A NOVEL CAMERA-BASED SYSTEM FOR COLLABORATIVE INTERACTION WITH MULTI-DIMENSIONAL DATA MODELS

M. Van den Bergh,
Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland;
vamichae@vision.ee.ethz.ch

J. Halatsch,
Chair of Information Architecture, ETH Zurich, Zurich, Switzerland;
halatsch@arch.ethz.ch

A. Kunze,
Chair of Information Architecture, ETH Zurich, Zurich, Switzerland;
kunze@arch.ethz.ch

F. Bosché, PhD,
Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland;
bosche@vision.ee.ethz.ch

L. Van Gool, Prof.,
Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland;
vangool@vision.ee.ethz.ch

G. Schmitt, Prof.,
Chair of Information Architecture, ETH Zurich, Zurich, Switzerland;
schmitt@ia.arch.ethz.ch

ABSTRACT: In this paper, we address the problem of effective visualization of and interaction with multiple and multi-dimensional data supporting communication between project stakeholders in an information cave. More exactly, our goal is to enable multiple users to interact with multiple screens from any location in an information cave. We present here our latest advancements in developing a novel human-computer interaction system that is specifically targeted towards room setups with physically spread sets of screens. Our system consists of a set of video cameras overseeing the room, and of which the signals are processed in real-time to detect and track the participants, their poses and hand-gestures. The system is fed with camera based gesture recognition. Early experiments have been conducted in the Value Lab (see figure 1), that has been recently introduced at ETH Zurich, and they focus on enabling the interaction with large urban 3D models being developed for the design and simulation of future cities. For the moment, experiments consider only the interaction of a single user with multiple layers (points of view) of a large city model displayed on multiple screens. The results demonstrate the huge potential of the system, and the principle of vision based interaction for such environments. The work continues on the extension of the system to a multi-user level.

KEYWORDS: Information Cave, Interaction, Vision, Camera..

1. INTRODUCTION

1.1 Product Information Models for Design and Simulations

Future cities, standing for evolving medium-size and mega-cities, have to be understood as a dynamic system – a network that bridges different scales, such as local, regional, and global scales. Since such a network comprises several dimensions, for example social, cultural, and economic dimensions it is necessary to connect active research, project management, urban planning as well as communication with the public to establish a mutual vision, or to map the desires of the involved participants.

In the last few decades, the use of computers, software and digital models has expanded within many fields related to the Architecture, Engineering, Construction and Facility Management (AECFM), where Facility may refer to commercial, industrial or infrastructure building assets, and also cities. However, it is only recently that researchers have started tackling the problems of the compartmentalization of this expansion within these different fields corresponding to the multiple stakeholders of such project. And, this expansion occurred without wider project integration. For example, in urban planning, multiple different digital models are often used to perform different analyses such as: CO_2 emissions, energy consumption and traffic load. Nonetheless, significant progresses have recently been made in the integration of information models into what are now commonly referred to *Building Information Models (BIM)*, *City Information Models (CIM)*, *etc*.

These integrated models enable earlier and more systematic (sometimes automated) detection of conflicts different multiple analysis and processes. However, the resolution of these conflicts still requires human negotiations, and effective methods and technologies for interacting collaboratively with the information in order to resolve detected conflicts are still missing. The main complexity here is that large projects, such as large scale planning projects, require the involvement of many technical experts and other stakeholders (*e.g.* owners, pubic) who approach projects from many different view points, which results in many different types of conflicts that must resolve collaboratively.

In order to address this problem, holistic participative planning paradigms (governing process management, content creation as well as design evaluation) have to evolve, and consider new software and hardware solutions that will enable the different stakeholders to effectively work collaboratively.

1.2 Example: Dübendorf Urban Planning Project

Today's urban planning and urban design rely mainly on static representations (*e.g.* key visuals, 3D models). Since the planning context and its data (for example scenario simulations) are dynamic, visual representations need to be dynamic and interactive too, resulting in the need for physical environments enabling such dynamic processes.

During spring semester 2009 students researched how to establish design proposals in a more collaborative manner. The focus was on an urban planning project, the rehabilitation of the abandoned Swiss military airport in Dübendorf. The main goal of this research project was to develop an interactive shape grammar model (Müller, 2009), which was implemented with the CityEngine (http://www.procedural.com/cityengine). In combination with real-time visualization using Autodesk Showcase (http://usa.autodesk.com/adsk/servlet/index?id=6848305&siteID=123112), a better understanding design interventions was achieved.

