Author :
Frédérick Gianni, Patrice Dalle
IRIT-TCI gianni@irit.fr, dalle@irit.fr

Key-Words:
Gestures Interaction, Gestures Interpretation, Computer Vision,

Visuo-Gestural Interaction with a video wall

This article introduce a design methodology for free-hand gestures based interactions in an environment composed of a wide diversity of information sources and a video-wall. This article is made of three parts. It defines in the first part the characteristics and singularities of our experiment. In the second part, it describes the methodology used to create the gesture language. Finally, it presents the current image processing results. We conclude on the evaluation perspectives for this interface.

1. Introduction

Man machine interaction is consider as a major problem by the computer-vision community, particularly with the aim at driving larges screens with gestures. One of those stakes in this context is to elaborate solutions at different levels from the language used to the video processing to implement. We will first present the conception of the gestural language, thanks to two wizards of Oz experienced on several persons; it is a mean to get unconstrained gestures realized in different ways. Then we will define the framework of use, whom implies some operating conditions. Computer engineering will next be detailed to explain the tracking of the speaker and the interpretation of his gestures. We will point the use of several model that allow us to retrieve relevant informations in order to interpret gestures with a monocular vision system. At last we will show the provided solutions to interpret the gestures of the command language defined.

2. Framework of use and constraints

2.1 The environment

The studied environment is characterized by a collective display system, or a video wall on which several sources of informations coming from computers of the room can be simultaneously displayed. The vision system used is based on a colour camera composed of three CCD. The purpose of the interaction proposed is to organise the displaying surfaces the room: the video wall and the different spatially distributed computers (see illustration 1) . This room can be employed for presentations of work, meeting of conception or as a crisis room, where informations must be able to be shown as fast as possible. Those cases are the motivations of use of gestures as commands. Those contexts of uses define the operating conditions and the constraints imposed by the vision system.

Ill.1: the room environment

2.2Constraints

One of the problem when designing a Visuo-gestural interaction system is the precise definition of the contextual constraints to which it will be subjected. These constraints will strongly influence the performances of the system and the definition of the gestural expressions. We try here to synthesize the

principal criteria which make possible to define the framework of a Visuo-gestural interaction system. For each criterion we illustrate our positioning. These criteria were established by extending the work of (Moeslund 2001, 231-268) and (Krahnstoever 2002, 203).

*Real Time – Interactivity:*the algorithms of tracking and image processing must allow an interaction whose reactivity does not exceed the second. This constraint will exploit the speed of execution and the complexity of the gestures.

Occlusions: when the hands cross or pass behind the body. We decided to use interactions without hand occlusion, as far as it is possible.

Environment variations: the environment can often be reconfigured (furnitures can be moved) and because of the presence of pits of natural light the luminosity can vary according to the period of the day and the year.

Image resolution: the needed resolution of the image depends on the position of the user in the room and depends of the size of the parts of the body performing the gesture. Here, the user is free to move in the space. However the different configuration of the hand are not considered in the current system, the hand could be too tiny.

Physical characteristics of the user: user clothes should have a different colour than his skin, and may have long sleeves. But we do not pose constraints on the appearance (skin colour, height, pilosity) of the user. We must be able to adapt to different contexts.

Sequence of gestures: in our case interactions are drowned in the context, the user will perform orders in an isolated and specific way but during other parallel and parasitic actions.

Segmentation: from the flow of gestures sequences, it is necessary to determine the markers of commands gestures to allow a temporal segmentation. We have, for the moment, considered several solutions which are currently in validation (gaze tracking, orientation, use of key gestures, immobilization or temporal marker, multi-modal or vocal markers).

Initialization and calibration: the system must be automatically initialized and does not have to be user dependent.

Multiple access: mono-user system or asynchronous multi-user.

General interaction features: dynamics gestures evolving in time, the user can use his whole body (tighten and bend the arms), interactions with two hands possible.

3.Language conception

A difficulty with which a gesture interaction system is confronted to is the definition of a gestures vocabulary which should be sufficiently intuitive to allow a fast training and sufficiently discriminative for a possible interpretation without confusions. As underlined by (Moeslund 2001, 231-268), it is difficult to create metaphors of adequate gestures. Majority of experiments impose the vocabulary to their users.

