INTERACTIVE ART SYSTEM FOR MULTIPLE USERS BASED ON TRACKING HAND MOVEMENTS

Byunghyung Kim[1], Danna Gurari[1], Hough O'Donnell[2], Margrit Betke[1]
Boston University
1 Image and Video Computing Group, Department of Computer Science,
2 School of Visual Arts, College of Fine Arts

ABSTRACT

We propose a camera-based human computer interface where hand movements directly influence the creation and manipulation of art in an interactive system. This unobtrusive, low-cost interface addresses limitations in previous interactive art systems by supporting multiple users simultaneously in an unconstrained, uncontrolled environment. We chose Markov Models for our computer vision algorithm to track multiple hands across multiple camera views in real time. We show how our algorithm can be applied as a tool for a team of users to interact with *Strange Attractors* and *Hidden Paintings. Strange Attractors* is an interactive art system that displays beautiful fractal structures whose location on a large screen reflects the users' hand movements. *Hidden Paintings* is an interactive game in which two almost identical versions of an artwork are displayed on a large screen and the movements of both hands of the users are superimposed as bounding boxes onto the display, allowing users to identify image regions in one version of the artwork that differ from the regions in the other. Experiments with twelve subjects show successful adoption of the proposed interactive art system and its two applications.

KEYWORDS

Human computer interaction, camera-based human computer interfaces, computer vision, art, games

1. INTRODUCTION

Over the past decade, various systems have been explored that enable people to immerse themselves in images or videos and interact through their body movements. These systems have contributed to innovations in the performing arts, education, and entertainment [3][5][9][11][12][14][17]. This paper proposes a camera-based human computer interaction (HCI) system where movements of users' hands directly influence the creation and manipulation of art.

Related innovations in the performing arts include dance performances where a dancer's movements are tracked and used to steer a virtual representation in real time on a large monitor behind the dancer. In one setup, dancers' movements are detected by monitoring the location of wearable sensors [15]. This approach is limited because the sensors potentially impair the dancer's ability to move naturally. Another system uses 16 cameras to record body movement [12].

Related interactive art includes applications for improving the educational experience of children. One example is the *KidsRoom*, a media installation for guiding children through a story by interpreting their movements [3]. The installation was built as a bedroom-sized structure in a laboratory that contains a mattress and rug and is configured with the appropriate cameras, lighting, and screens. Another example is *SMALLab*, an immersive media environment used for kindergarten-through-12th-grade education to support active learning [2]. In this freestanding environment, teachers and groups of students interact with one another and their composed media worlds through free play, structured movement, and vocalization. A limitation for both systems is the complex setup required. *KidsRoom* requires students to visit the laboratory structure and *SMALLab* requires equipment for 3D-vision-based movement tracking, an audio microphone array, a video projector, and a multi-channel surround audio system.

The proposed HCI system is camera-based where a computer captures video of peoples' movements. The contribution of the proposed system, as an alternative to existing work, is a new, fun interface that addresses the limitations of the previous systems in a single system:

- **Multiple Users Supported** - several people can use the system simultaneously, encouraging interaction between users in addition to between the user and the system.
- **Unconstrained User Space** - any number of cameras can be used to increase the field of view and therefore allow for user movement across larger areas.
- **Unencumbered User Activity in Physical Space** – there are no wearable sensors; people can move naturally and interact with or create artworks using simple hand movements,
- **Easy to Use –** users can vary the direction and tempo of hand movements, making it easy and intuitive to create and interact with art pieces.
- **Simple installation -** minimal setup requirements makes the system suitable for everyday use (e.g., homes, offices, classrooms) and high-traffic spaces (e.g., galleries, museums).

Computer vision is used as the bridge between users' movements and art. The computer vision algorithm tracks hand motions (Section 2) and maps that movement graphically to the display in a way that is determined by the specific application (Section 3). The physical setup involves users, one or several low-cost web cameras, and a low-cost computer (Figure 1).

