The Journal of DISPLAYS, Vol. 15, No. 3, pp. 141-148, July 1994, UK.

IMPROMPTU CONDUCTOR: A VIRTUAL REALITY SYSTEM FOR MUSIC GENERATION BASED ON SUPERVISED LEARNING

Rung-huei Liang, Ming Ouhyoung
Communications & Multimedia Laboratory, Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, ROC

ABSTRACT

This paper presents the "Impromptu Conductor" virtual reality system which combines computer graphics and music while emphasizing on computer music. We introduce the supervised learning method mentioned in the field of pattern recognition into the reproduction and the organization of music.

We have proposed and implemented a practical way of capturing human's motion to create music as well as generating the corresponding images on a screen by using a 6-D tracker to simulate a conductor's hand in real time. However, the mapping between music and hand motion in our system is not a simple one to one function, and is constrained and properly modified by music styles collected from supervised learning. That is, the music style produced by an interactive user strongly depends on the movement of one's hand gesture. The system we implemented shows that the system designer can give different styles of feedback to different patterns of a user's behavior easily.

Keywords: virtual reality, computer music, computer graphics, pattern recognition.

1. Introduction

We have established a virtual environment in which there are several kinds of musical instruments set up in the virtual space(Plate 2). A user will feel as if there are instruments in the space and can use a 6-D tracker to activate or touch them, while music is produced simultaneously[1](Plate 1). This real time system simulates an impromptu drummer by presenting two hands holding drumsticks on a computer screen while letting one manipulate ten musical instruments and some kinds of chords. As depicted in Fig. 5(a), this system employs a simple mapping function from the geometry of the 6-D tracker to a unique sound of musical instrument. One shortage of this system is that all sounds are arbitrary in their pitches, in another word, there are no music styles and background music patterns, as indicated by Chabot[7].

In a user's point of view, the sound produced by his own interaction with a computer is, speaking in two totally different views, either all by the system's design[10] or arbitrary by his action[1,7], without any coordination and organization. We hope that there are styles in the produced music, inviting the user to do some degree of interaction. In general, there is always a strong relation between the motion of the music conductor and the conducted music. Since we want to simulate a conductor's hand motion, the music style must be dependent on his behavior. Because the supervised learning method is convenient in classifying a new coming sample into a proper class in the assumption that there exists a supervised learning set, we first divide music into several classes. Based on the hypothesis

that one's behavior is strongly related to the music style he prefers, we conducted an experiment which got one training sample from each subject's behavior. After supervised learning, an arbitrary new coming user's behavior can be automatically classified into one class corresponding to a specific music style. Details are given in sections 5 and 6.

Once we have a large representative training set and a classification method, a user will produce Blues as he "acts" like a blues fan, for instance, and can produce classical as he "acts" like those who prefer classical. One advantage is that one can always make music according to his own action, in which he might not be aware of the corresponding music style embedded, even if he doesn't have any experience in playing musical instrument. Therefore, the sound generated is not just "noise" anymore, and music styles and harmony are already blended in.

2. Music Classification and Measurements

According to traditional classification[2,3], we can divide music into these classes: Medieval, Renaissance, Baroque, Classic, Romantic, Modern, Blues, Jazz, Rock, Pop. Then statistical work of every class of music is collected in the existing available music score in unit of every note considering the following measurements: average pitch, average duration, weighted divergence to the chord, average distance to the previous note and chord progression. Next step is to generate typical motifs[5,6] which meet those measurements for every corresponding class respectively. Currently, we use a predesigned motif for every class.

