The 33rd Annual Conference of the IEEE Industrial Electronics Society (IECON)
Nov. 5-8, 2007, Taipei, Taiwan

Vision-Based Hand Gesture Recognition for Understanding Musical Time Pattern and Tempo

Hongmo Je
Dept. of CSE, POSTECH
Hyoja-Dong, Pohang,
South Korea, 790-784
Email: invu71@postech.ac.kr

Jiman Kim
Dept. of CSE, POSTECH
Hyoja-Dong, Pohang,
South Korea, 790-784
Email: jmk@postech.ac.kr

Daijin Kim
Dept. of CSE, POSTECH
Hyoja-Dong, Pohang,
South Korea, 790-784
Email: dkim@postech.ac.kr

***Abstract*— We introduce a method of understanding of four musical time patterns and three tempos that are generated by a human conductor of robot orchestra or an operator of computer-based music play system using the hand gesture recognition. We use only a stereo vision camera with no extra special devices. We suggest a simple and reliable vision-based hand gesture recognition with two naive features. One is the motion-direction code which is a quantized code for motion directions. The other is the conducting feature point (CFP) where the point of sudden motion changes. The proposed hand gesture recognition system operates as follows: First, it extracts the human hand region by segmenting the depth information generated by stereo matching of image sequences. Next, it follows the motion of the center of the gravity(COG) of the extracted hand region and generates the gesture features such as CFP and the direction-code. Finally, we obtain the current timing pattern of beat and tempo of the playing music by the proposed hand gesture recognition using either CFP tracking or motion histogram matching. The experimental results on the test data set show that the musical time pattern and tempo recognition rate is over 86.42% for the motion histogram matching, and 79.75% for the CFP tracking.**

I. Introduction

Gesture recognition, as a part of pattern recognition and analysis, is human interaction with a machine (definitely including computer) in which human gestures, are recognized by the machine. Recognizing gestures as input might make machines more accessible for the physically-impaired and make interaction more natural for young children. It could also provide a more expressive and nuanced communication with a machine. Numerous works on gesture recognition have already been conducted [1], [2].

Hand gesture recognition, as one of gesture recognition problem, is so important that the motion of human hands can provide abundant information of human intention and implicit meaning to the machines in real world. Many reports on intelligent human machine interaction using hand gesture recognition have already been presented [3], which can be mainly divided into "Data Glove- based" and "Vision-based" approaches.

The "Data Glove-based" methods use a special input device named "hand data sensor glove" for digitizing hand and finger motions into multi-parametric data. It is possible to analyze 3D space hand motion with the sensing data. However, the device is too expensive and the users might feel uncomfortable when they communicate with a machine [4].

Without specialized tracking devices, one of the greatest challenges of the system is to reliably detect and track the position of the hands using computer vision techniques. The "Vision-based" methods use only the vision sensor; camera [5]. In general, the entire system of the vision-based hand gesture recognition must be more simple than the Data Glove-based approach, and it makes human-friendly interaction with no extra device. The vision-based hand gesture recognition is a challenging problem in the field of computer vision and pattern analysis, since it has some difficulties of algorithmic problems such as camera calibration, image segmentation, feature extraction, and so on.

Conducting a music band is a highly sophisticated art that has been matured over centuries [6]. Recently, some researchers in the filed of human computer interaction also have been concerned about creating a machine-based music play system, which includes intelligent robots or computers, and considers the conductor's desired beat and the tempo. The first electronic orchestra with a complex performance database and Musical Instrument Digital Interface (MIDI) controllers responds to the gestures of the conductor through a sensor glove. Also, a special purpose electronic baton was introduced in [7]. It can identify 4 over 4 beat timing pattern (4/4) by following the motion of electronic baton in the right hand, while recognizing a play speed by tracking the 3D motion of the left hand wearing a sensor glove. Another methodo of using sensor glove had been proposed by Winker [8]. In Modler [9], neural networks for mapping hand gestures into the music play parameter in the "virtual musical system" is described. In addition, a research of mapping gesture into music using 3D motion data captured by a commercial 3D motion capture system (Vicon 8) has been reported [10].