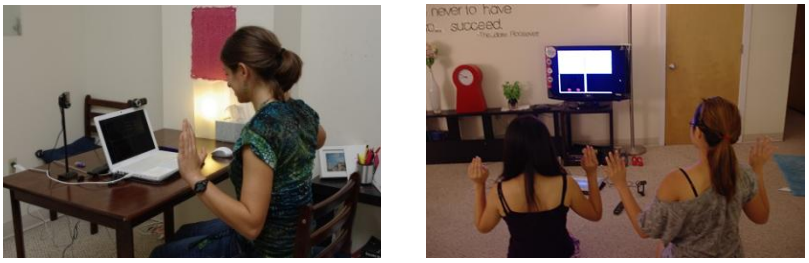

Figure 1. Interactive art system in use in two home environments with minimal hardware setup.

The key technical challenge in designing the proposed system was the development of a tracking algorithm that addresses the following requirements. The algorithm must have the ability to *simultaneously* track *multiple* hands of *different* users in *real time*. This means it must associate hands observed in any of the camera views with the hands tracked across previous views and discriminate against false hand detections. Many tracking methods, such as optical flow [1] and *CAMshift* [8], track objects based on features within shared fields of view between successive video frames making them unsuitable for tracking across multiple camera views. We employ a Markov Model to solve this data association problem, extending methods used at the University of Adelaide [6]. Our approach differs in that we track human hands instead of the human body and we permit human hands to move in various directions rather than just one direction. A Markov-Model-based approach is ideal because it uses observable states (not like hidden Markov models [7]), is computationally fast and accurate when tracking multiple objects, and distinguishes false hand detections. In addition, it handles tracking across multiple, non-overlapping camera views since discontinuities in visibility between two or more fields of view are accurately reflected in the Markov Model.

We experimented with two applications (Section 3). In *Strange Attractors*, our interface enables people to *create* beautiful fractal structures with quick energetic hand motions. A related application defines the form of the fractal structure with mouse input [10]. In *Hidden Paintings*, our interface enables players to actively *observe* artwork by pointing to different areas on pictures where expected objects are missing. Both applications encourage human-computer interactions based on a wide range of hand motions from quick, energetic gestures to carefully directed pointing. They demonstrate how the proposed interface can serve as a platform for engaging applications that support playful, creative interactions between users.

2. METHOD

A design goal of our interactive art system is that it can be used inexpensively for a wide range of applications in uncontrolled environments. It must therefore be resilient to real-world challenges such as

changing illumination conditions, poor quality video, different backgrounds, many hand poses, a wide range of skin tones, and occlusions. It must also provide real-time feedback. A flowchart of the proposed system is shown in Figure 2. The computer vision part of the system involves *detection of multiple hands* and *tracking of hands* and is followed by an *interactive art application.*

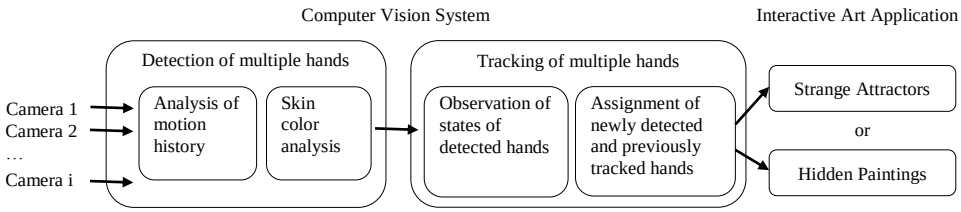

Figure 2. Overview of system for detection and tracking of multiple hands

2.1 Detection of Hands

Previous work demonstrated that motion and skin color are good indicators of hands in image sequences [18]. The proposed system therefore uses these cues for detecting hands. Motion can be analyzed in a computationally inexpensive way with a Motion History Image (MHI) [3]. MHI is a scalar-valued image where the pixel intensity is a function of how recently the motion occurred, with brighter pixels corresponding to the most current timestamps. The length of motion history retained is tunable. Our system captures a MHI from the past four images and creates a binary image with all static areas set to "0" (white in Fig. 3) and all movement areas set to "1" (blue in Fig. 3).