While this research project showed the feasibility of collaborative interactive design, the experiments, then conducted in the Value Lab (see Section 2) showed that the interactivity offered by such information caves did not always meet the expectations of the users (see analysis in Section 2).

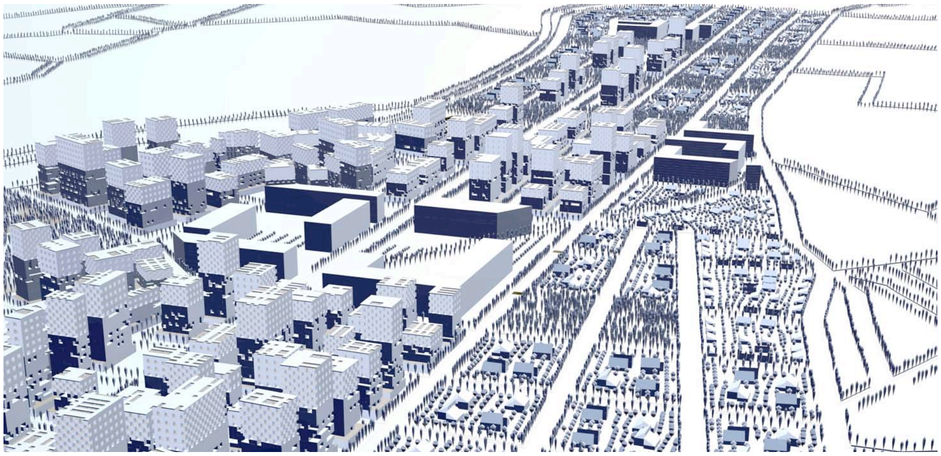


FIG. 1: As a result of collaborative city design workshops a new use for an abandoned military airport in the outskirts of Zurich had been implemented with the collaborative interaction tools that are available inside the Value Lab.

1.3 Information Visualization Caves

Information visualization caves have been investigated in order to enable stakeholders to sit in a single room and collaboratively solve conflicts, during planning, construction or operation. Such caves are typically designed with complex multimedia settings to enable participants to visualize the project model, and the possible conflicts at hand, from multiple points of view simultaneously (*e.g.* owner vs. user vs. contractor, contractor A vs. contractor B).

Traditional human-computer interaction devices (*e.g.* mouse, keyboard) are typically focused to fulfill single user requirements, and are not adapted to work with the multiplicity of participants and the multi-dimensionality (as well as multiplicity) of the data sets representing large projects. Solutions have however been proposed to improve interactivity. A multi-screen setup can drastically enhance collaboration and participatory processes by keeping information present to all attendees, and such setup is common in information caves (Gross et al., 2003, König et al., 2007). Additionally, (multi-) touch screens are now available as more intuitive multi-user human-computer interaction devices. However, despite their definite advantages for interactions with multiple users, particularly in table settings, multi-touch screens remain inadequate for use in rooms with physically spread sets of screens, as they require the users to constantly move from a screen to the other.

2. VALUE LAB

The ETH Value Lab (see Figure 2) is a special kind of information visualization room, and was designed as a research platform to guide and visualize long-term planning processes while intensifying the focus on the optimization of buildings and infrastructures through new concepts, new technologies and new social behaviors to cut down CO2 emissions, energy consumption, traffic load, and to increase the quality of life in urban environments (Halatsch and Kunze, 2007). It helps researchers and planners to combine existing realities with planned propositions, and overcome the multiplicity (GIS, BIM, CAD) and multi-dimensionality of the data sets representing urban environments (Halatsch et al. 2008a and 2008b).

The Value Lab consists of a physical space with state-of-the art hardware (supercomputer), software (*e.g.* urban simulation and CAD/BIM/GIS data visualization packages) and intuitive human-computer interaction devices. The interface consists of several high-resolution large area displays including:

1. *Five large screens* with a total of 16 mega pixels and equipped with touch interface capabilities; and
2. *Three FullHD projectors*. Two projectors form a concatenated high-resolution projection display with 4 Megapixel in resolution. That particular configuration is for example used for real-time landscape visualization. The third projector delivers associated views for videoconferencing, presentation and screen sharing.

The computing resources, display and interaction system produces a tremendous amount of possible configurations especially in combination with the connected computing resources. The system manages all computing resources, operation systems, displays, inputs, storage and backup functionality in the background as well as lighting conditions and different ad hoc user modes.