Bien and Kim [11] have suggested a vision-based method for understanding human's conducting action for chorus with a special purpose light baton and infra-red camera. They proposed a vision system which captures the image sequences, tracks each end-point of the baton which is a stick having a distinguished color feature to be detected easily by a camera, and analyzes a conducting action by fuzzy-logic based inference. Lately, Watanabe and Yachida [12] have proposed a real-time interactive virtual conducting system using the Principle

1-4244-0783-4/07/$20.00 ©2007 IEEE

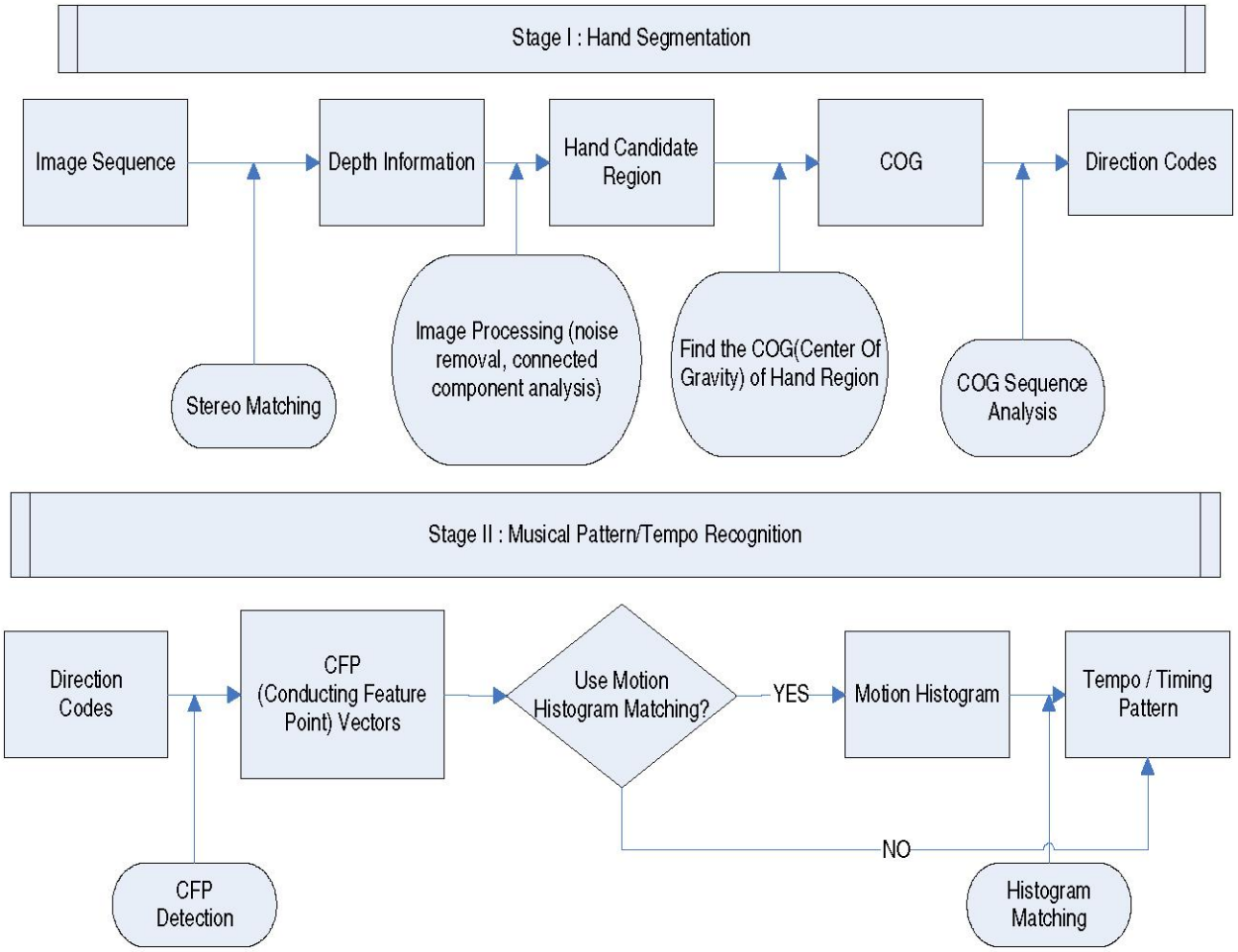


Fig. 1. Block diagram of the proposed system.

Component Analysis (PCA)-based gesture recognition that can identify only 3/4 time pattern.

