

Sketch on Dynamic Gesture Tracking and Analysis Exploiting Vision-based 3D Interface

Woontack Woo, Namgyu Kim, Karen Wong and Makoto Tadenuma

ATR MIC Labs

Kyoto 619-0288, Japan

E-mail:{wwoo,ngkim,karen,tadenuma}@mic.atr.co.jp

ABSTRACT

In this paper, we propose a vision-based 3D interface exploiting invisible 3D boxes, arranged in the personal space (*i.e.* reachable space by the body without traveling), which allows robust yet simple dynamic gesture tracking and analysis, without exploiting complicated sensor-based motion tracking systems. Vision-based gesture tracking and analysis is still a challenging problem, even though we have witnessed rapid advances in computer vision over the last few decades. The proposed framework consists of three main parts, *i.e.* (i) object segmentation without bluescreen and 3D box initialization with depth information, (ii) movement tracking by observing how the body passes through the 3D boxes in the personal space and (iii) movement feature extraction based on Laban's Effort theory and movement analysis by mapping features to meaningful symbols using time-delay neural networks. Obviously, exploiting depth information using multiview images improves the performance of gesture analysis by reducing the errors introduced by simple 2D interfaces. In addition, the proposed box-based 3D interface lessens the difficulties in both tracking movement in 3D space and in extracting low-level features of the movement. Furthermore, the time-delay neural networks lessens the difficulties in movement analysis by training. Due to its simplicity and robustness, the framework will provide interactive systems, such as ATR I-cubed Tangible Music System or ATR Interactive Dance system, with improved quality of the 3D interface. The proposed simple framework also can be extended to other applications requiring dynamic gesture tracking and analysis on the fly.

Keywords: vision-based 3D interface, dynamic gesture tracking, movement analysis, Effort Theory

1. INTRODUCTION

3D virtual environments stir up new challenges for human-computer interface/interaction (HCI). The choice of the HCI techniques strongly influence the acceptability and efficiency for the user within virtual environments. Over the last few years, many other researchers have reported a number of promising new ideas for improving the HCI. Consequently, traditional HCI methods, such as the keyboard or mouse, will increasingly be replaced by intuitive mechanisms such as speech, posture of hands, direction of view, gesture, etc.

In this paper, we focus on 3D space as an interface and provide a way to communicate with a computer by simply making dynamic gestures, instead of manipulating a keyboard and/or a mouse. In general, we can bring HCI closer to human-human interaction by providing computers with the ability to understand gestures, speech and facial expressions. In particular, gesture recognition could provide a more expressive interaction with a computer than speech or text does. A large amount of research on gesture analysis has been reported over the last few decades due to its importance in communication between the human and the computer, as well as between humans [1]. However, the available 3D movement tracking and analysis schemes are far from being a comfortable and/or natural HCI.

In general, gestures in 3D space can be traced based on the information captured by a camera. 2D vision-based approaches, however, due to the lack of 3D information, have inherent weaknesses in tracking 3D gestures. Therefore, conventional vision-based approaches have mainly been focused on *static* (or pose) gesture recognition, rather than *dynamic* (or time-dependent) gesture recognition. Only a few research works on dynamic gesture analysis have been reported. A simple 2D box-based interface been applied to track/analyze dance gestures in real video-rate (10-30 Hz). [2-4]. Meanwhile, the movements of a user can be tracked by attaching special reflective material to the person's joints and limbs*. However, the attached markers on the body encumber the user and hinder free expression.

*The "DOZO" system (Kleiser-Walczak Construction Company) tracks detailed movements by attaching special reflective material and then translates the motions onto a computer-generated avatar.

Movement in 3D space also can be traced using fiber optic sensors or magnetic trackers. One of the first attempts to capture gestures based on sensors was to use a glove, wired up with sensors to detect the orientations of the hand and fingers, as shown in Figure 1 (a) [†]. The following attempt was the "DataSuit", which used the same fiber optic flex-sensing technology. As shown in Figure 1 (b), this full body suit can track the movement of all the user's limbs such as arms, legs, feet, and torso[‡]. However, the weakness of marker or sensor-based approaches stem from an uncomfortable interface, which encumbers the user and hinders free expression, because they have to be worn or attached on the body and connected to computers with wires (or radio links).

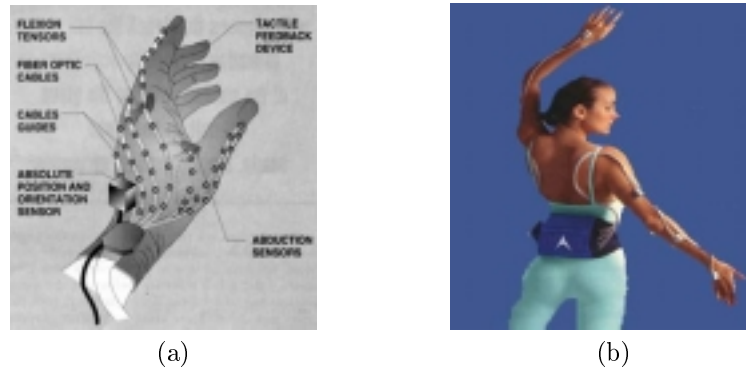


Figure 1. Sensor or marker-based movement tracking. (a) DataGlove (b) DataSuit. These images were downloaded from the webpages at http://www.niee.ufrgs.br/~claudia/interface/body_interface.html and <http://www.ipct.pucrs.br/leticia/suit.html>, respectively.

