Application of MML to Motor Skills Acquisition

C. Sun [*] F. Naghdy [†]
D. Stirling [‡]

[*]University of Wollongong,
[†]University of Wollongong, fazel@uow.edu.au
[‡]University of Wollongong, stirling@uow.edu.au

This article was originally published as: Sun, C, Naghdy, F & Stirling, D, Application of MML to Motor Skills Acquisition, International Conference on Computational Intelligence for Modelling, Control and Automation 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, Sydney, November 2006, 77. Copyright IEEE 2006.

Application of MML to Motor Skills Acquisition

Chao Sun Fazel Naghdy David Stirling
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, NSW, Australia
cs055@uow.edu.au, fazel@uow.edu.au, stirling@elec.uow.edu.au

Abstract

Study on modeling human psychomotor behaviour based on tracked motion data is reported. The motion data is acquired through various integrated inertial sensors, and represented as Euler angles and accelerations. The Minimum Message Length (MML) algorithm is used to identify frames of intrinsic segmentations and to acquire a classification basis for unsupervised machine learning. The classification model can ultimately be deployed in recognizing certain skilled behaviors. The prior results are analyzed as FSMs' (Finite State Machines) to extract the potential rules underlying behaviors. The progress made so far and plan for further work is reported.

1. Introduction

Acquisition of the human psychomotor behaviour has become a popular research area. In addition to self-discovery, learning of skills in humans generally takes place through training by an instructor in the psychomotor domain, where 'motor' is an observable movement response to a stimulus. According to Smith and Smith [21], there are three types of movements, postural, locomotor and manipulative movements. In this work we focus on locomotor movements, which translate and rotate a body. The aim is to capture predefined behaviours in the inertial sensors data, and subsequently analyze the multidimensional patterns, employing an unsupervised Minimum Message Length (*MML*) encoding — a machine learning method, in order to build a model then to distinguish different behaviours with this model.

In this paper, Minimum Message Length (*MML*) encoding is deployed to build a primitive model for three predefined human arm behaviours. This is subsequently validated on new and unseen data were the model is employed to identify new examples of each behaviour. In Section 2 a review of the previous work will be carried out. The experimental details and considerations will be described in Section 3. In Section 4 some aspects of the related theory will be discussed and in Section 5 a range of experimental results will be presented. These results are further considered and discussed in Section 6, followed by an outline of future work in Section 7.

2. Background

The study for human motion modeling has become of particular interest in the robotics and other relative fields. Generally, in order to analyse human behaviour, an approach is to define and segment these into motion primitives [13], and describe or generate new behaviours using such primitives. These can be defined in various patterns, according to different methods and theories.

In the work conducted by Nakazawa [15, 16], the primitives are defined as composition of "motion base + motion style". Both the motion-base and -style are graphic based and calculated from monochromatic video presenting human dance motions. An alternative approach is to use the number of Degree of Freedoms (DoFs) to define the motion primitives. Amit and Matari [1] assumed a set of innate base primitives to control the DOFs in their motion learning framework. The choice of primitives is required for animating the robot as well as being able to characterize the skilled patterns of motion observed. HaiBing and colleagues [10] modeled motion primitives by utilizing Gaussian Mixture Models and their distribution densities in their study of human actions. In this work the primitives are not constrained to relate to any particular mode, they can be associated with any type of motion segment and can be used to describe any manoeuvre.

In order to acquire motion data of human behaviour, various types of sensors have been employed. Optical and inertial sensors are popular choices in this area, but they require significant postprocessing of image data in order to deduce motions of fixed points. Optical sensors are mostly unobtrusive, practical and will not impact on the motion of the subject. Such systems are widely used for industrial, and or, public monitoring purposes [1, 8, 26, 27]. Although

International Conference on Computational Intelligence for Modelling
Control and Automation,and International Conference on
Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

in certain cases the 2-D information provided by monocular camera vision has proved to be sufficient for monitoring purposes, 2-D vision systems do not function as ideal sensors for complex human behaviour study. To make motion feature extraction simpler, reflective markers are often mounted on the subject [12]. Multiple cameras are used to extend 2-D image perception for 3-D space in order to cope with more demanding or complex motions and scenarios [19].