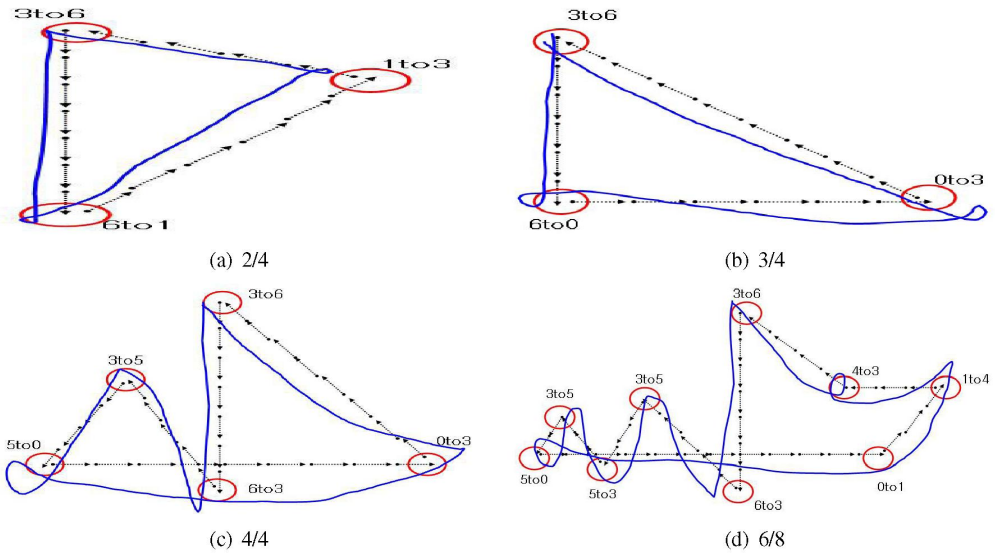
In general, conductors perform various music using both hands and natural conducting actions may be very difficult to represent. Hence, we take a few assumptions to alleviate the problem as follows: 1) the conductor uses only one-side hand ,2) the conducting action must be in the view range of the camera, 3) the conductor may indicate four timing patterns (2/4, 3/4, 4/4, 6/8) with three tempos (Andante, Moderato, Allegro) by his/her hand motion, 4) the conductor needs no special devices.

We propose a very simple but reliable vision-based hand gesture recognition of the music conductor with no extra devices. Unlike the previous vision-based hand gesture recognition, we use the depth information, instead of using the intensity or the color information of image, generated by a stereo vision camera to extract human hand region that is the key region of interest (ROI) in this application. Our proposed system can obtain the motion velocity and the direction by tracking the center of gravity (COG) of the hand region, which provides the speed of any conducting time pattern. We introduce two methods to recognize the musical time pattern. One is the *CFP tracking* which uses only special features like conducing feature point and another is the *motion histogram matching* which can identify the time pattern and the tempo at once, where the "Mahalanobis distance" is chosen as the distance metric of motion histogram matching.

The remainder of this paper is organized as follows. Section II describes the proposed hand gesture recognition system to understand music time pattern and tempo in detail. Section III presents the experimental results of both simulation and real world videos. Finally, section IV draws conclusion and discusses future work.

II. The Proposed Hand Gesture Recognition System for Understanding Music Time Pattern and Tempo

Fig. 1 illustrates the block diagram of our proposed system. The system has two stages which are *'hand segmentation'* and *'music time pattern and tempo recognition'*. To make understand the proposed system easy, data (or information) are represented as rectangles, action (or processing) units are represented as rounded-rectangles, and arrow lines means the



(a) 2/4 (b) 3/4 (c) 4/4 (d) 6/8

Fig. 2. The real trajectories and the approximated directions of each conducting pattern. (solid line - real trajectory, dashed line - motion direction, red circle - conducting feature point).

flow of data or control information. More details on each stage are described in following subsections.

A. Hand Segmentation Stage

Hand segmentation is the first step of our proposed hand gesture recognition system, which separates the human hand region from the others. Most methods for the hand segmentation uses the skin color information to extract the hand region. The skin color-based hand region extraction is quiet simple, but it is sensitive to light condition change and complicated and cluttered background which has many skin-like colored objects such as wood and wall papers. We use the depth information of a stereo image instead of the 2D pixel image. The depth information might not only be insensitive in any light condition but also robust even if there is a complicated background.