Although sensor or marker-based 3D motion tracking provides relatively accurate position and movement information over vision-based tracking, they have several inherent problems. The devices need complicated calibration and conversion procedures to get accurate movement data. Given accurate position and movement data, there still remains several other hindrances obstructing widespread usage of 3D interfaces [5]. First of all, proper low-level features have to be selected from the obtained movement data and processed into meaningful features. Then, a robust yet simple movement analysis framework, mapping features to meaningful symbols, is required. This is necessary for the 3D interface to be utilized for applications requiring *dynamic* gesture analysis on the fly.

In this paper, we propose a unified vision-based dynamic gesture tracking/analysis framework. As shown in Figure 2, the proposed framework consists of three main parts, *i.e.* (i) object segmentation without bluescreen and 3D box initialization with depth information, (ii) movement tracking by observing how the body passes through the invisible 3D boxes in the personal space and (iii) movement feature extraction based on Laban's Effort theory and movement analysis by mapping features to meaningful symbols using time-delay neural networks. Without using bluescreen, we first segment the user from the background and estimate depth information using multiview images. Given depth information we estimate the center of the segmented body to initialize the personal space and allocate invisible 3D boxes to the personal space. Next, we track the movement of the body by observing how the body passes through the invisible 3D boxes in the personal space, instead of tracking the body itself. Finally, we extract movement features based on Laban's Effort theory and then analyze by mapping the features to meaningful symbols using time-delay neural networks.

The proposed box-based 3D interface lessens the difficulties in both tracking movement in 3D space, and extracting features of the movement. Obviously, exploiting depth information using multiview images improves the performance of gesture analysis by reducing the errors introduced by simple 2D interfaces. Such vision-based invisible sensors in the space also reduces tracking errors resulting from the unpredictable movement of the user's skirt or clothing. Unlike such complicated motion data acquisition equipment (such as heavy headsets, data gloves, tethers)

[†]First "dataglove" developed in the late 1970s at the University of Illinois, and then it became commercially available in 1980 as a product from VPL, which used optical fibers to measure finger bending, and an electromagnetic sensing system for hand orientation. A redesigned version of the dataglove, using conductive ink and ultrasonic position sensing, was later marketed by Mattel as the "PowerGlove", for use with Nintendo video games.

[‡]DataSuit uses up to fifty different sensors on the users joints and with four position sensors (Polhemus trackers, two for the hands, one for the head, one for the back of the suit).



Figure 2. Basic structure for movement tracking and analysis using vision-based 3D interface. In the proposed vision-based 3D interface, an object is segmented out from background scene and then movement is tracked by observing the invisible 3D boxes in the segmented user’s personal space. The movement features, extracted based on the Laban’s movement theory, are mapped to meaningful symbols by time-delay neural networks.

that attach infrared or magnetic sensors to the user, the proposed vision-based 3D interface does not distract the user since it tracks the movement based on invisible 3D boxes exploiting depth information estimated using multiview images/video. In addition, the difficulties in gesture analysis, mapping of the movement features to meaningful symbols, can be reduced by adopting the proposed time-delay neural networks. Consequently, the proposed framework can be easily used in every place. The framework will provide interactive systems, such as ATR I-cubed Tangible Music System or ATR Interactive Dance system, with improved quality of the 3D interface. In addition, the proposed simple framework can be extended to other applications requiring dynamic gesture tracking and analysis on the fly.

This paper is organized as follows. In Section 2 we explain how to initialize and design box-based 3D interface, given segmented user with depth information. In Section 3, given 3D boxes in personal space, the low-level features extraction and movement analysis schemes are explained. Explanation on applications exploiting the proposed 3D interface and possible extension of the proposed framework are given in Section 4 and 5, respectively.

2. VISION-BASED 3D INTERFACE EXPLOITING 3D BOXES

2.1. User Segmentation and 3D Box Initialization

The first step is to segment the user from the natural background scene, in order to simplify 3D gesture analysis. In general, object segmentation from natural scenes is still considered to be an open problem [6–8]. Even though the user can be separated from background using bluescreen techniques, in this paper we are not considering such a special environment. Instead, the used scheme segments the user from natural scene, without using bluescreen, by exploiting multiple cues such as intensity (or color), edge, motion and depth [9, 10].