As a result, The Value Lab forms the basis for knowledge discovery and representation of potential transformations of the urban environment, using time-based scenario planning techniques in order to test the impact of varying parameters on the constitution of cities. It shows how the combination of concepts for hardware, software and interaction can help to manage digital assets and simulation feedback as well as promoting visual insights from urban planners to associated stakeholders in a human-friendly computer environment (Fox, 2000).

However, as discussed earlier, we found out that beside the direct on-screen manipulation of information, a technology was needed to steer larger moderated audiences inside a project, and that offers a more integrated navigation and usability behavior as well as permitting a wider overview on the main contents presented.

Therefore we are investigating a novel touch-less interaction system with camera-based gesture recognition. This system is presented below and early experimental results are presented in Section 4.

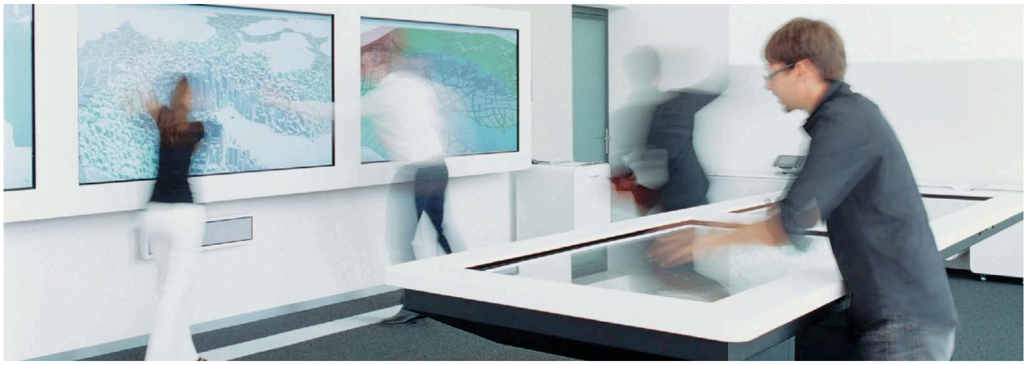


FIG. 2: The Value Lab represents the interface to advanced city simulation techniques and acts as the front-end of the ETH Simulation Platform.

3. VISION SYSTEM

In this section, we describe our vision system, that detects and tracks hand gestures of a user in front of a camera mounted on top of a screen as shown in figure 3. The goal of the system is to enable the interaction of the person with a 3D model.

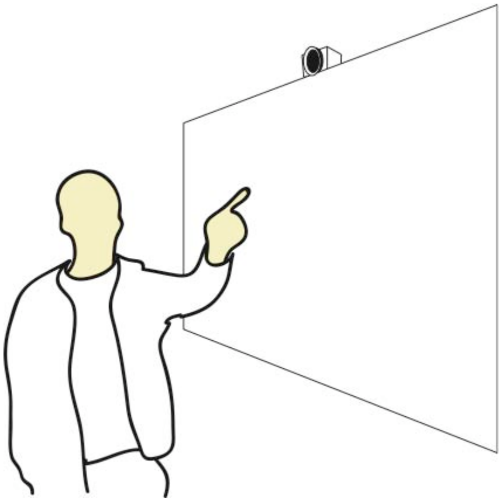


FIG. 3: Person interacting with a camera and screen.

3.1 Skin Color Segmentation

The hands of the user are located using skin color segmentation. The system is hybrid, combining two skin color segmentation methods. The first is a histogram-based method, which can be trained online, while the system is running. The advantage of this system is that it can be adapted in real-time to changes in illumination and to the person using the system. The second method is trained in advance with a Gaussian mixture model (GMM). The benefit of the offline trained system is that it can be trained with much more training data and is more robust. However, it is not robust to changes in illumination or to changes in the user.

3.1.1 Online model

Every color can be represented as a point in a color space. A recent study (Schmugge *et al*., 2007) tested different color spaces, and concluded that the HSI (hue, saturation and intensity) color space provides the highest performance for a three dimensional color space, in combination with a histogram-based classifier.

A nice characteristic is that the histograms can be updated online, while the system is running. Two histograms are kept, one for the skin pixel color distribution (H_{skin}), and one for the non-skin colors ($H_{non\text{-}skin}$). For each frame in the incoming video stream, the face region is found using a face detector such as the one in OpenCV (http://opencvlibrary.sourceforge.net/), and the pixels inside the face region are used to update H_{skin}. Then, the skin color detection algorithm is run and finds the face regions as well as other skin regions such as the hands and arms. The pixels which are not classified as skin are then used to update $H_{non\text{-}skin}$.