Normally, the members of orchestra must concentrate their attention on the conductor's face and hands to read his/her intention and his/her hands are placed in front of conductor's body when conducting. Based on this fact, we utilize a face detector to detect the human face, which allow us to find the hand candidate region easily because we know that the depth of hand region must be closer than the face region to the stereo camera. Fig. 3 shows the intermediate results through hand segmentation stage.

The exact hand region can be completely segmented after postprocessing on the extracted hand candidate region. Fig. 3 shows several postprocessing stages such as the morphological operator and the connected component analysis to remove the non-hand region [13].

To track the motion of the hand, the COG of the hand region needs to be computed. We approximate the COG by computing the mean coordinates of the segmented hand region as

$$X_{cog} = \frac{\sum_i x_i}{N}, Y_{cog} = \frac{\sum_i y_i}{N} \quad (1)$$

where x_i, y_i are the x and y coordinates at the ith pixel position, respectively, and N is the number of pixels of the hand region.

B. Musical Time Pattern and Tempo Recognition Stage

In contrast to the hand sign language recognition, the hand gesture recognition for understanding a musical time pattern and tempo does not have to be accurate. While a slight posture or movement of hands in the hand sign language represents an independent and important meaning, only salient features like beat transition point of hand motion are the most important information in the conducting gesture.

1) The direction code of the hand motion: The easiest way to find the trajectory of the conducting gesture is to track the motion direction of the hand. We obtain the direction angle of the hand motion by computing the difference between the previous COG of hand region and the current COG of it as

$$\begin{aligned} \Delta X_{cog}(t) &= X_{cog}(t) - X_{cog}(t-1), \\ \Delta Y_{cog}(t) &= Y_{cog}(t) - Y_{cog}(t-1), \\ \theta(t) &= \arctan \frac{\Delta Y_{cog}(t)}{\Delta X_{cog}(t)}, \end{aligned} \quad (2)$$

where $\theta(t)$ is the direction of the hand movement on time t. To represent the direction-code, the real value of the hand direction should be quantized in eight directions. Fig. 4 shows three-bit codes for the eight dominant direction of hand movement.

2) Conducting feature point: Instead of analyzing all the sequences of motion directions, we implement simple finite state machine (FSM) to track the salient features of conducting