There has been a growing interest in object segmentation to provide various applications with object-based functionalities such as interactivity and scalability. In spite of tremendous progress in computer vision, object segmentation schemes have lagged behind mainly because of the limited usage of visual cues (*e.g.* using only intensity or both intensity and motion information). The difficulties of automatic segmentation mainly come from defining *semantically meaningful areas* because the definition of *meaningful areas* itself is not so clear in many cases [11]. Even in cases where we do have a clear definition about the areas of interest, an accurate and reliable segmentation is a challenging and demanding task because those areas are not homogeneous with respect to low-level features such as intensity, color, motion, disparity, etc.

To alleviate these difficulties, several hybrid schemes have been proposed [12–15]. Among these schemes, motion information has been widely accepted as a crucial cue, under the assumption that the objects of interest (OOI) can be characterized by a coherent motion. Note, however, that this scenario only works for the case in which the OOI have motion in the scene. Even in the case of moving objects, motion similarity may not work well due to various error sources including occlusions and inaccurate motion estimation. Therefore, additional information is necessary to detect accurate boundaries of objects.

We use a robust segmentation scheme jointly exploiting color (or intensity), edge, motion and disparity information [10, 16]. The proposed moving object segmentation scheme consists of three steps, which are (1) static background scene capture and estimation of its statistics (2) disparity estimation (3) moving object segmentation from natural scene. We first capture a static background scene and estimate corresponding statistics for each pixel in the image,

$N(m, \sigma)$, where m and σ denote mean and standard deviation, respectively. The statistics $N(m, \sigma)$ are used as threshold values in the initial segmentation process. Next, we estimate a smooth disparity of current scenes [8–10, 16–19]. Finally, we segment moving objects from the static background scene by comparing intensity (color), edge, motion and disparity. To segment only objects of interest we assume that these objects have a limited range of disparities, *i.e.* depth, and that the disparities with the objects are smoothly changing. The segmented object containing depth information is ready to be used in gesture analysis and then mixed into another image/video using z-keying.

According to our experimental results [10], the proposed hybrid segmentation scheme efficiently separates the user not only from bluescreen but also from real scene, as shown in Figure 3. Note that the obtained depth information of the segmented object is exploited in 3D gesture analysis and then z-keying with the pre-captured background video or rendered 3D virtual environment [16].

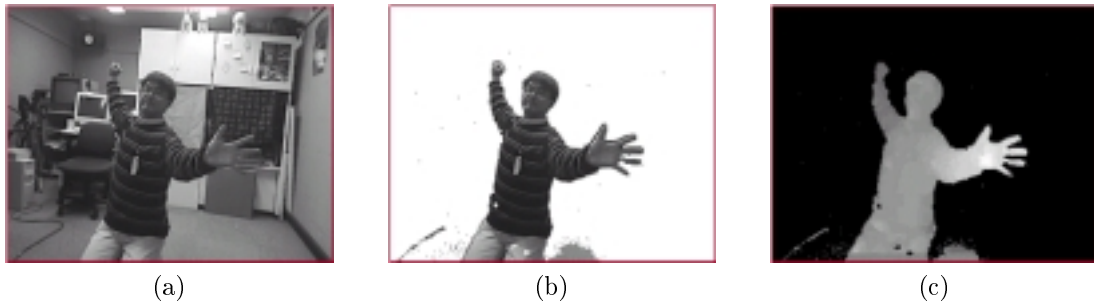


Figure 3. Object Segmentation from natural scene. (a) image with a static background (b) segmented object image (c) disparity image of the segmented object. The size of the image is 320×240 .

Given a segmented user with depth information, we estimate the center of the segmented body, $C_0(x_0, y_0, z_0)$, in 3D space. The 3D coordinate C_0 is considered as the center of the user’s personal space. We allocate invisible 3D boxes to the space surrounding the user, C_0 , to track movement in 3D space. In this experiment, we adopt eight 3D boxes in the personal space. The more 3D boxes, the more accurate the tracking of movement will be. However, there is a tradeoff between accuracy and computation time because the increased latency or time-delay caused by an increased number of boxes will distract the user.

2.2. Design of box-based 3D Interface

Given the segmented user with depth information, a personalized 3D interface is designed for tracking the body movement in 3D space on the fly. As explained, a simple 2D interface can be used to track and analyze dynamic gestures in real video-rate (10-30 Hz) [2–4]. However, the cost of saving processing time is the loss of the tracking accuracy. As a substitute for the 2D interface, we adopt what is called a box-based 3D interface, which improves tracking accuracy while maintaining its simplicity. The proposed box-based interface is focused on body movements in the personal space. Note however that the proposed interface can be used in tracking/analysing movements in both the personal and general space. For example, we can keep tracking the body movement in the general space by recording the movement of the body center in the 3D space .