Figure 3. Visible images (top) and binary versions of corresponding motion history images (bottom)

In the binary image, clusters of pixels are grouped using connected component labeling and marked as hand candidates. Classification of hand candidates as hands is based on the dominant color and the size of the candidate. The skin color model *F* is defined during a training phase where a range of skin-colored hue values is determined. The hue component describes color in the form of an angle between 0 and 360 degrees (0^o = red, 120^o = green, 240^o = blue, 60^o = yellow, 300^o = magenta). The skin-color model is defined as:

$$F(x) = \begin{cases} \text{True}, & T_{min} \leq \text{hue value of pixel } x \leq T_{max} \\ \text{False}, & else \end{cases}$$

where $T_{min} = 70^o$ and $T_{max} = 256^o$ are the minimum and maximum hue values, respectively, that indicate skin color. For each hand candidate region, if the peak value of its hue histogram falls within the pre-defined range of skin colored hues and its size is greater than a pre-defined minimum size, it is classified as a hand. This feature-based approach is ideal for our application because it is simple, fast, and accurate.

2.2 Multiple-Hand Tracking across Several Non-overlapping Camera Views

The next challenge addressed by the system is tracking the left and right hands of the interacting users that are in the field of view of one of the cameras. The proposed solution is based on a Markov Model which

characterizes how each hand will move in the future using knowledge of how the hand has moved in the past. At time *t*, our system uses *i* cameras with non-overlapping views $C_1, \ldots, C_i$, has detected *j* hands $D_1, \ldots, D_j$ in these *i* views, and maintains a list of *m* tracked hands $H_1, \ldots, H_m$,. The task to assign the *j* currently detected hands to the *m* previously tracked hands is solved at each time step *t* by selecting the most probable set of assignments based on the transition probabilities of the Markov model. In the proposed system, an observable Markov Model state is a region in the camera view. Each video frame is divided into 16 states resulting in 32 states for a two camera setup as shown in Figure 4. The number of states must be selected such that a balance is struck between tracking accuracy and system speed. More states return improved tracking accuracy at the cost of slower performance and vice versa. The parameters defining the Markov Model are the state transition probability matrix and initial state probability vector, which are estimated during a training phase. When the proposed system is in the tracking phase, the transition probabilities of the Markov Model guide the tracking process.

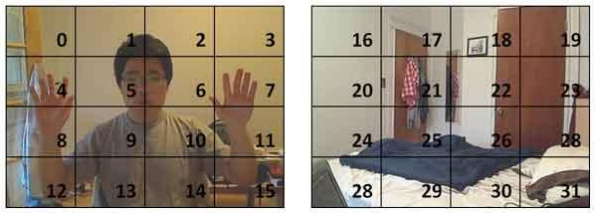

Figure 4. Each image is divided into 16 states resulting in 32 states across two camera views

Training Phase: Estimates for the initial probability vector are assigned such that each state has an equal probability. The Markov Model used is ergodic, which means that any state can be reached from any other state in a finite number of steps. The state transition probability matrix is therefore initialized with equal probabilities. During training, its entries are adjusted using a series of observations collected as users move their hands in front of the cameras. The resulting matrix stores the probabilities that hands will move from a given region in an image (state) to another region in this image or in the image recorded by an adjacent camera (state). The transition probabilities trained for state 5, which represents a region in the center of the first image, are shown in Figure 5. The data shows that the hands did not move far during training and most likely appeared in surrounding image regions. This reflects the nature of human movement. As a result, in practice, we only need to be concerned with a localized area for state transitions.