On the contrary we have decided to adopt a process taking into account the users in the specification of those commands gestures. The definition of the language was done in three stages. Initially we have defined a list of five commands which the user will be able to perform: display, move, remove, resize, zoom. This is the definition of the statement of the command.
In a second time we sought to know how users would translate these statements into gestures. In this way we organized an experiment in the form of wizard of oz. We invited eight voluntary users to perform each one the same scenario using all the orders, in various contexts, without constraints on the realization of the gestures. The effects of the gestural orders were simulated by an operator present in the room. Each experimentation was recorded by a contextual camera and a camera focused on the user. We could then evaluate the realisation and choose the gestures according to their intuitiveness (the most used), complexness (simplest) and singularity (uniqueness). Our estimate of the complexity and the discriminants of the gestures observed was done on partly empirical criteria. We thus had a generic corpus of gestures corresponding to the commands defined.
In the third and last step we wanted to evaluate the intuitiveness of our corpus and to note possible interpretations of formal descriptions of the gestures. We organized a new wizard of oz around two new scenarios of presentation implementing in particular these gestures in extreme situations. Nine new users took part in this meeting and nine others observed and evaluated the use of the corpus. From this new session we refined the corpus, starting from the study of the performed gestures and the parasitic gestures. Here is the list of the commands and their realisations:

- display: point a computer, point a part of the video wall, hand return back to a rest position
- move: point a window, move in the area of the video wall, hand return back to a rest position
- remove point a window, move it out off the video wall bounds, hand return back to a rest position
- resize: point a window, indicate the scale ratio with two hands (distance between the two hands), hands return back to a rest position
- zoom: point a window, moving the hand closer to the body to perform a zoom in, further for a zoom out, hand return back to a rest position

We defined three separates cases of commands production. The first one is the case where commands are emitted one at a time we will call this “isolated command”. In this situation the command statement is composed of three part :

- the pre-command, where the hand is in a resting position and start to move to indicate the object on which the command will be apply to
- the command, which is the action to perform
- the post-command, where the hand move back to its rest position.

The second one is where several commands are issue with the same subject of command (understand a window on which the command will be applied to), here the command will has the form: the pre-command, the command, another command, ..., post-command. The last one is where multiple commands are perform on several subject, the command statement will be: pre-command, command, pre-command, command, ..., post-command. This two last cases will be call “chained commands”.

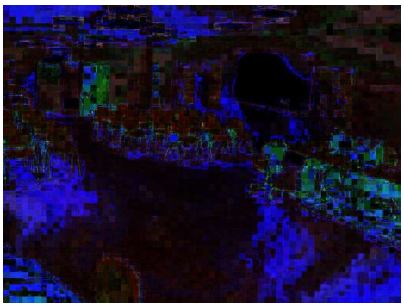
4.The vision system

From the constraints identified and the complexity of the vocabulary, the operators of image processing can be implemented. We need to use several model as *a priori* knowledge, those models are: the background model of the scene, the skin colour model of the user and an anthropometric model of the

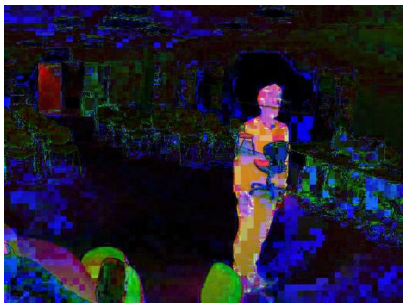
user. We will see now the whole sequence of operators used for user detection, segmentation, tracking and gestures interpretation.

4.1Initialisation

The purpose of our first operator is to separate the background and the foreground from an image sequence produced by a fixed camera. For this, we model each pixels of a background image sequence, in a statistical manner, in the HSV (hue, saturation, value) colour space in order to learn its luminosity and chromatic variations (illustration 2). After applying an adaptive threshold, we can retrieve the silhouette of the user (see illustration 3, 4). Using only the chrominance information we can eliminate the silhouette shadows. This background model can be calculated at any time and so allow adaptation to the variations of the environment. As specified above, this step permit us to free ourselves from the obligation of a uniform background (Wren 1997, 780-785). Other methods such as modelling background by mixtures of Gaussian (Stauffer 1999, 246-252) shows certain limitations: the number of Gaussian is arbitrarily set and their initialization remains a problem. Han (Han 2004) proposes, always with the aim of finding a person, to use the *mean-shift* algorithm but this one is limited by its execution time and its memory occupation even in its optimized version.