It is natural to use 3D space as an integral part of communicating with computers. The user can use his/her body in 3D space the way a painter uses his/her paint brush on paper. The body movement, as user moves through space, creates movement patterns in the air. These patterns of movement in 3D space have a special meaning or reason of the movement. Like dances, such dynamic gestures in 3D space can be considered as choreographed sequences and thus analyzed based on choreography. In choreography, space refers to where the body is moving. There are two kinds of 3D space: general space and personal space. General space is the space in which the user travels around. Meanwhile, personal space (or kinesphere) is the space which the user can reach in any or all directions, while standing in place without traveling, much like a personal bubble surrounding the user. Thus, the user takes his/her personal space with him/her as he/she travels through general space.

As explained previously, as a tradeoff between accuracy and processing time, we chose to allocate eight 3D boxes into the space surrounding the user. The top view of the general space, as shown in Figure 4, shows how the 3D

boxes are initialized to the user’s personal space. First, we measure the width ($2x$) and height (h) of the segmented user while he/she is standing with their back straight and their arms against their sides. The size of the other 3D boxes are determined relative to the values of x and h . Then we allocate eight normalized 3D boxes into the personal space in order.

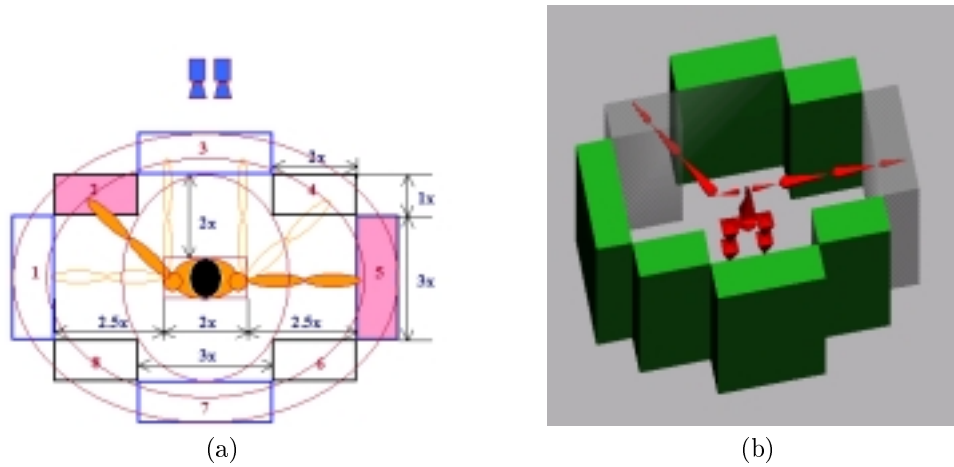


Figure 4. Basic structure of box-based 3D interface. (a) top view (b) side view. The top view of the general space shows how the 3D boxes are initialized in the personal space. The eight 3D boxes are allocated into the personal space in order, based on the width of the segmented user, *i.e.* $2x$.

Let the state of the 3D box be $S = \{s_0, s_1, \dots, s_N\}$, where N denotes the number of 3D boxes surrounding the user in his/her personal space, as well as the 3D box directly enclosing the user. The status of i -th box, s_i , is denoted by 1 or 0, where 1 indicates that the user is touching the box when it is observed. Unlike other vision-based approaches, we keep track of the state of 3D boxes, S , to track body movement, instead of tracking the movement of all the user’s limbs such as arms, legs, feet and torso. We detect whether a 3D box is touched by the user by observing the state of the 3D boxes, S . Simultaneously, as shown in Figure 5, the position of the body within a touched 3D box, $p_i(x_i, y_i, z_i)$, is estimated by calculating the mean coordinate of the object within the 3D box.

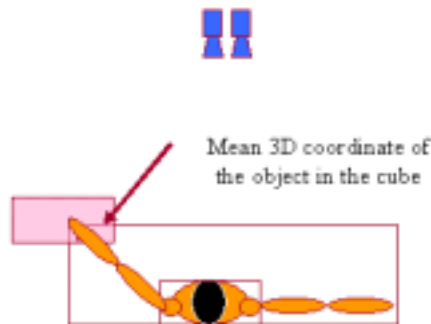


Figure 5. Position measure. The center position of the object within the touched 3D box is estimated by calculating a mean coordinate of the object within the 3D box.

Given the segmented participant with depth information and the box-based 3D interface, the body movement can be tracked by observing how the body moves through the eight 3D boxes in the personal space. For example, several low-level features (gesture, velocity, and acceleration of the movement) are extracted as corresponding to the movement features. In this scenario, the tracking of body movement in the personal space is relatively easy compared to other schemes tracking each part of the body using markers, sensors or geometrical models.