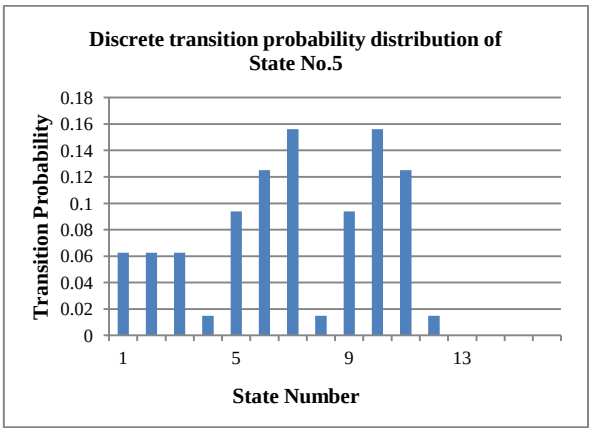

Figure 5. Discrete transition probability distribution of a trained Markov Model for state 5

Tracking Phase: During tracking, the system assigns the *j* detected hands to the *m* previously tracked hands. At each time step *t*, every detected, unmatched hand is allocated to its corresponding state using the location of its center in the image where it appears for the mapping. For each detected hand D, the most probable assignment to a tracked hand is calculated. In this calculation, a sliding time window of duration *k* is used. For each tracked hand H, the likelihood that it ended up in the image region (state) where D has been

detected is computed by multiplying the probabilities of the last *k* transitions of hand H, ending with the transition into D. The assignment process has quadratic time complexity *O(j m)* at each time step.

2.3 *Strange Attractors:* Creating Art by Gesturing with Both Hands

The proposed application, *Strange Attractors*, is an interactive art system that displays fractal structures whose location, shape, and color reflect the users' hand motions (Figure 6). It was designed for an installation in the lobby of a high-traffic dormitory building at our university. Hand movements captured by the primary camera are passed as parameters driving the display of a new fractal structure. The fractal structure is generated by a dynamic system described by a differential equation [10]. Hand movements captured by the secondary camera change the dominant color of the rendered fractal structures. The *Strange Attractors* application encourages users to playfully interact with the computer and learn how they can manipulate the apparent chaos of displayed fractal structures.

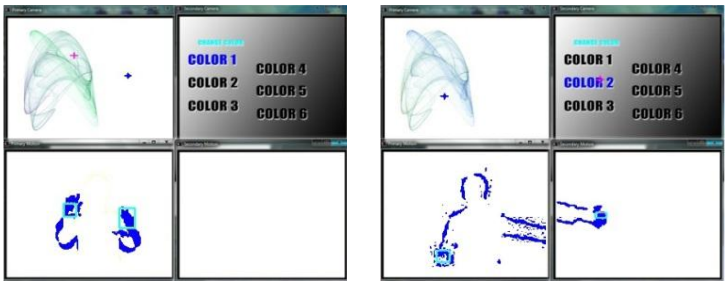

Figure 6. *Strange attractors*: The interface contains windows that displays the art (top left), the controller of the attractor color (top right) and the tracking results (cyan boxes) overlaid on the motion history image (blue) for the two camera views (bottom). The positions of the tracked hands in the primary camera view (bottom left) are input parameters to create the shape and location of the attractor. Tracked hands in the secondary camera view (bottom right) change the attractor color (here from green to blue).

2.4 *Hidden Paintings:* Interacting with Art by Gesturing with Both Hands

A second application that uses our proposed multiple-hand tracking system is an interactive game called *Hidden Paintings*. It is an educational, media-based application for locating differences in two almost identical versions of a picture using hand gestures. Multiple players can play together to find the hidden objects in the picture. A screen shot of a game is shown in Figure 7.

Figure 7. *Hidden Paintings*: The interface contains windows that display two versions of a picture* and tracking results. The detected hand locations (cyan) are shown overlaid onto the binary version of the motion energy image. The tracked locations of the left and right hands are mapped to corresponding positions in the original version of the picture (blue and green rectangle). The goal of the user is to move the rectangles onto the objects that do not appear in the other version of the picture, here the fence and the red apple. (* The original picture was obtained from Google Image Search and manipulated with image editing software to create a version with "hidden" objects.)