Ill. 2: Background model



Ill. 3: pixels detected as foreground (in red and green)

Ill. 4: after adaptive threshold

Next we locate the head of the user, from his silhouette we are looking for skin coloured pixels. We use a probabilistic skin model to segment skin tones pixels. It has been shown (McKenna 1997, 140-151) that human skin tones can be characterised by a multivariate normal distribution in the HSV colour space. Providing a human skin image we estimate the distribution in a parametric form as a gaussian model to obtain the skin colour model . Using this model we compute the likelihood of any pixel of the silhouette belonging to the skin model (see illustration 5). We then select the most probable pixels and regroup them in area by connexity. The biggest area will be identified as the head of the user (see illustration 6). The areas of the left and right hand are identified in the same way we just use one more model: an anthropometric model in order to validate the position of the hands relatively to the head . This last model permit us to compute the expected size of the hands and define the areas of search around the head in the silhouette (see illustration 7).

Ill. 5:skin pixels in the foreground

Ill. 6: Detection of the user head

Ill. 7: Detection and identification of the right hand

4.2 Hands tracking

At time t a new image arrive, we already have detected and identified the head, the left and right hands of the user. We use the position of the old bounding box, at time t-1, to re-detect the hand and update its trajectory. It can happen that more than one skin coloured area can be found in this bounding box, we select the one minimising the variations of size, of direction, of speed. In a case where the hand move between the video camera and the hand of the user, a merge situation arise which will be followed by a split situation: the pixels heap representing the hand will merge with the pixels heap representing the head and then the hand heap will split from the the head heap.

4.3 Gestures interpretation

Our method for interpreting the gesture rely on the spatial nature of our command language. During a phase of initialisation the user has to point the corner of the objects of interest out, see illustration 8. The objects of interest are the computers providing the informations to be displayed and the video wall. A command is interpreted if the corresponding sequences of action as been performed. For instance the display command is realised once a hand has been detected in the area of a computer and then in the area of the video wall, with a smooth trajectory from a computer to the video screen (see illustration 9). In order to have a good differentiation between the co-verbal gestures produced during a work presentation and the command gestures, the user have to use some temporal markers: each time a designation is produce he has to hold is gesture in the area of the object a few seconds.

Ill. 8: Initialisation, positioning of the video wall in the user space

Ill. 9: interpretation of the command display, the hand has been detected in the area of a computer (in purple), and is now detected in the right bottom part of the video wall (in blue)

5. Conclusion

In this paper we have presented a process of conception of an ambitious system of visuo-gestural interaction. We have proposed a set of criteria which permit to define the constraints and singularity of a such system compared to existing research. We also have presented the different elaboration steps of an intuitive gestural language followed by the presentation of the image processing operators. These operators have been selected on their capacity to extract the data respectively to the temporal constraints needed for the interaction. An architecture permitting to add in a simple way some other image processing operators is in conception. This will allowed some evolution in the image processing system in order to possibly modify the command language and the interactive context. The next step in this study is to validate the system using the chained commands situation. It will be interesting in particular to study the evolution of the realisation of the commands gestures during the use.

Bibliography

Krahnstoever, N. Schapira, E. Kettebekov, S. Sharma, R. (2002). Multimodal Human Computer Interaction for Crisis Management Systems, *IEEE Workshop on Applications of Computer Vision* (pp. 203).

McKenna, S. Gong, G. and Raja, Y. (1997) Faces recognition in dynamic scenes. In Clark A F, ed. *British Machine Vision Conference,* (pp140-151).

Moeslund,T.B. and Granum,E. (2001). A survey of computer vision-based human motion capture. In *Computer Vision and Image Understanding*, 81(3):231-268.

Han, B. Comaniciu, D. et Davis, L. (2004). Sequential kernel density approximation through mode propagation : application to background modelling, Proc. *ACCV - Asian Conf. on Computer Vision*.

Stauffer, C. Grimson, W.E.L. (1999). Adaptive background mixture models for real-time tracking, Proc. *of Computer Vision Pattern Recognition,* (pp. 246-252).

Wren, C. Azarbayejani, A. Darrell, T. and Pentland, A. (1997). Pfinder:Real-time Tracking of the Human Body. *IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7):780-785.*