3. Supervised Learning

The Fisher approach[4] proposed in pattern recognition is based on the projection of d-dimension data onto a line. The goal is that these projections onto a line will be well separated into classes. To simplify, we use this approach with two classes, Classical and Blues, in our system currently. The training set

$$H = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\} = \{H_1, H_2\} \quad (1)$$

is obtained from two classes of subjects' hand motion records, where H_1, H_2 mean two classes. Our goal is to find a $\underline{w}$ that maximize

$$J(\underline{w}) = (\underline{w}^T S_B \underline{w})/(\underline{w}^T S_W \underline{w}) \quad (2)$$

where $\underline{w}$ is a line's direction we want to project onto, S_B is the *between-class scatter matrix* and S_W is the *within-class scatter matrix*. The numerator of (2) may be understood as:

$$\underline{w}^T S_B \underline{w} = \underline{w}^T (\underline{m}_1 - \underline{m}_2)(\underline{m}_1 - \underline{m}_2)^T \underline{w} = (m_1 - m_2)^2 \quad (3)$$

where $\underline{m}_1$ and $\underline{m}_2$ are the mean vectors of H_1 and H_2 respectively ,and m_1 and m_2 are the projection of $\underline{m}_1$ and $\underline{m}_2$ on the line $\underline{w}$ respectively.

and the denominator may be thought as variance of within-class data. The criterion Eq. 2 is the ratio (difference of means)/(variance of within-class), and to maximize it is to achieve the well-separated projection. Forming $\partial J / \partial w = 0$, we will use the result in[4]:

$$\underline{w} = S_W^{-1}(\underline{m}_1 - \underline{m}_2) \quad (4)$$

4. System Configuration

The following are our system requirements in hardware: an SGI Indigo^2 Extreme, an IBM PC486, an MIDI interface MPU401, a tone generator (*Yamaha*'s TG100), and a 6-D tracker(Ascension Co.). The system configuration is given in Fig. 1. We used assembly language for device drivers such as MIDI[8,9](Musical Instrument Digital Interface) driver, and C language for supervised learning and motion-to-music transformation.

Our system is consisted of two parts: real-time image rendering and music generation. We use an SGI Indigo^2 Extreme to generate image and an IBM PC486 to generate music, while an RS232 line is used to build connection between them. A user wears stereo glasses (Crystal Eyes) and watches the screen to get stereo image, while wearing a headphone with a tracker attached on top of the headphone to hear the music and to change the view point(Fig. 1). The following are short descriptions of two applications based on this system.

4.1. Virtual Drum System

This system simulates a drummer playing a drum set in front of the screen(Plate 1). One interacts with this system by controlling two drumsticks with trackers attached. The images of drumsticks will be updated as the real drumsticks move and the sound of the corresponding drum piece will be generated simultaneously if a collision happens between any drumstick and a drum piece(Plate 2.). Three sets of tracker data are sent to the Indigo^2, which performs the collision detection from those of drumsticks and then generates the corresponding image while translating the collision information into MIDI codes to an IBM PC for sound generation. It is not quit straitforward to detect a collison in the space, however, we have tried two different methods for collision detection: maximal deceleration and bounding box.

In the first method, we take the behavior of drumstick in playing into consideration. General speaking, if a drumstick hits the surface of a drum, it will bounce back (Fig. x). What we want is to detect the lowest point in the path of a moving drumstick which is performing a hit on the drum. The offline analysis of the path of the 6-D tracker attached on drumstick is simple, but the key problem is to know the user's intention, whether to hit or just to move the drumstick, in real-time. To achieve this goal is to detect the occurrence of a maximal deceleration and recognize it as a hit immediately. However, our solution is simply to memorize the last two locations reported from the 6-D tracker and compare them with the current one. In Fig. XXX, both (a) and (b) are recognized as hits immediately after the current position *C* is available, based on the condition that vector *AB* is greater than a reasonably large positive threshold and vector *BC* is smaller than a reasonably small positive threshold. The reason of all above is to avoid the jittering phenomena of the 6-D tracker. The second one is much simpler. If the drumstick moves accross the top surface of the bounding box of a drumpiece, a collision happens. This method only keeps track of the last one position and the current one (Fig. YYY). A constrain of point *A* and *B* being separated by a threshold away from the drum surface also avoids the jittering.