3. EXPERIMENTS AND RESULTS

Two sets of experiments were conducted using two web cameras streaming visible video of two adjacent, non-overlapping views. Cameras were positioned approximately 25-40 cm apart from each other with their

optical axes aligned approximately parallel to one other. In the first set of experiments, the computer vision system was tested; in the second set of experiments, the focus of the tests was the HCI component.

3.1 Evaluation of Multiple-User Hand Tracking across Camera Views

For these initial experiments, learning of the Markov Model was performed off-line through manual annotations of three video sequences in which users were asked to move "just for fun." As a result, the trained Markov model represents arbitrary movements in space. In one system set-up, the hands of a single participant moved freely within the field of views of two cameras. Experiments determined that the system successfully tracked motion of hands between the two fields of views. In a second system setup, the system successfully tracked two people's four hands which moved freely within the fields of view of two cameras (Figure 8). The main challenge of this setup is assigning hands correctly when they are close to each other or occluding each other (Figure 8, row 6). This problem can be solved by dividing each field of view into more regions resulting in additional states in the Markov Model. The downside of increasing the number of states is slower system performance. In testing, by increasing the number of total states from 50 to 98, the overall tracking accuracy increased from 68.95% to 86.64% based on 1430 observations.

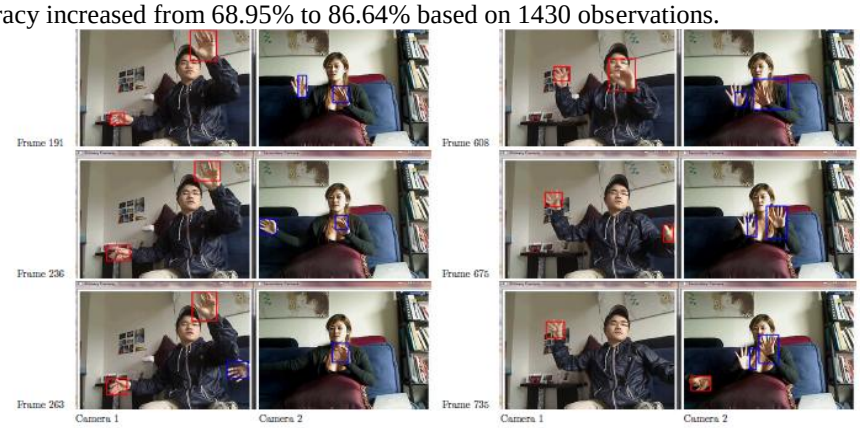

Figure 8. The system tracks the hands of two participants who move them in either camera's field of view.

3.3 Evaluation of the Interactive Art System

Experiments were conducted with twelve participants playing *Strange Attractors* and *Hidden Paintings*, both which are based on camera-detected hand movements. Participants included Caucasians, Asians, and African Americans with different skin color tones. Their ages ranged from 5 to 60. Participants cooperated with the system by not creating occlusions of hands. The experiments were run in a variety of settings including public places such as coffee shops and private places such as apartments.

One goal was to test whether the system sufficiently would track quick, energetic motions as well as precisely articulated motions as users move naturally in a variety of environments. Participant feedback about *Strange Attractors* was that the fractal structures looked beautiful, but that it was difficult to recognize the connection between the shape of the structures and the users' motions. Fittingly, users made a range of lively hand movements as they tried to learn how the system would respond to their actions. *Hidden Paintings*, in contrast, encouraged defined hand movements from users. The experiments showed users enjoyed creating fractal art with *Strange Attractors* and playing the *Hidden Paintings* game, indicating system performance was subjectively sufficient for users as they moved in a variety of ways. Moreover, consumer-level cameras provided adequate video quality and our system successfully detected various skin tones.

