop into direct communication, using a gestural language. In addition to providing for a natural transition from novice to expert, such a gestural language makes no assumptions about naturalness or semantics of gestures, because it is defined by the menu structure. In principle, if not in practice, the command set is unlimited. A further advantage is that the demands put on the computer vision algorithms are reasonable. Very fast and stable tracking of the hand will be required, however.

A PROTOTYPE FOR HAND GESTURE INTERACTION

The prototyping and experimental work is still in an early stage and only a brief overview and some early impressions can be given here. Inspired by Freeman et al. [6], we chose remote control of appliances in a domestic environment as our first application. However, so far, we have only designed a first example of a hierarchic menu system for controlling some functions of a TV, a CD player, and a lamp.

The Computer Vision System

We have chosen a view-based representation of the hand, including both color and shape cues. The system tracks and recognizes the hand poses based on a combination of multi-scale color feature detection, view-based hierarchical hand models and particle filtering. The hand poses, or hand states, are represented in terms of hierarchies of color image features at different scales, with qualitative interrelations in terms of scale, position and orientation. These hierarchical models capture the coarse shape of the hand

poses. In each image, detection of multi-scale color features is performed. The hand states are then simultaneously detected and tracked using particle filtering, with an extension of layered sampling referred to as hierarchical layered sampling. The particle filtering allows for the evaluation of multiple hypotheses about the hand position, state, orientation and scale, and a likelihood measure determines what hypothesis to choose. To improve the performance of the system, a prior on skin color is included in the particle filtering step. In fig. 1, yellow (white) ellipses show detected multi-scale features in a complex scene and the correctly detected and recognized hand pose is superimposed in red (gray). A detailed description of the algorithms is given in [3].

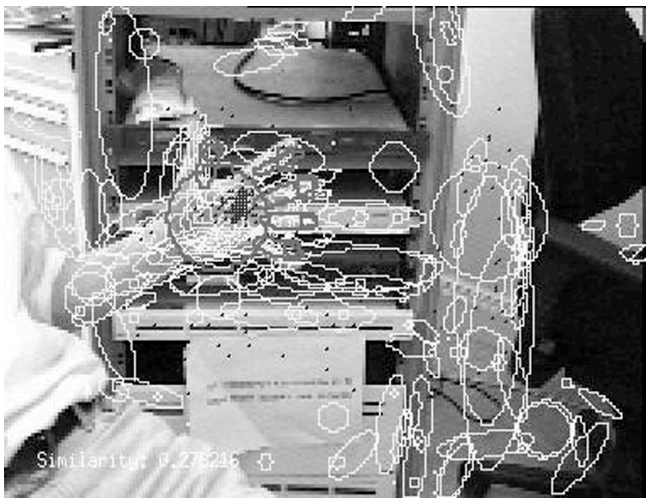

Fig. 1 Detected multi-scale features and the recognized hand pose superimposed in an image of a complex scene.

As the coarse shape of the hand is represented in the feature hierarchy, the system is able to reject other skin colored objects that can be expected in the image (the face, arm, etc). The hierarchical representation can easily be further extended to achieve higher discrimination to complex backgrounds, at the cost of a higher computational complexity. An advantage of the approach is that it is to a large extent user and scale (distance) invariant. To some extent, the chosen qualitative feature hierarchy also shows view invariance for rotations out of the image plane (up to approx. 20-30 degrees for the chosen gestures).

There is a large number of works on real-time hand pose recognition in the computer vision literature. Some of the most related to our approach are, e.g., Freeman and Weissman [6] (see above) who used normalized correlation of template images of hands for hand pose recognition. Though efficient, this technique can be expected to be more sensitive to different users, deformations of the pose and changes in view, scale, and background. Cui and Weng [5] showed promising results for hand pose recognition using an appearance based method. However, the performance was far from real-time. The approach closest to ours was presented by Triesch and von der Malsburg [14] representing the poses as elastic graphs with local jets of Gabor filters computed at each vertex.

In order to maximize speed and accuracy in the prototype, gesture recognition is currently tuned to work against a uniform background within a limited area, approximately 0,5 by 0,65 m in size, at a distance of approximately 3 m from the camera, and under relatively fixed lighting conditions.

Menu System

A menu system with three hierarchical levels and four choices in each menu currently exists. Only a few of choices are active, however: *TV On/Off*, *Previous/Next channel*, *CD Play/Stop/Back/Forward*, *Lamp On/Off*. A hand pose with the index finger and thumb outstretched is used as the start pose for activating the menus, corresponding to pen-down in a pen-based interface. A hand with five fingers outstretched is used as the selection pose, corresponding to pen-up. Evidently, any two hand poses could be used for these purposes. Menus are activated when the start hand pose is detected by the computer vision system in the active area. The hand is tracked as long as the start pose is held. If the hand is moved over the periphery of a sector that has a submenu, the parent menu disappears, and the submenu appears. Showing the selection pose in an active field, e.g., *TV on*, makes a selection. All other ways of ending the interaction are ignored. The menus are currently shown on a computer screen, placed by the side of the TV (fig. 2). This is inconvenient, and in the future menus will be presented in an overlay on the TV screen.