3. FEATURE EXTRACTION FROM BOX-BASED 3D INTERFACE

3.1. Movement Theory

In this section, we introduce basic elements of movement based on Laban’s movement theory [20]. According to choreography, the types of body movements in 3D space include locomotor (traveling) and non-locomotor. The locomotor denotes the movement of traveling from place to place, usually identified by weight transference on the feet such as basic movement (e.g., walk, jump, slide, roll, run, leap, hop) and combined movement (e.g., step-hop, waltz-run or triplet). Meanwhile, the non-locomotor denotes any movement that occurs in one location in space using the available space in any direction (e.g., curl, stretch, spin, etc.) or movement organized around the axis of the body (e.g. bending, twisting, stretching, swinging, etc.), rather than designed for travel from one location to another.

Such movement can be analyzed by Laban Movement Analysis (LMA), a method for observing, describing and interpreting human movement in 3D space[§]. In general, LMA is considered as a system for observing, analyzing, and classifying movement. The LMA has been used by diverse ranges of groups and individuals including researchers, dance performers, athletes, actors, therapists and educators. The LMA encompasses four main categories: *BODY*, *EFFORT (dynamics)*, *SHAPE*, and *SPACE* [20].

The *BODY* aspect deals with principles such as the initializing and sequencing of movement from different parts of the body, and the connection of body parts to each other. The *EFFORT* dimension is concerned with movement qualities and dynamics, or how someone moves as opposed to what they do. The *EFFORT* factor is subdivided into four elements: weight, space, time and flow. *SHAPE* is about the way the body interacts with its environment, not just where they move. There are three types of shape change: shape-flow (growing and shrinking, folding and unfolding, etc.), directional (spokelike or arclike) or shaping (molding, carving and adapting). The *SPACE* involves the study of moving in connection with the environment and is based on spatial patterns, pathways and lines of spatial tension.

Most widely used in gesture synthesis/analysis is his concept of *EFFORT/SHAPE*. We also focus on the “*EFFORT*” and “*SHAPE*” aspects, which are concerned with movement qualities and dynamics. As explained previously, according to Laban’s Effort Theory, the movement in 3D space, including dynamic gestures such as dance, can be analyzed by disassembling the movement into basic elements of movement such as {space, time, weight, flow}. The principles of the basic elements are no less important for the body movement than color, line and form for a painter.

The element *space* measures the size (magnitude of body movement on a continuum of small/large, closed/open or direct/indirect) to make framework simple. The element *time* refers to how the body moves in relation to time (temp), which measures the speed (a continuum from very fast to very slow) with which a movement travels spatially. The element *weight* measures energy along a light to strong (or heavy) continuum. The element *flow* measures the rhythm.

In general, the element *space* is one of the key features in static gesture recognition because the *space* also measures direct or indirect body movement in the space. A direct movement can be likened to an arrow traveling straight to its target, i.e., the use of space is economical and restricted. Meanwhile, an indirect movement means a flexible curving, roundabout and plastic movement. In addition, the *space* may be described in terms of various other factors according to applications:

- *Directions*: forward, backward, sideways, diagonal, upward or downward.
- *Focus*: where the eyes or the intention of the movement is directed, or the body is facing.
- *Levels*: high, middle and low or deep, e.g., on floor, kneeling, elevation.
- *Plane*: horizontal, vertical, sagittal
- *Pathways*: the patterns or designs made in the air or on the floor by the person’s movements, e.g. zigzag, curved, spiral, circle, straight.

[§]R. Laban (1879-1958), known as the father of modern dance theory, made many contributions to modern dance and dance therapy. He developed a detailed system of describing movement called Labanotation, used to record movement and choreography. He studied and developed choreutics, “the relationship of the body to space around it”, eukinetics, “the formulation of all possible types and directions of bodily movement”, and “the connection between psychology and motion.”

- *Shape*: the design of the body’s position
- *Size*: the magnitude of the body shape or movement.

The number of possible combinations and permutations of the elements is virtually endless. For example, as shown in Figure 6, the combinations of {space, time, weight} creates eight basic movements. The eight basic movements include floating, flicking, slashing, wringing, pressing, gliding, dabbing and thrusting or punching [2, 20].

- *Floating*: {INDIRECT, SLOW, LIGHT} - Flexible,
- *Flicking*: {INDIRECT, SLOW, HEAVY} - Flexible, Sustained, Strong Bound flow.
- *Slashing*: {INDIRECT, FAST, HEAVY} - Sudden, Strong, Flexible.
- *Wringing*: {INDIRECT, FAST, LIGHT} - Flexible, Sudden, Free flow.
- *Pressing*: {DIRECT, SLOW, HEAVY} - Sustained, Strong, Bound flow.
- *Gliding*: {DIRECT, SLOW, LIGHT} - Sustained.
- *Dabbing*: {DIRECT, FAST, LIGHT} - Direct, Sudden.
- *Thrusting or punching*: {DIRECT, FAST, HEAVY} - Sudden, Strong. Usually performed with free flow but can also be performed with bound flow. Sustained.