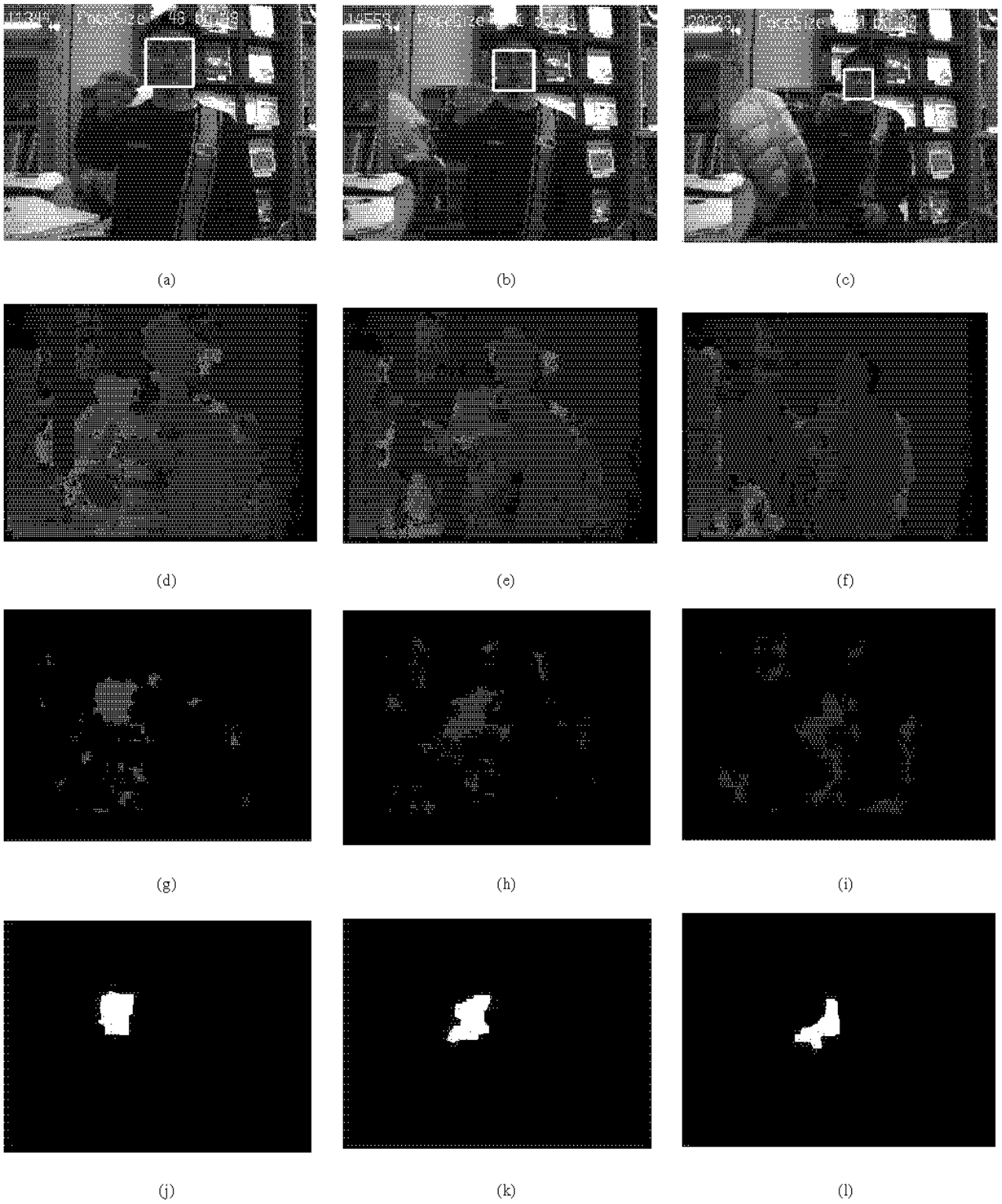


Fig. 3. (a),(d), (g) ,(j) A human conductor stands at 1.0m from the camera, (b),(e), (h), (k) 1.5m from the camera, (c),(f), (i), (l) 2.0m from the camera. ; the first row (a), (b), (c) show images ; the second row (d), (e), (f) show the depth maps ; the third row (g), (h), (i) show the noise removal and segmentation; the fourth row (j), (k) (l) show the result after connected component analysis and morphological operation

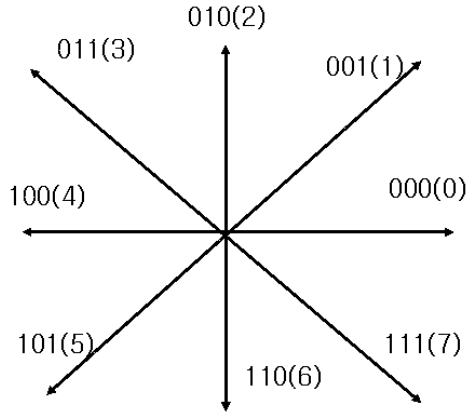


Fig. 4. Eight direction codes.

TABLE I
THE SEQUENCES OF CFPS FOR TIME PATTERNS

Time Pattern	CFP Sequences
2/4	*3to6–6to1–1to3*
3/4	*3to6–6to0–0to3*
4/4	*3to6–6to3–3to5–5to0–0to3*
6/8	*3to6–6to3–3to5–5to3–3to5–5to0–0to1–1to4–4to3*

gestures and to recognize the musical time patterns. Fig. 2 illustrates the representative features which are called as "Conducting Feature Point (CFP)", of each musical time pattern. For example, a musical time pattern of 2/4 has three CFPs which are *3to6*, *6to1*, and *1to3*. Assuming that the new coming CFP is *6to1* while the previous CFP is *3to6* or the start point of the initial gesture, then the gesture recognition system expects the next CFGs *1to3* following *3to6*. Thus, the recognition system can identify the time pattern by observing the sequence of CFPs. Table I explains the CFP sequences for musical time patterns.

3) Motion histogram matching: Although the analysis of CFP sequences is reliable for identifying musical time patterns, it can fail when the system misses an important CFP. This can be occur for a complicated time pattern like 6/8 which has a large variation among the different human conductors. To avoid this problem, we propose a motion histogram matching based on the musical time pattern tempo analysis.