3.1.2 Offline model

In the GMM-based approach, the pixels are transformed to the *rg* color space. A GMM is fitted to the distribution of the training skin color pixels using the expectation maximization algorithm as described in (Jedynak *et al*., 2002). Based on the GMM, the probabilities P(*skin*|color) can be computed offline, and stored in a lookup table.

3.1.3 Post processing

On the one hand, the histogram-based method performs rather well at detecting the skin color pixels under varying lighting conditions. However, as it bases its classification on very little input data, it has a lot of false positives. On the other hand, the GMM-based method performs well in constrained lighting conditions. Under varying lighting conditions it tends to falsely detect white and beige regions in the background. By combining the results of the histogram-based and the GMM-based methods, many false positives can be eliminated. The resulting segmentation is improved further in additional post processing steps, which include median filtering and connected components analysis.

3.2 Hand Gesture Recognition

The hand gesture recognition algorithm is based on the full body pose recognition system using 2D *Haarlets* described in (Van den Bergh *et al*., 2009). Instead of using silhouettes of a person as input for the classifier, hand images are used.

3.2.1 Classifier input

The hands are located using the skin color segmentation algorithm described in section 4.1. A cropped grayscale image of the hand is extracted, as well as a segmented silhouette, which are then concatenated into one input sample, as shown in figure 4. The benefit of using the cropped image without segmentation, as shown on the right, is that it is very robust for noisy segmentations. Using the silhouette based on skin color segmentation only, as shown on the left, the background influence is eliminated. Using the concatenation of both gives us the benefit of both input sample options.

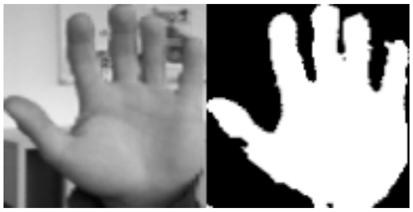


FIG. 4: Example of an input sample.

3.2.2 Haarlet-based classifier

For details about the classifier we refer to (Van den Bergh *et al*., 2009). It is based on an average neighborhood margin maximization (ANMM) transformation **T**, which projects the input samples to a lower dimensional space, as shown in figure 5. This transformation is approximated using *Haarlets* to improve the speed of the system. Using nearest neighbors search, the coefficients are then matched to hand gestures stored in a database.

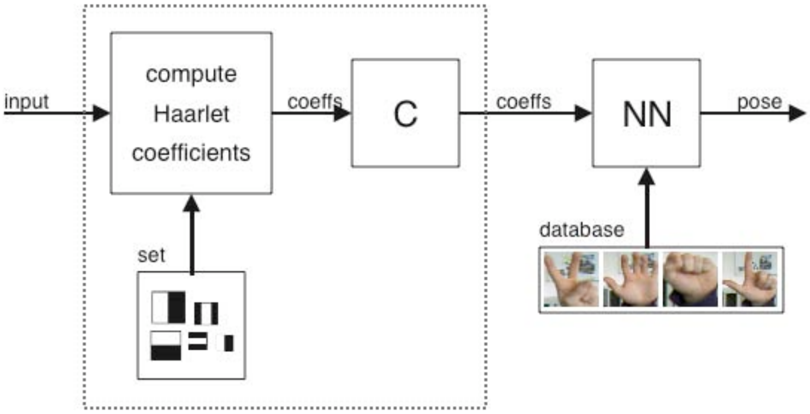


FIG. 5: Structure of the classifier illustrating the tranformation **T** *(dotted box), approximated using Haarlets. The Haarlet coefficients are computed on the input sample. The approximated coefficients (that would result from* **T***) are computed as a linear combination* **C** *of the Haarlet coefficients.*

4. EXPERIMENTS

In this section, we describe the demo application that allows for the visualization of 3D models that can be loaded into the program. Using hand gestures, the user can zoom in on the model, pan and rotate it.

4.1.1 Gestures

The hand gesture classifier is trained based on a set of training samples containing the gestures shown in figure 6. An example of the system detecting these static gestures is shown in figure 7.