The dynamic gestures (or action) can be considered as movement phrase, *i.e.* the combination (or permutation) of such basic movements. Through the sequence of dynamic gestures the users express their intention or emotion. For example, dynamic gestures in modern dance can be classified into seven motives [4, 10]. In case of general gestural expression, obviously, the expression can be classified into three groups of emotions, {positive, neutral, negative}. As shown in Figure 6, the positive and negative expressions, respectively, further can be classified into {Peace/Love, Joy/Happiness, Surprise} and {Anger, Sadness, Fear} which consist of “seven basic emotional categories” with {Neutral}:

- *Peace/Love*: {Floating, Gliding} - quiet, gentle, open and positive, slow, dreamy state.
- *Joy/Happiness*: {Gliding, Dabbling, Flicking} - openness, lively, prancing all about
- *Surprise*: {Wringing} -
- *Anger*: {Pressing, Punching, Slashing} - storming around, much fast, violent movement, punching, kicking, temper tantrum
- *Sadness*: {Floating} - sulking, closed, little movement.
- *Fear*: {Wringing, Slashing} -
- *Neutral*: { } - no movement at all.

3.2. Feature Extraction and Emotional Mapping

In this section, we describe how dynamic gestures or sequence of dynamic gestures can be recognized. As explained, action or dynamic gesture, consists of eight basic movements that all consist of four basic elements of movement. Conversely speaking, dynamic gesture (or action) can be analyzed by disassembling the movement phrase into basic movement and further into basic elements of movement.

In the proposed framework, we decompose the dynamic gestures into basic elements of movement through the proposed 3D box-based interface. As shown in Figure 2, we first segment the user from the background and then track the movement in the personal space. Then, we track the movement and extract corresponding local features, based on Laban’s Theory [20], using eight 3D boxes. We measure the quantity of dynamic movement based on the four basic elements of movements, such as space (open or closed), time (fast or slow), weight (heavy or light) and flow. In case of the proposed box-based analysis, we define four elements of movement as follows.



Figure 6. Eight basic elements of movement. The combinations of based on {space, time, weight} creates eight basic movement, which include floating, flicking, slashing, wringing, pressing, gliding, dabbing and thrusting (or punching).

- *space*: openness, *i.e.* the shape of the gesture can be measured by counting the number of boxes touched simultaneously and which boxes are touched.
- *time*: velocity, *i.e.* the number of touched boxes in a given time period.
- *weight*: acceleration, *i.e.* the number of different boxes touched in a given time period.
- *flow*: voting based on movement history, *i.e.* keep previous movement for a few seconds (3-5 sec) and refer them to analyze the rhythm of the movement.

After extracting the quantity of movement, we analyze the quality of movement by mapping the low level features to meaningful symbols such as intention or emotion. To assign intention of movement from low-level features, we adopt a time-delay neural networks (TDNN). It has long been postulated that neural networks might provide the most sound basis for approximating any (linear or nonlinear) function with a finite number of discontinuities [21]. In particular, multi-layer perceptron (MLP) might provide the most valid way to map any nonlinear relation, given sufficient neurons in the hidden layers. After proper training on a representative set of input and output vectors, MLP tends to lead a new input vector (that the MLP has never seen) to a similar output (to the correct output for the close input vector used in training).

However, the MLP may have limitations in applying for applications requiring time-dependent nonlinear mapping such as dynamic gesture or emotion analysis. These problems can be alleviated by adopting time-delay to the input of MLP (TD-MLP). Note however that there is a tradeoff between accurate analysis and proper response time. The delayed response may distract the user, especially when it is used in real time applications [3].

The proposed TD-MLP consists of three layers including input, hidden and output layers [3]. The adopted TD-MLP performed in two step, *i.e.* learning and then mapping. First, TD-MLP learns the nonlinear relationship between the local features and the user’s intention (or emotion) through examples, rather than complicate motion analysis and mapping. After proper training, the TD-MLP associates the time-delayed input vectors with specific emotional category, *i.e.* categorizes the dynamic gestures to emotional categories, in real time.

In this framework, a set of local features of personal space (*i.e.* space, time, energy and flow) are mapped directly to a set of symbolic representation of emotion *i.e.* {Peace/Love, Joy/Happiness, Surprise, Anger, Sadness, Fear, Neutral}, as input and output vectors of the TD-MLP. Note that alternatively, the proposed 3D box-based interface can be used to make the analysis procedure even simpler. For example, as shown in Figure 7, without

extracting elements of movement, the state of 3D boxes, $S = \{s_0, s_1, \dots, s_8\}$ can be directly connected to the input layer of TDNN. The status of i -th box, s_i , is denoted by 1 or 0, where 1 stand for that the user is touching the box when it is observed. Also, the relative height of object within the box, $\{0, \dots, 1\}$, can replace binary input S and be trained to be mapped to the proper output categories. In all cases, the time-delay factor allows incorporating the element *flow* into interpretation of the movement [3].