Because we think that the crash cymbal being hit is an important change in a piece of music, we switch the key of the background music chord to the relative perfect forth of

current key if it is hit. In the same way, a ride cymbal being hit will produce a sustained chord of current key. Hitting the floor tom-tom will generate a root of the current key playing on the electronic bass, and the rest pieces of drums played will produce a randomly selected note of current key on piano. All the music heard above is generated through MPU401 and TG100.

4.2. Molecular Docking System

Another application is in scientific visualization of molecular binding[11]. We have developed a molecular docking system, which shows the binding of molecules and allows a user to hold a tracker to move and rotate one drug molecule to fit into another molecule(around the receptor site). We provide two important cues to improve the visualization: stereo images and audio. Stereo images provide complete information in three dimensions while audio provides the current status of docking. We divide the docking into four status and give corresponding sound respectively: far, collision, near without collision and fit. One can easily know the current status without focusing on the images to inspect details of the molecule. We find that this system facilitates the docking procedure.

The above two systems have their specific purposes and do not really have styled music. In the following section, we give an experiment that shows how we can embed style into music.

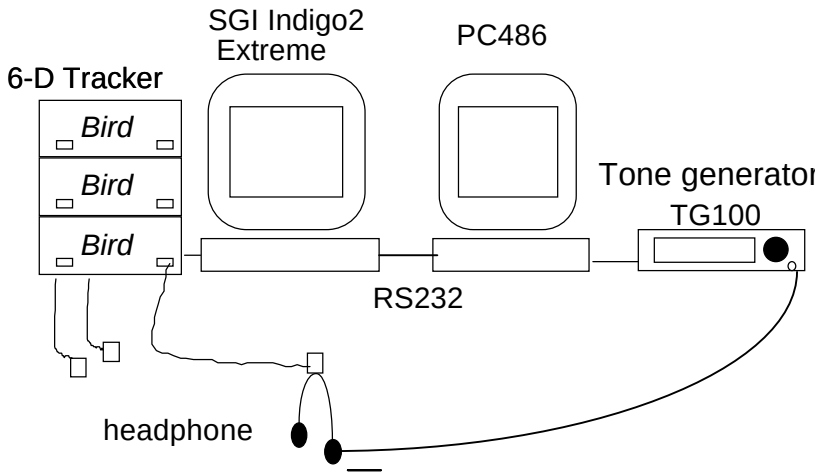


Figure 1. System configuration. A Tone Generator (TG100) is connected to a PC with MPU401 interface and three 6-D trackers are connected to the serial port of an SGI Indigo^2 Extreme. An SGI and a PC are connected via RS232.

5. Style Music Experiment

5.1. Subjects

Eight young adults served as volunteers. Six of them are graduate students and two are undergraduates. All of them had normal hearing capability and were naive regarding the purpose of the experiment. Four of them preferred Classical music while the others preferred Blues.

5.2. System Implementation

As for now, we divide music into only two categories: Blues and Classical, and their motifs are based on Blues Dorian Scale and Natural Major Scale, described in Fig. 2, respectively. We regard music as two components: background chord and melody. One chord is generated at a time, which is in the form of dispersion. A constant *tempo* (68 quarter notes per minute) is kept and the members of one specific chord are distributed into quarter notes evenly. For example, in C Major, a sequence of C, E, G, E is repeated. The above is generated in the timbre of piano, furthermore, there is a root of that chord produced in the timbre of wood bass in every beginning of a bar. Currently we always employ a major chord. The background chord depends on the rotation of the 6-D tracker along *Z*-axis (*Rz*) and the pitch of the melody is on the position in *X*-axis, as depicted in Fig 3. Twist in *Z*-axis will result in changing a chord to its neighborhood in fifth-circle[1] shown in Fig. 4, that is, if rotation about *Z* axis (*Rz*) is positive and greater than 90 degrees, C major will change to F major, or F major to B♭ major, etc., and on the other hand, a negative *Rz* smaller than -90 degrees makes C major go to G major, or G major to D major, and so on. Both Classical and Blues use the same way of background chord generation, but they are different in melody generation. The more toward left the 6-D tracker is, the lower and the more left pan notes are heard, and vice versa. Details are given in the Appendix. All melodies of both types are in the tone of piano.