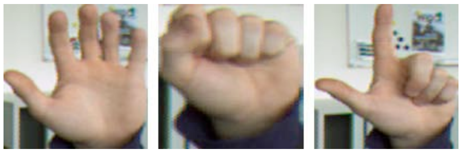
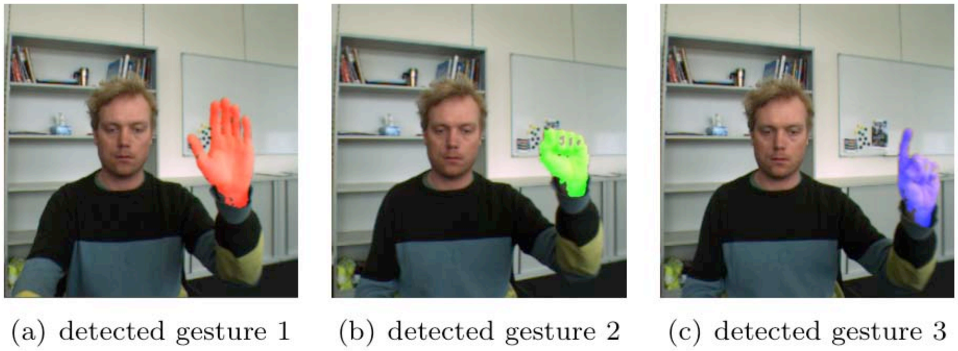


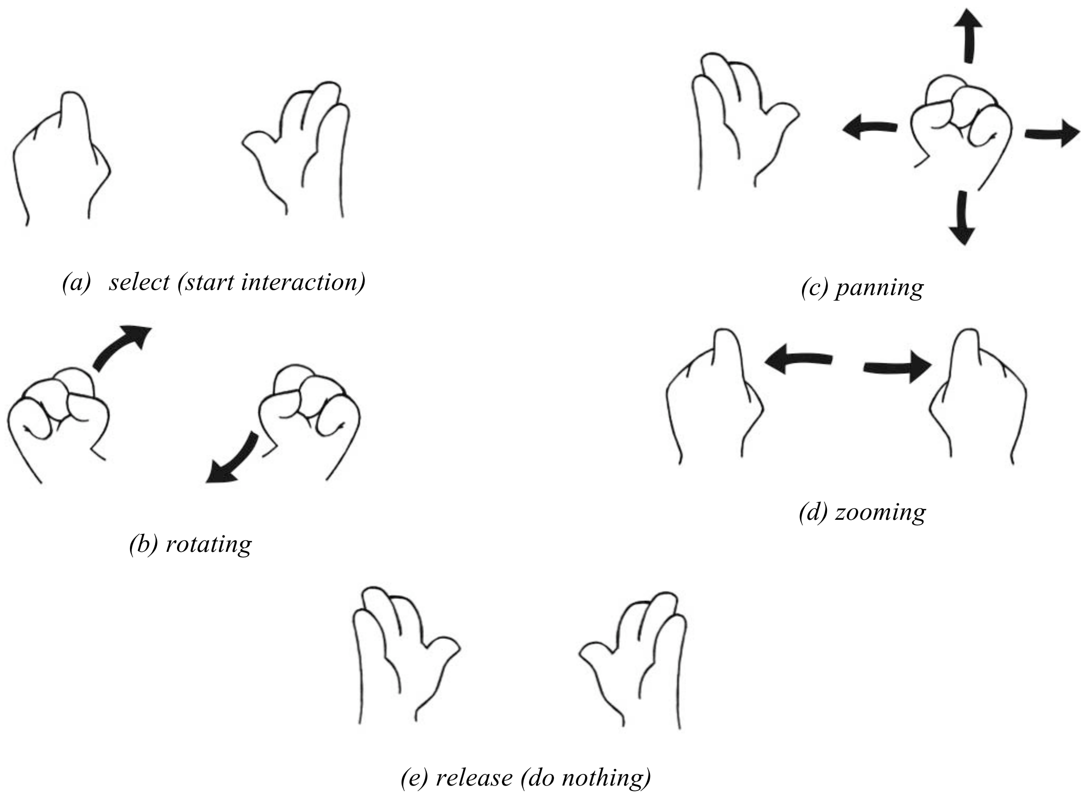
FIG. 6: The gestures that are trained in the hand gesture classifier.



(a) detected gesture 1 (b) detected gesture 2 (c) detected gesture 3

FIG. 7: Examples of the hand gesture recognition system detecting different hand gestures.

The hand gesture interaction in this application is composed of the hand gestures shown in figure 6. It recognizes the gestures and movements of both hands to enable the manipulation of the object/model. Pointing with one hand selects the model to start manipulating it. By making two fists, the user can grab and rotate the model along the *z*-axis. By making a fist with just one hand, the user can pan through the model. By making a pointing gesture with both hands, and pulling the hands apart, the user can zoom in and out of the model. The open hands release the model and nothing happens until the user makes a new gesture. An overview of these gestures is shown in figure 8.