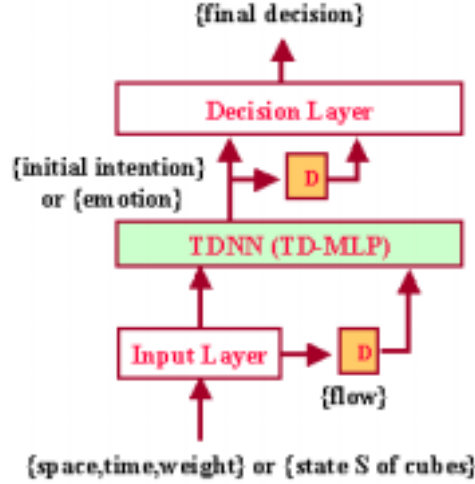


Figure 7. Basic structure of movement analysis using time-delay neural networks. "D" represents a time-delay. The carefully selected elements of movement, {space, time, weight} can be a set of input. Also, the state of 3D boxes $S = \{s_0, s_1, \dots, s_8\}$ can directly be a set of input of the TDNN. The status of i -th box, s_i , is denoted by 1 or 0, where 1 stand for that the user is touching the box when it is observed. In all cases, the time-delay factor allows incorporating the element *flow* into interpretation of the movement.

As shown in Figure 7, the time-delay factors are also used in output stage to make the output consistent. The more delay, the more consistent analysis results we get. However, we need a balance between consistency and latency of the system. In the experiment, we choose 3-5 delays in both input and output layers.

4. POSSIBLE APPLICATIONS: GESTURE-BASED MUSIC PLAY

Recently, gesture-based interfaces have been applied for music performance as well as for a comfortable and natural communication medium between human and computer. Obviously, interpreting and responding to non-verbal dynamic gestures is quite important in music performances. Since 1970s, there have been good initial framework mappings between specific gestures and their intended musical results. For example, in 1970, a conducting system based on the "Generating Realtime Operations On Voltage-controlled Equipment(GROOVE)" using knob inputs was developed by M. Mathews, a distinguished pioneer in computer music (including invention of sound synthesis) and gesture-based input for human-computer interaction [22,23]. A computer music system that follows a human conductor, was designed by Morita et al [24]. The electronic orchestra system, with a complex performance database and MIDI controllers, responds to the gestures of a conductor through a CCD camera and a sensor glove.

The proposed box-based 3D interface also can be used for music performance without exploiting complicated sensor-based motion tracking systems such as gloves or suits. To show the effectiveness of the proposed vision-based 3D interface, we apply the box-based 3D interface into "ATR I-cubed Tangible Music System (ITMS)[¶]." In the system, as shown in Figure 3, we first segment a user from natural background and measure the distance between the camera and the segmented user, using a multiview camera (*e.g.* TriClops) in front of the user. As shown in Figures 4, eight 3D boxes are allocated to extract body movement information in the personal space, without complicated motion capture facilities. Using the estimated depth information and the center body-coordinate of segmented user, we initialize and divide the personal space using eight 3D boxes to track the user's movement.

In ITMS, each 3D box is assigned to a different musical instrument and the corresponding music is played when the 3D box is touched by the user. Therefore, music with different instruments can be played according to the user's dynamic gestures. Several instruments can be played simultaneously according to the user's gestures. We also control the volume of the instrument by measuring the mean height of the object in the touched box, as shown in Figure 5. Numerous insights have been gained from building and testing the ITMS, and a preliminary theoretical framework will be the basis for future works in this area.

[¶]the ITMS was demonstrated during the 13th ATR Research Exposition, November 1-2, 2000, Kyoto, Japan.

The demonstrated ITMS can be extended to “I-cubed Visual Music System (IVMS)” by adding gesture/emotion analysis scheme proposed in Section 3. The IVMS integrates music performance with emotional visualization by allowing the user interactively reflecting user’s intention onto virtual environment. The user’s movement is analyzed based on the proposed analysis scheme, *i.e.* local features corresponding to {time, space, weight} are extracted and then mapped onto predetermined emotional categories, *i.e.* {Peace/Love, Joy/Happiness, Surprise, Anger, Sadness, Fear, Neutral}. The resulting emotional information and dynamic gestures are used to construct a virtual environment and to interactively control virtual objects in the environment. Thus the resulting virtual environment provides the user with immersive multimedia experience by playing music and interactively reflecting the emotional factors onto virtual environment on the fly.

5. SUMMARY AND DISCUSSION

We proposed a simple and robust dynamic gesture analysis and interpretation framework exploiting non-contact vision-based 3D interface. In this paper, we showed that robust motion tracking in 3D and its analysis can be achieved on the fly, without using complicated sensor-based tracking systems. The main contribution of the proposed non-contact vision-based framework is in providing both simple feature extraction and robust personal space analysis tools. The box-based feature extraction scheme helps extract local features without adopting complicated motion capture facilities. In addition, it can analyze the personal space without tracking detailed part of human body such as head, hands or feet.