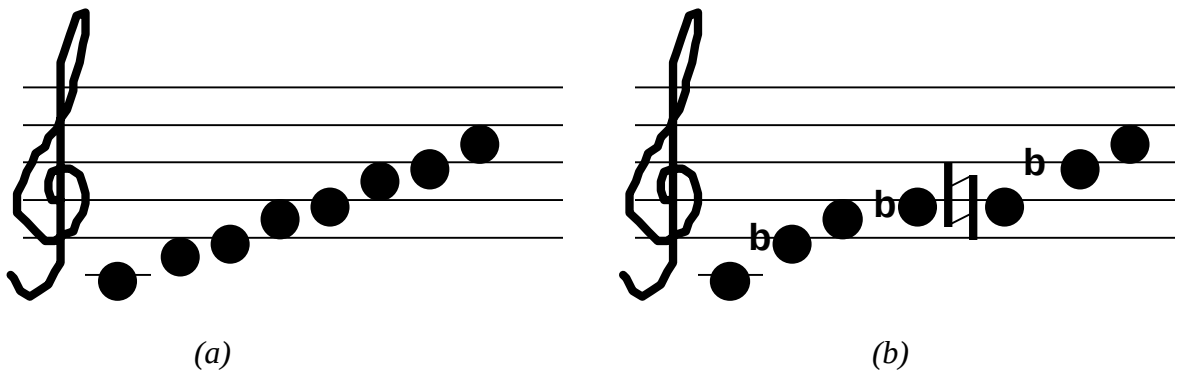


Figure 2. (a) Natural major scale and (b) Blues Dorian scale beginning on note c.

In the supervised learning phase, as depicted in Fig. 5(b), the sequence of *Rz* and the motion velocity in *X*-axis (*Vx)* of every subject is saved and then a calculation of $\underline{w}$ is performed. In the classifying phase, the music style of a new user is classified every four bars by the projection:

$$y = \underline{w}^T \underline{x} \tag{5}$$

comparing to the threshold:

$$y_T = \underline{w}^T (\underline{m}_1 + \underline{m}_2)/2 \tag{6}$$

where $\underline{m}_1$ and $\underline{m}_2$ are the mean vectors of H_1 and H_2 respectively, in the assumption that priori probabilities P(Blues) and P(Classical) are equal and the shape of their

corresponding distribution is the same. For example, if $\underline{y}_1 = \underline{w}^T \underline{m}_1$, $\underline{y}_1 > \underline{y}^T$, and $\underline{y} > \underline{y}^T$, $\underline{y}$ is classified into H_1. The exact strategy used in the learning phase, as described above, is used to reproduce music according to different classes.

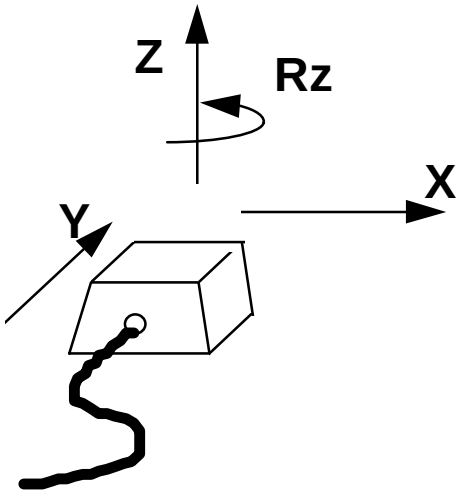


Figure 3. The 6-D tracker and its coordinate.