(a) select (start interaction)

(c) panning

(b) rotating

(d) zooming

(e) release (do nothing)

FIG. 8: The hand gestures used for the manipulation of the 3D object on the screen.

4.1.2 Application

The interaction system above has been implemented as an extension of an open-source 3D model viewer, the *GLC Player* (http://www.glc-player.net/). This enables us to 1) load models in multiple formats (OBJ, 3DS, STL, and OFF) and of different sizes, and 2) use our hand interaction system in combination with standard mouse and keyboard interaction. Pressing a button in the toolbar activates the hand interaction mode, after which the user can start gesturing to navigate through the model. Pressing the button again deactivates the hand interaction model and returns to the standard mouse-keyboard interaction mode.

We conducted experiments by installing our system in the Value Lab and tested with multiple 3D models, and in particular with a model created as part of the Dübendorf urban planning project. This model represents an area of about 0.6 km2 and is constituted of about 4000 objects (buildings, street elements, trees) with a total of about 500,000 polygons. Despite this size, our system achieved frame rates of about 30fps (frame per second), which is sufficient for smooth interaction. Examples of the user zooming, panning and rotating through the 3D model are shown in figures 9, 10 and 11 respectively. In each figure, the left column shows side and back views of the system in operation at the beginning of the gesture, and the right columns the same views but at the end of the gesture.

The hand interaction mode is currently only available for model navigation (rotation, panning and zooming), all the other features of the viewer being only accessible in mouse-keyboard interaction mode. Nonetheless, our implementation enables simple extensions of the hand interaction mode. In the near future, we for instance aim to enable the hand interaction mode for object selection (to view its properties).

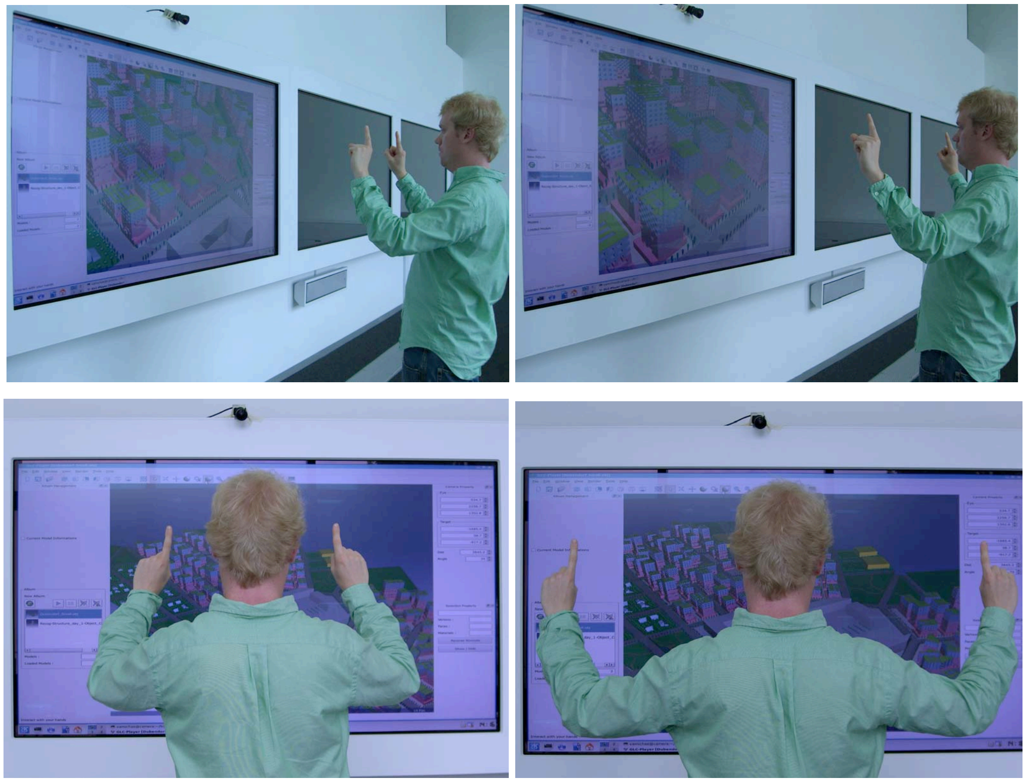


FIG. 9: Zooming into the model.