We can obtain a cycle of each time pattern, where one cycle means a sequence of the direction-code from the start point of time patterns to thier end point (usually both the start point and the end point are the same). In general, most musical time patterns have *"3to6"* type of the CFP as the start point of thier action.

In the training stage, we collect the histogram vectors $H = [h_0, h_1, ..., h_7]$ where h_0 to h_7 are the number of each direction code for the cycle and obtain the statistics (mean, variance) of motion histogram for all combinations of time patterns (2/4, 3/4, 4/4, 6/8) and tempos (Andante, Moderato, Allegro). Then, the mean and variance vectors H_μ and H_Σ of the histogram vectors can be computed as

$$H_\mu = \frac{1}{N}\sum_i H_i, \qquad (3)$$
$$H_\Sigma = \frac{1}{(N-1)}\sum_i (H_i - H_\mu)^2,$$

where N is the number of training data.

Thus, we have twelve profiles of motion histogram. Table II denotes the profile index for the time patterns and tempos. For example, H_μ^1 represents the mean of 2/4 with moderato tempo and H_Σ^{11} represents the variance of 6/8 with allegro tempo. We selected the "Mahalanobis distance" as a metric of motion histogram matching. By Eq.(4), the similarity scores for all profile are evaluated. The proposed musical time pattern and tempo recognition system identify the time pattern and tempo by taking the profile whose similarity score is the minimum.

$$MD^k = \sqrt{(H_c - H_\mu^k)^T {H_\Sigma^k}^{-1} (H_c - H_\mu^k)} \qquad (4)$$

$$ProfileIndex = \arg\min_k MD^k, \qquad (5)$$

where H_c is the current motion histogram and k is the profile index given in Table. II.

TABLE II
THE PROFILE INDEX FOR THE TIME PATTERNS AND TEMPOS.

Index	**0**	**1**	**2**	**3**	**4**	**5**
Time pattern	2/4	2/4	2/4	3/4	3/4	3/4
Tempo	And.	Mod.	Alle.	And.	Mod.	Alle.
Index	**6**	**7**	**8**	**9**	**10**	**11**
Time pattern	4/4	4/4	4/4	6/8	6/8	6/8
Tempo	And.	Mod.	Alle.	And.	Mod.	Alle.

III. EXPERIMENTAL RESULTS

We used the *'BumbleBee stereo vision camera'* [14] as input sensor. Since it automatically provides depth information for each frame of stereo images, we needed not perform stereo matching with heavy computational complexity. We collected they conducting gesture data for each time pattern and tempo which consists of 300 cycles respectively. We divided them into 200 cycles for training the recognition system and 100 cycles for testing the recognition system. Fig. 5 represents the recognition rate of the experiments for the profile indices. As a result, the average recognition rate of using the CFP sequence analysis is 79.75% and that of using the motion histogram matching is 86.42%.

IV. CONCLUSION AND FUTURE WORK

This paper presented a vision-based hand gesture recognition. As a simple application, we implemented a system for understanding musical time pattern and tempo that was generated by a human conductor. We only used the stereo vision camera with no extra devices such as motion capture system, data gloves, and etc. Instead of using the color pixel image, we used the depth information of the current stereo

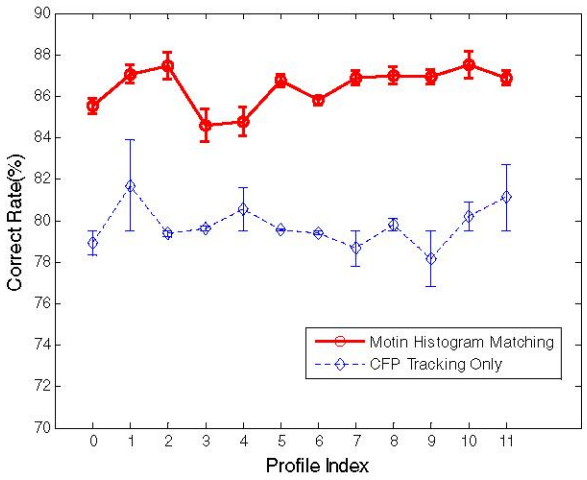


Fig. 5. The plot of recognition rate ; red solid line for *'Motion Histogram Matching'*, blue dashed line for *'CFP tracking only'*

image. It is easy to extract an interesting region on the depth map generated by stereo matching. In addition face detection helped us to find the hand candidate regions because we assumed that the hand motion of the conductor usually is made in front of the human face. With this assumption, we could extract the hand region faster and easier.