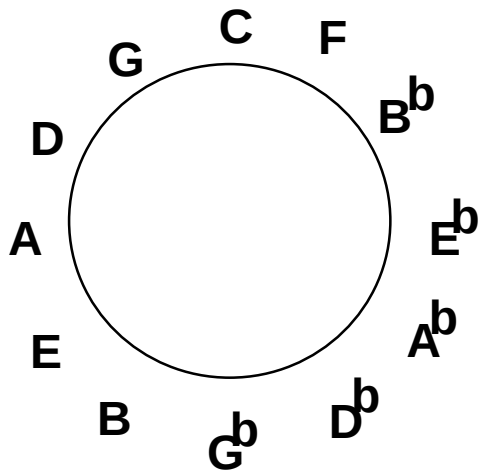
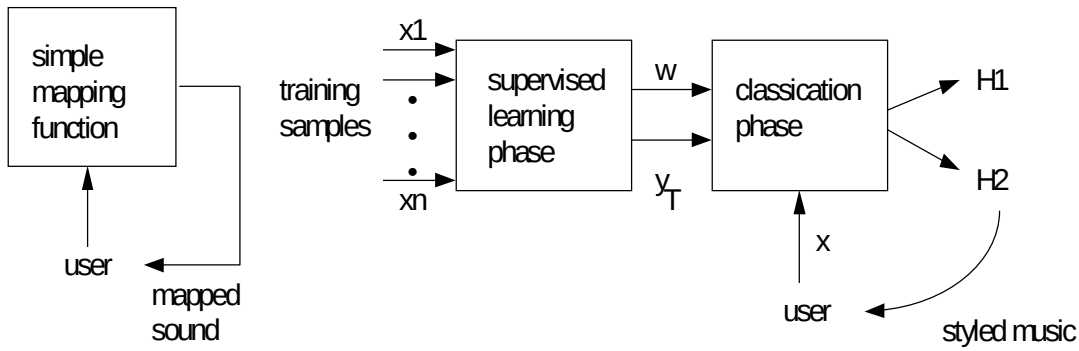


Figure 4. The fifth circle of chords.

5.3. Procedure

The experiment was conducted for each subject in a quiet environment individually. Every subject was asked what kind of music he preferred between Classical and Blues at



(a) *(b)*

Figure 5(a) the sound produced by a simple one to one function and (b) with supervised learning.

first. The experimenter recorded this and explained the usage of this system in five minutes. Then, the subject put on a headphone and held a 6-D tracker to practice and get familiar with the relation between his hand motion and the music produced. Notice that one subject was already classified into one of the two kinds of music, therefore, his behavior would be a

supervised learning sample in his corresponding type of music. Only the kind of music he preferred was produced during the whole experiment. The practicing of a subject lasted at most two minutes and a one-minute recording began immediately.

6. Result

The mean vectors of eight subjects are given in Table 1. According to these, there are two sets of learning samples, $\underline{w}^T$ = (*0.062785, 0.283614*). These data and their projection line are drawn in Fig. 6.

The simplified Fisher approach mentioned above has met the real time constraint and adds background music patterns and styles into our system with the capability of automatically classifying and reproducing music. A user not only won't be limited to accept only one music type, but also won't produce a sequence of random "noise". This reminds us that a user's behavior is an important cue in an interactive real-time system, while the learning method in the field of pattern recognition will a good solution.

Comments from subjects

"The music is really interesting", commented subject D. "The change of chords is much too obvious", said subject C.

Criticism also came from subject F, "it's oversimplified", who is a well trained musician, "Blues has its own chords progression."

	mean *Vx*	mean *Rz*		mean *Vx*	mean *Rz*
subject A	*0.175048*	*0.015905*	subject E	*0.432559*	*0.085436*
subject B	*0.114713*	*0.009029*	subject F	*0.425153*	*0.071797*
subject C	*0.196290*	*0.021911*	subject G	*0.378940*	*0.079429*
subject D	*0.097083*	*0.011643*	subject H	*0.478009*	*0.072719*

Table 1. The data on the left side are those of H_1 (Classical) and on the right side are of H_2 (Blues). Vx and Rz are from the 6-D tracker and in unit of foot and radian respectively. Note that we have taken the absolute values of Vx and Rz in the statistical calculation.

"In Jazz and Blues, the variation in rhythm is also interesting, not just in the style of syncopation.", commented subject E.