In a quantitative set of experiments, we compared the time it took participants to complete a task with camera-detected hand movements with the time it took them to perform the same task with the traditional computer mouse. The assigned task was to finish eight stages of the *Hidden Paintings* game. The total time per stage was calculated for each player and the results were grouped according to the players' ages (Figure 9). Ages for each group are 5-19 for Group 1, 20-40 for Group 2, and 41-60 for Group 3. A common factor for most players was that less time was taken with experience; the interaction time typically

decreased from game 1 to 5. It is concluded that people become accustomed to the interface through repeated experiences. Another interesting result is that younger generations comparatively performed the tasks faster than older generations, independent of the interface. We also measured that the time to finish the game was lower for mouse-based interaction than for camera-based interaction by approximately a factor of 2/3. In an extreme case, it took a participant more than twice as long to finish a game with the camera-based interface than with the mouse-based interface.

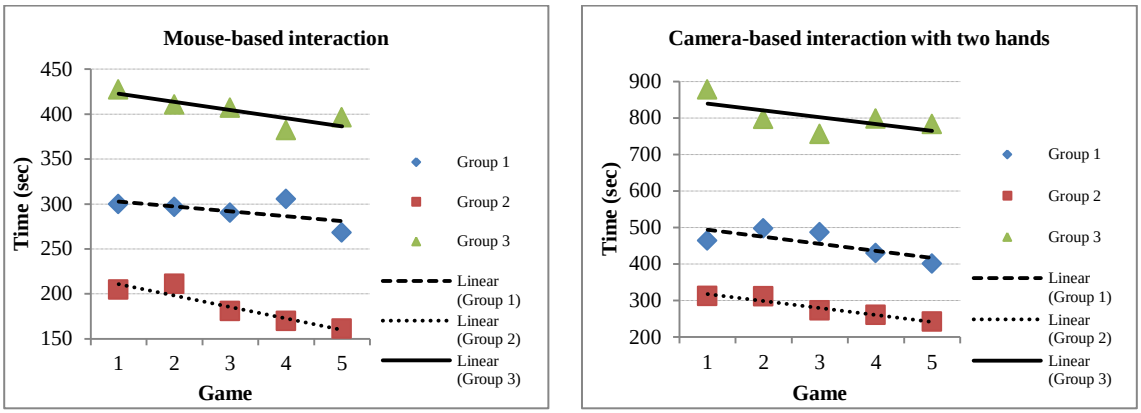

Figure 9. Timing results for mouse-based interactions (left) and camera-based interaction with two hands (right).

We also computed the ratio of average mouse-based and camera-based task-completion times per game and per group in two experiments. The plot in Figure 10 indicates that the rate of the task-completion time stayed flat for groups 2 and 3 and increased for group 1. This means that the younger participants were able to speed up their interaction with the previously unknown camera-based interface much more than with the common mouse-based interface. The older participants might have also experienced this speed-up if they had more practice. The question remains open whether training could result in camera-based interaction times that are similar to mouse-based interactions. The difference in interaction times becomes irrelevant for art installations where user access to a computer mouse is impractical (e.g., shopping mall installations).

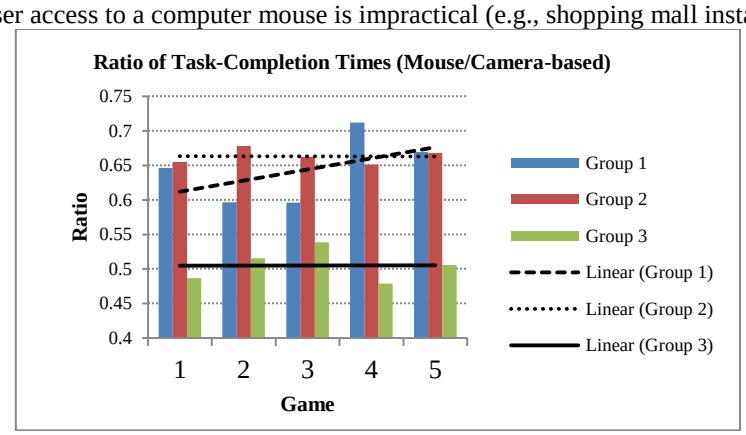

Figure 10. Comparison of mouse-based and camera-based task-completion times for *Hidden Paintings*

4. CONCLUSIONS AND FUTURE WORK

The proposed interface enables users to create and manipulate artwork with movements of their hands. It sufficiently tracks and interprets a wide range of hand motions from quick, energetic gestures to carefully directed pointing. Users may interact with each other while moving naturally in a variety of environments. The interactive system is low cost and easy to use, requiring only inexpensive web cameras and computers with minimal setup. It currently supports two example applications, *Strange Attractors* and *Hidden Paintings*. Participants of HCI experiments with the interface enjoyed creation and manipulation of artworks.