We introduced the conducting feature points with the direction-codes which simply indicated the musical time pattern and tempo that was generated by hand motions. When the human conductor made a conducting gesture, our proposed system tracked the COG of the hand region and encoded the motion information into the direction-code. We also suggested the motion histogram matching which could identify the current musical time pattern and tempo simultaneously by finding the best matched distribution of direction-code in a cycle. From the numerous experiments, the recognition accuracies were 79.75% and 86.42% using the CFP sequence analysis and the motion histogram matching, respectively.

The proposed conducting gesture recognition method currently has two limitations. First, it cannot recognize the decoration conducting actions such as *cresendo*(gradually stronger), *decresendo*(gradually weak), *staccato*(seperate sound), and various dynamics signs(*mf, f, mp, p*). We will extend the gesture recognition algorithm by using a powerful gesture recognition algorithms such as HMM (Hidden Markov Model), and DBN (Dynamic Bayesian Networks which are the promising state space models that can deal with the complicated motions. Second, it can only treat the the frontal view of conductor, which means the conductor always needs to confront the vision sensor while some members of robot orchestra might sit on not the exact front side but another angle side. We can deal with an unrestrict hand motion in the 3D space considering the structural connectivity of human body. If we have the hand motion information at any positions, then the conductor may not have to be face with camera.

It is also a challenging problem to implement the robot orchestra, which has an ability to perform a virtual instrument, and play music by following the conducting actions of the human conductor.

ACKNOWLEDGMENT

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs partially funded by the Ministry of Science and Technology of Korea. Also it was partially supported by the Ministry of Education and Human Resources Development(MOE), the Ministry of Commerce, Industry and Energy(MOCIE) and the Ministry of Labor(MOLAB) through the fostering project of the Lab of Excellency.

REFERENCES

[1] R. Watson, "A survey of gesture recognitiontechniques.," Tech. Rep. TCD-CS-93-11, 1993.
[2] Joseph J. LaViola Jr., "A survey of hand posture and gesture recognition techniques and technology," Tech. Rep. CS-99-11, 1999.
[3] Ying Wu and Thomas S Huang, "Vision-based gesture recognition: A review," *LNCS:Gesture-Based Communication in Human-Computer Interaction: International Gesture Workshop*, vol. 1739, pp. 103, 2004.
[4] A. Mulder, "Hand gestures for hci," *Technical Report 96-1*, vol. Simon Fraster University, 1996.
[5] F. Quek, "Toward a vision-based hand gesture interface," in *Proceedings of Virtual Reality Software and Technology*, Singapore, 1994, pp. 17–31.
[6] Ingo Grull, "Conga: A conducting gesture analysis framework," *ULM University*, 2005.
[7] Hideyuki Morita, Shuji Hashimoto, and Sadamu Ohteru, "A computer music system that follows a human conductor," *IEEE Computer Society*, vol. 24, pp. 44–53, 1991.
[8] Todd Winkler, "Making motion musical : Gesture mapping strategies for interactive computer music," in *Proceedings of the 1995 International computer Music Conference*, Canada, 1995, pp. 27–31.
[9] Paul Modeler, *Trends in Gestural Control of Music*, pp. 301–313, IrCam, 2000.
[10] Christopher Dobrian and Frederic Bevilacqua, "Gestural control of music using the vicon 8 motion capture system," 2002.
[11] Zeungnam Bien and Jong-Sung Kim, "On-line analysis of music conductor's two-dimensional motion," San Diego, CA, USA, 1992, pp. 1047–1053.
[12] Takahiro Watanabe and Masahiko Yachida, "Real-time gesture recognition using eigenspace from multi-input image sequences," *IEEE Computer and System in Japan*, vol. 30, no. 13, pp. 810–821, 1999.
[13] R. Bloem, H. N. Gabow, and F. Somenzi, "An algorithm for strongly connected component analysis in $n \log n$ symbolic steps," pp. 37–54, Nov 2000, LNCS 1954.
[14] "http://www.ptgrey.com/products/stereo.asp," .