7. Conclusion and Discussion

We have established a system with $c = 2$ classes and $d = 2$ dimension. But as described in Section 2, when $c = 10$ then the dimension of the learning sample's feature vector (i.e. hand motion) should be larger in order to represent different stereotypes more precisely. It won't be a problem in time complexity of classifying as d becomes large because the advantage of Fisher's approach is that the classifying is always performed on the projected line no matter how high the dimension is.

In our experiment, there is obvious clustering phenomenon based on only eight subjects, in another word, the two sets are linearly separable. The situation will be much more complicated, maybe not linearly separable when there is a nonzero classification error; if the learning set is getting large. Although Fisher's solution might not be exactly optimal, it provides a simple and effective one. As long as most efforts are done in the supervised learning phase and a simple projection function is needed in decision phase, any other supervised learning method[4] satisfying these conditions will also meet the real time constraint and can be considered as an alternative solution.

Another way to improve the Fisher's approach is to insert the new coming user's data including his favorate music and his behavior vectors into the training set right after the classification is done. A new learning phase begins immediately to get the new projection line and its corresponding threshold for classification. The purpose of doing so is to achieve the assumption step by step that there exists a large, representative learning set mentioned in Section 1.

In Section 5, we have not considered the *tempo* as an element in music generation yet, which is an important factor of how one feels. Another important feature in music is the progression of chords, which is almost a stereotype in Blues, for example, the twelve bars Blues progression. We are still working on these.

However, the system we implemented shows a way that the system designer can give different styles to different patterns of user's behavior easily. As shown in Fig. 5(b), sufficient learning samples can be used to derive a simple projection function $\underline{w}$ and a threshold y_T can be found for music style classification. Once that these parameters are available, the system can give feedback of different styles to a user. The virtual drummer system mentioned in Section 4 is a good candidate to use supervised learning.

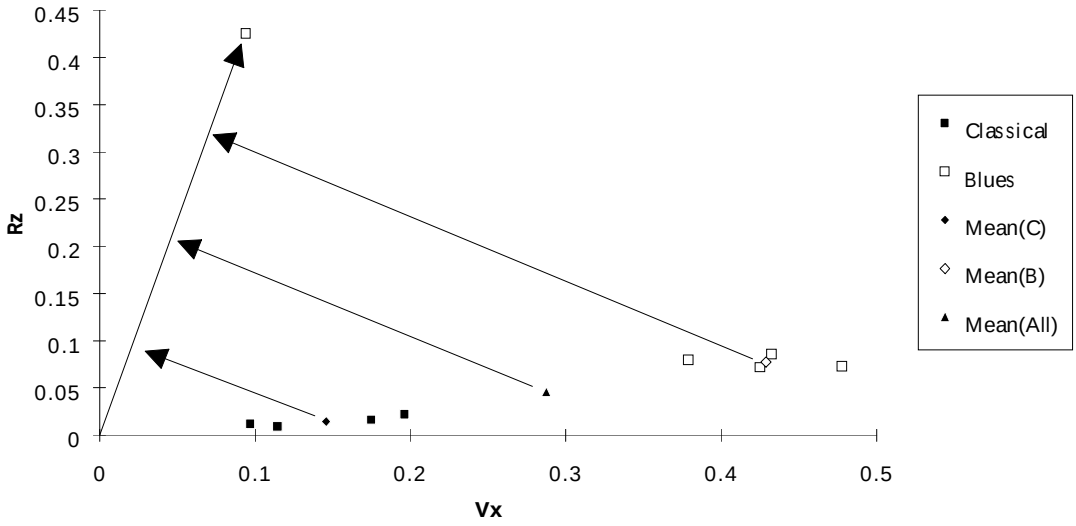


Figure 6. Two clusters of data from Table 1, and the projection line which corresponds to the direction $\underline{w}^T$ = (0.062785, 0.283614), where Vx means translation speed horizontally, while Rz means rotation about the vertical axis. Mean(C) represents the mean of set H_1(Classical) and Mean(B) the mean of set H_2. The projection of the mean of all data is a threshold.