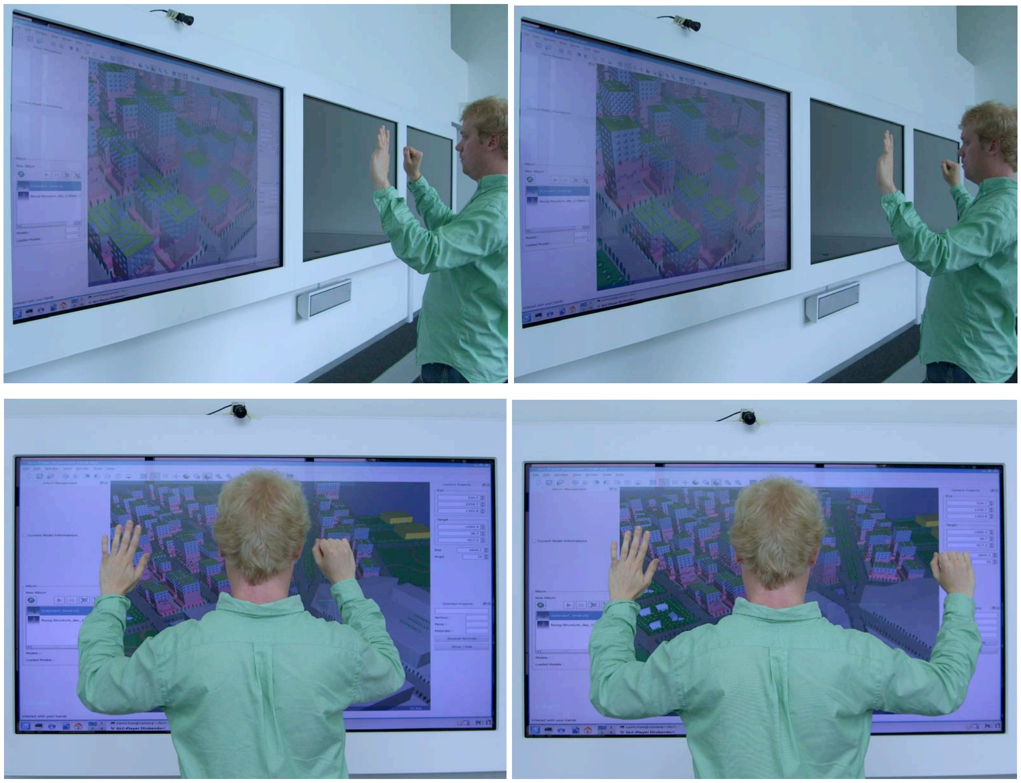


FIG. 10: Panning the model.