Future work will be focused on three tasks. First, the computer-vision part of the system will be enhanced so that the creation of the Markov Model is automated. The system can then be used more easily in new environments. Second, we plan to add other applications to the system that enable users to create and manipulate artworks. Third, we will study user reaction to the system when it becomes an art installation in the lobby of our university dormitory.

ACKNOWLEDGEMENTS

The authors thank the test subjects for their time and gratefully acknowledge NSF funding (IIS-0910908, IIS-0855065, and IIS-0713229).

REFERENCES

[1] Adiv, G., 1985. Determining Three-Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence:* 7(4):384-401.

[2] Bircheld, D. et al., 2006. Sound and interaction in K-12 mediated education. *Proceedings of International Computer Music Conference*. New Orleans, USA.

[3] Bobick, A. F. et al., 1999. The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. *In PRESENCE: Teleoperators and Virtual Environments*, Vol. 8, No. 4, pp. 369-393.

[4] Bobick, A. F. and Davis, J. W., 2001. The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*: 23(3):257 -267.

[5] Castellano, G. et al., 2007. Expressive Control of Music and Visual Media by Full-Body Movement. *Proceedings of the Conference on New Interfaces for Musical Expression*. New York, USA, pp. 390-391.

[6] Dick, A. and Brooks, M., 2004. A Stochastic Approach to Tracking Objects Across Multiple Cameras. *Proceedings of Australian Conference on Articial Intelligence*. Sydney, Austrailia, pp. 160-170.

[7] Elmezain, M. et al., 2008. A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory. *Proceedins of 19th International Conference on Pattern Recognition*. Florida, USA, pp. 1-4.

[8] Exner, D. et al., 2010. Fast and robust CAMShift tracking. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*. San Francisco, USA, pp. 9-16.

[9] Freeman, W. T. et al., 1998. Computer vision for interactive computer graphics. *IEEE Computer Graphics and Application*, 18(3): 42-53.

[10] Grant, S., 2008. Attractors and Fractals. http://www.cc.gatech.edu/~phlosoft/attractors.

[11] Hämäläinen, P. and Höysniemi, J., 2002. A Computer Vision and Hearing Based User Interface for a Computer Game for Children. *Proceedings of the 7th ERCIM Workshop "User Interfaces For All"*. Paris, France, pp. 23-25.

[12] James, J. et al., 2006. Movement-based interactive dance performance. *Proceedings of the 14th International Conference on Multimedia*. Santa Barbara, USA, pp. 470-480.

[13] Kakumanu, P. et al., 2007. A Survey of Skin-Color Modeling and Detection Methods. *Pattern Recognition*, 40(3): 1106-1122.

[14] Levin, G., 2006. Computer Vision for Artists and Designers: Pedagogic Tools and Techniques for Novice Programmers. *In Journal of Articial Intelligence and Society*, Vol. 20, No. 4, pp.462-482.

[15] Park, C. et al., 2006. A wearable wireless sensor platform for interactive art performance. *Proceedings of IEEE International Conference on Pervasive Computing and Communications*. Pisa, Italy.

[16] Pickover, C., 1995. *The Pattern Book: Fractals, Art, and Nature*. World Scientic Publishing Company.

[17] Sparacino, F. et al., 2000. Media in performance: Interactive spaces for dance, theater, circus, and museum exhibits, *IBM Systems Journal*, 39: 479-510.

[18] Zhu, Q. et al., 2004. Adaptive learning of an accurate skin-color model. *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*. Seoul, Korea, pp. 37-42.