Appendix

For the classical, the notes of melody are restricted to the notes of the natural major scale related to its background chord, which are in different octaves of C, D, E, F, G, A, B, for example, in C major, and every note is divided into quarter note evenly. For the blues, Blues Dorian scale restricts the notes of melody, that is, in C major, A, C, D, E^b,E, G in different octaves are considered, and every two adjacent notes are separated in the way of syncopation like *swing*.

Acknowledgment

This paper is partially supported by National Science Council under grant NSC83-0425-E002-140 and NSC83-0408-E002-006. We would like to thank graduate students Yuong-wei Lei and Jiann-Rong Wu in the system implementation.

References

1. Rung-Huei Liang, Ming Ouhyoung, *Impromptu Conductor: A Virtual Reality System of Computer Graphics and Music*, Proceedings of National Computer Symposium, Taiwan, **Vol. 1**, (1993), pp. 474-480.
2. Stanley Sadie, Alison Latham, *The Cambridge Music Guide*, Cambridge University Press.
3. Claude V. Palisca, *Norton Anthology of Western Music*, 2nd edition, W. W. Norton.
4. Robert Schalkoff, *Pattern Recognition Statistical, Structural, and Neural Approaches*, Wiley.

5. Voss, R. F., and J. Clarke. *1/f Noise in Music: Music from 1/f Noise.* Journal of Acoustical Society of America, (1978), 63(1), pp. 258-263.
6. Charles Dodge, *Profile: A Musical Fractal,* Computer Music Journal, **Vol. 12**, No. 3, (Fall 1988), pp. 10-14.
7. Xavier Chabot, *Gesture Interfaces and a Software Toolkit for Performance with Electronics*, Computer Music Journal, **Vol. 14**, No. 2, (Summer 1990), pp. 15-27.
8. Jack Hsieh, *MIDI and Computer Music*, The Third Wave, Taiwan, (1990).
9. Jim Conger, *MIDI Sequencing in C*, M & T Books.
10. Jun-ichi Nakamura, Tetsuya Kaku, Tsukasa Noma, and Sho Yoshida, *Automatic Background Music Generation Based on Actors' Emotion and Motions*, Proceedings of Pacific Graphics '93, **Vol. 1**, (1993), pp. 147-161.
11. Frederick P. Brooks Jr., Ming Ouhyoung, James J. Batter, P. Jerome Kilpatric, *Project Grope-- Haptic Displays for Scientific Visualization*, ACM Computer Graphics, **Vol 24**, No. 4, pp. 177-185(SIGGRAPH 1990).

Plate 1. Impromptu Conductor system in usage. Three trackers are used: two on the ends of drumsticks to report the hand motion and one on top of the headphone to detect the head motion. A user wears a stereo glasses to get stereo images and a headphone to listen to the generated stereo music.

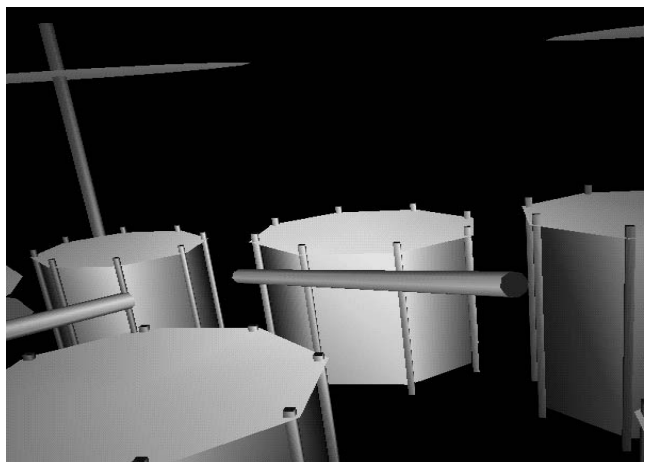


Plate 2. Image generated on the screen. Three trackers are used to control two drumsticks and the drum set viewing angle respectively. 552 triangles were used in the drum model.