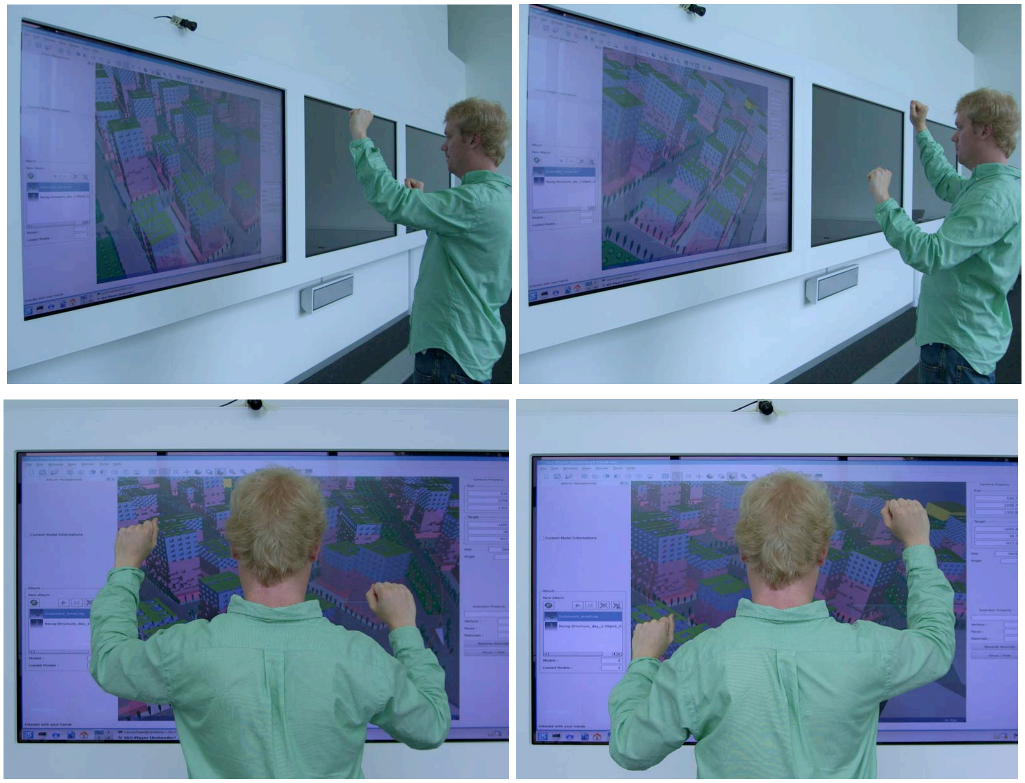


FIG. 11: Rotating the model.

5. CONCLUSION

In this paper, we first described the need for novel human-computer interaction tools, enabling users in information visualization caves to simultaneously interact with large amounts of information displayed on multiple screens spread around the cave. Today's urban design tasks could be significantly enhanced in terms of interaction especially when different stakeholders are involved. Currently available interaction devices, such as mouse-keyboard or screen (multi-)touch capabilities, are often not adapted to such requirements, and this was confirmed in an urban design project conducted in the Value Lab at ETH Zurich.

A novel solution for human-computer interaction was then introduced that is based on vision. Compared to currently existing systems, it presents the advantage of being marker-less. Experiments, conducted in the Value Lab, investigated the usability of this system in a situation as realistic as possible. For these, our interaction system has been integrated to a 3D model viewer, and tested with a large 3D model of an urban development project. The results show that our system enables a smooth and natural interaction with 3D models.

Nonetheless, these results remain preliminary. The system is not always as robust as it should be, and its applicability to enable multiple users to simultaneously interact with multiple screens remains to be demonstrated. Future work will thus be targeted to: (1) extend the set of viewing features accessible through hand gesture (in particular object selection and de-selection); (2) further improve the robustness of the system, particularly with respect to different users; and (3) develop a larger system containing multiple cameras and enabling the interaction of multiple users with different screens.

6. REFERENCES

Fox, A., Johanson, B., Hanrahan, P., and Winograd, T. (2000). Integrating information appliances into an interactive workspace, *IEEE Computer Graphics & Applications*, 20:3, May/June, 54-65.

Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Van Gool, L., Lang, S., Strehlke, K., Moere, AV., and Staadt, O. (2003). Blue-c: a spatially immersive display and 3D video portal for telepresence. *In ACM SIGGRAPH 2003 Papers, San Diego*.

Halatsch, J., and Kunze, A. (2007). Value Lab: Collaboration In Space. *Proceedings of 11th International Conference Information Visualization (IV07)* 4-6 July 2007, Switzerland Zurich, IEEE, pp 376 – 381

Halatsch, J., Kunze, A., Burkhard, R., and Schmitt, G. (2008a). ETH Value Lab - A Framework For Managing Large-Scale Urban Projects, *7th China Urban Housing Conference, Faculty of Architecture and Urban Planning*, Chongqing University, Chongqing.

Halatsch, J., Kunze, A., and Schmitt, G. (2008b). Using Shape Grammars for Master Planning, *Third conference on design computing and cognition (DCC08)*, Atlanta.

Jedynak, B., Zheng, H., Daoudi, M., and Barret, D. (2002). *"Maximum entropy models for skin detection," Technical Report publication IRMA*, vol. 57, no. 13.

König, W. A., Bieg, H.-J., Schmidt, T., and Reiterer, H. (2007). Position-independent interaction for large highresolution displays, *IHCI'07: IADIS International Conference on Interfaces and Human Computer Interaction 2007*, IADIS Press, p. 117-125.

Müller, P., Wonka, P., Haegler, S., Ulmer, A., and Van Gool, L. (2006). Procedural Modeling of Buildings. In *Proceedings of ACM SIGGRAPH 2006 / ACM Transactions on Graphics (TOG)*, ACM Press, Vol. 25, No. 3, pages 614-623.

Schmugge, S. J., Jayaram, S., Shin, M. C., and Tsap, L. V. (2007). *"Objective evaluation of approaches of skin detection using ROC analysis." Computer Vision and Image Understanding*, vol. 108, 41–51.

Van den Bergh, M., Koller-Meier, E., and Van Gool, L. (2009). *"Real-time body pose recognition using 2D or 3D Haarlets." International Journal on Computer Vision, vol. 83, 72-84.*