The proposed non-contact vision algorithm is more cost-effective than available motion capture/tracking systems. Due to its simplicity and robustness of the framework, the proposed framework will enable wide-ranging applications requiring dynamic movement analysis on the fly. For example, it can be extended to the applications such as dancing, because dance is considered a typical way to deliver a dancer’s intention through distinctive user-controlled gestures [1, 3, 4, 20]. Furthermore, we can analyze the intended emotion of the user by adopting time-delay neural networks (TDNN) [9]. In addition, we explained how a consistent emotional analysis can be achieved based on time-delay neural networks. The difficulties in analyzing nonlinear relationships between movement and emotional intention can be alleviated by mapping low-level features extracted from 3D boxes in the personal space to predetermined emotional categories, by training.

One of the remaining challenges is combining statistics into a 3D dynamic gesture model to achieve a more natural 3D interaction in real-time. Another challenging task is to add other perceptual cues to build a more natural and robust interface, e.g. making the computer aware of the participants voice and face as well as whole body.

REFERENCES

1. A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Chmpbell, Y. Ivanov, A. Schutte, and A. Wilson, “The KidsRoom: A perceptually-based interactive and immersive story environment,” *Presence: Teleoperator and Virtual Operator* **8**, pp. 367–391, Aug. 1999.
2. A. Camurri, M. Ricchetti, and R. Trocca, “Eyesweb-toward gesture and affect recognition in dance/music interactive system,” in *Proc. IEEE Multimedia Systems*, pp. 643–648, June 1999.
3. W. Woo, J. Park, and Y. Iwate, “Emotion analysis from dance performance using time-delay neural networks,” in *Proc. IEEE JCIS-CVPRIP*, pp. 374–377, 2000.
4. R. Suzuki, Y. Iwate, M. Inoue, and W. Woo, “MIDAS: MIC Interactive Dance System,” in *Proc. IEEE SMC*, pp. 751–756, 2000.
5. A. Mulder, *Human movement tracking technology*, Simon Fraser University: Technical Report 94-1, 1994.
6. B. Horn, *Robot Vision*, The MIT Press, 1986.
7. R. Franich, *Disparity Estimation in Stereo Digital Images*, Ph.D. thesis, TUDelft, 1996.
8. H. Jeong, W. Woo, C. Kim, and J. Kim, “A unification theory for early vision,” in *Proc. First Korea-Japan Joint Conf. on the Computer*, pp. 298–309, Oct. 1991.
9. W. Woo and Y. Iwate, “Object-oriented hybrid segmentation using stereo images,” in *Proc. SPIE IVCP*, pp. 487–495, Jan. 2000.
10. W. Woo, N. Kim, and Y. Iwate, “Object segmentation for z-keying using stereo images,” in *Proc. WCC*, pp. 1249–1253, Aug. 2000.
11. M. Kunt, A. Ikonomopoulos, and M. Kocher, “Second-generation image coding techniques,” *Proc. of the IEEE* **73**, pp. 549–574, Apr. 1985.

12. M. Chang, M. Tekalp, and I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. on IP* **6**, pp. 1326–1333, Sept. 1997.
13. R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," *IEEE Trans. on CSVT* **8**, pp. 526–571, Sept. 1998.
14. E. Francois and B. Chupeau, "Depth-based segmentation," *IEEE Trans. on CSVT*, pp. 237–239, Feb. 1997.
15. E. Izquierdo, "Image analysis for 3D modeling, rendering and virtual view generation," *CVIU* **71**(2), pp. 231–253, 1998.
16. N. Kim, W. Woo, and M. Tadenuma, "Photo-realistic 3d virtual environment using multiview video," in *Proc. SPIE VCIP*, Jan. 2001.
17. W. Woo and A. Ortega, "Stereo image compression based on the disparity compensation using the MRF model," in *Proc. SPIE VCIP*, vol. 2727, pp. 28–41, Mar. 1996.
18. W. Woo and A. Ortega, "Stereo image compression based on the disparity field segmentation," in *Proc. SPIE EI-VCIP*, vol. 3024, pp. 391–402, Feb. 1997.
19. W. Woo and A. Ortega, "Modified overlapped block matching for stereo image coding," in *Proc. SPIE EI-VCIP*, vol. 3653, Jan. 1999.
20. R. Laban, *Modern Educational Dance*, Trans-Atlantic Publications, Inc., 1988.
21. D. Rumelhart and J. McClelland, *Parallel distributed processing: Explorations in the microstructure of cognition*, MIT Press, 1986.
22. R. Boulanger, "Conducting the MIDI orchestra," *Computer Music Journal* **14**(2), pp. 34–46, 1990.
23. L. Spiegel, "Graphical groove: Memorium for a visual music system," *Organised Sound* **3**(3), pp. 187–191, 1998.
24. H. Morita, S. Hashimoto, and S. Ohteru, "A computer music system that follows a human conductor," *Computer Magazine* **24**, pp. 44–53, July 1991.