Fig. 2 The demo space at CID.

Results and Discussion

Only a limited number of informal user trials have been performed so far. We have not yet been able to bring the technical performance of the system to a level where true gesture-based control without feedback can be accomplished. However, observations with the current system indicate that gesture-based control without feedback is feasible with single-level pie menus, but that gestures based on hierarchical menus create some problems. It is difficult to perform gestures sufficiently distinct, relying only on feedback from the proprioceptive system of the arm. Thus, in order to support development of autonomous command

sets, computer algorithms for recognition of fuzzy gestures might be required.

The current setup, with subjects seated facing the TV and making gestures with one arm and hand held out by the side of the body without support, is not suitable from an articulatory point of view. It is inconvenient and fatigue quickly sets in. The problem of fatigue is known from earlier attempts with gesture-based interfaces and must be addressed. In the current application much could be gained by providing support for the arm, by making gestures smaller, and by making the recognition system more tolerant as to the whereabouts of the user and the hand.

FUTURE WORK

As to the computer vision algorithms, there is ongoing work to increase the speed and tracking stability of the system, to acquire more position independence for recognition of gestures, to increase the tolerance for varying lighting conditions, and to increase recognition performance with complex backgrounds. The main effort, however, is currently aimed at the design and organization of menus. We are experimenting with reducing the number of choices at each level in the menu structure, i.e., trade breadth for depth, in order to reduce the demands on precision of gestures. Recently we have begun development of flow menus, a version of hierarchical marking menus in which successive levels of the hierarchy are shown in the same position [7]. This would greatly reduce the area which the gestures have to cover when the hierarchy is deep and thus diminish the problems of fatigue. We are also planning to introduce control menus [12] for continuous functions. With control menus, repeated control signals are sent as long as the hand is kept in a selection pose by the end of a gesture.

We are also considering a different scenario in which a few gestures (hand poses or pointing gestures) are used for direct control of common functions, such as the sound level or lighting, and gestures based on hierarchical marking menu structures are used for more complex selections.

ACKNOWLEDGMENTS

We thank Björn Eiderbäck, CID, who performed the Smalltalk programming for the menu system. Olle Sundblad, CID, did the Java programming for the application control server.

REFERENCES

[1] Baudel, T. & Beaudouin-Lafon, M. (1993) Charade: Remote Control of Objects using FreeHand Gestures. *Communications of the ACM*, vol 36, no. 7, pp. 28-35.

[2] Bolt, R.A. (1980) Put-that-there: Voice and Gesture in the graphics interface. *Computer Graphics*, 14(3), pp. 262-270.

[3] Bretzner, L., Laptev, I., & Lindeberg, T. (2002) Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. Paper presented at the *5th International Conference on Automatic Face and Gesture Recognition*, Washington, D.C:, May 2002.

[4] Callahan, J., Hopkins, D, Weiser, M. & Shneiderman, B. (1988) An Empirical Comparision of Pie vs. Linear Menus. *Proceedings of CHI'88*, pp. 95-100.

[5] Cui, Y. and Weng, J. (1996) Hand sign recognition from intensity image sequences with complex background. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 88-93.

[6] Freeman, W.T. & Weissman, C.D. (1994) Television Control by Hand Gestures. In *1st Intl. Conf. on Automatic Face and Gesture Recognition*.

[7] Guimbretière, F. & Winograd, T. (2000) FlowMenu: combining Command, Text and Data Entry. *Proceedings of UIST'2000*, pp. 213-216.

[8] Kurtenbach, G. & Buxton, W. (1994) The Limits of Expert Performance Using Hierarchic Marking Menus. *Proceedings of CHI'94*, pp. 482-487.

[9] Lenman, S., Bretzner, L. & Thuresson, B. Computer Vision Based Recognition of Hand Gestures for Human-Computer Interaction. Technical report TRITA-NA-D0209, CID-report, June 2002

[10] Oviatt,, S., Cohen, P, Wu, L. Vergo, J, Duncan, L. Suhm, B, Bers, J, Holzman, T, Winograd, T. Landay, J. Larson, J. Ferro, D. (2000) Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. *Human-Computer Interaction*, Vol 15, pp. 263-322

[11] Pavlovic, V.I., Sharma, R. & Huang, T.S. (1997) Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 19 (7) 677-695.

[12] Pook, S., Lecolinet, E., Vaysseix, G. & Barillot, E. (2000) Control Menus: Execution and Control in a Single Interactor. In *Extended Abstracts of CHI2000*, pp. 263-264.

[13] Starner, T, Weaver, J. & Pentland, A. (1998) Real-time American Sign Language recognition using desk and wearable computer-based video. *IEEE Transactions on Pattern. Analysis and Machine. Intelligence*.

[14] J. Triesch and C. von der Malsburg. (2001) A system for person-independent hand posture recognition against complex backgrounds. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 23, no 12.

[15] Turk, M. & Robertson, G. (2000) Perceptual User Interfaces. *Communications of the ACM*, vol. 43, no. 3, pp. 33-34

[16] Wexelblatt, A. (1995) An Approach to Natural Gesture in Virtual Environments. *ACM ToCHI*, Vol 2., No. 3, September, pp. 179-200.