There are several different types of inertial sensors now being utilized for motion capture, such as *Micro-Electro-Mechanical Systems (MEMS), solid state accelerometers, gyroscopes, magnetometers*. These, being relatively unobtrusive, can be mounted externally on the subject at the precise points of interest, providing direct and accurate measurements of motion and posture, often in real-time. The electrical signals generated by a single inertial sensor, are mostly a direct analogue of some specific aspect of the motion observed. However, consistent singular types of inertial data are often insufficient for motion study and it is necessary to augment their type. Sensor fusion [7, 14, 20] is one procedure to combine different types of sensory data and integrate them to form useful motion features.

Different machine learning and data mining methods have been applied in this area, in order to segment the observed human motions into various primitive modes from sensory data. As such primitives could be combined in various permutations to form plausible to useful segments, almost all the popular classification methods have been used by different authors.

A Support Vector Machine (SVM) algorithm was employed by Sukthankar and Sycara [22] in their military manoeuvre recognition project. Kumar et al. [11] were able to successfully classify and recognized human hand gestures using an Artificial Neural Network (ANN) and a Motion History Image (MHI) [2] in order to characterize the motion from a high dimensional space into a low dimensional space, and established a recognition criterion through a nearest neighbour technique. Fuzzy logic is also been endorsed by a number of authors, Nascimento et al. [17] developed the *Fuzzy Clustering Multiple Prototype* (FCMP) approach, based on FCM seeking to provide improved performance in fitting various proposed models. Haibing et al. [10] developed an alternative *Primitive-based Coupled Hidden Markov Model* (PCHMM) method based on a traditional Hidden Markov Model (HMM), endeavouring to recognize complex, but natural human actions within a smart classroom [9].

3. Experimental Setup

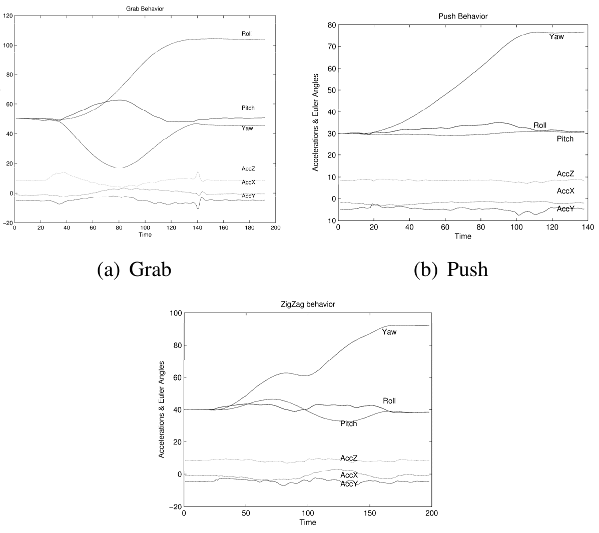
The sensor used in this work to capture the dynamic motion behaviours is an integrated inertial unit the MTx. The MTx itself combines nine individual MEMS sensors to provide drift-free 3D orientation as well as kinematics data: 3D acceleration, 3D rate of turn (rate gyro) and 3D magnetometers [5]. Embedded DSP within the MTx unit provides Euler angles, kinematics and the orientation matrices as outputs. The focus, however, is primarily on the Euler angles and accelerations as motion training data. Euler angles present the posture of the body at the point the sensor is attached, and the accelerations are relative to current movement. Both Euler angles and accelerations have 3 orientations, Roll, Pitch and Yaw for the Euler angles and X, Y, Z axes for accelerations. The input for our system is a mixed stream of these 6 features.

Three specific skilled tasks, principally locomotor and manipulative movements involving the arm and hand are chosen as target training behaviours, which would be to be performed several times, and hopefully later recognized by the resultant model. These are called Grab, Push and Zigzag. The goal for each of these behaviours is the same: to move an object (small tube of glue) from one location to another on a flat surface (desktop). The difference between each task is the trajectory used to complete it. Grab requires the subject, using their hand, to pick up the glue tube, translate and deposit it at the end point; Push, alternatively requires the subject again using their hand to push and to slide the glue along the desktop in a straight path until the end point is achieved; and Zigzag requires the subject to alternatively push/slide the object along the desktop, following an S-shaped path between start- and end- points.

This work is based on the idea that the same style of behaviours corresponds to similar trajectories within the 6-dimensional feature space. Example trends of the changes occurring for each these six features for all three behaviours are illustrated in Figure 1. One can readily appreciate from the trends in Figure 1 notable variations between behaviours. For example the variation in Pitch is pronounced in the "Grab", manoeuvre whilst relatively flat in the case of Push. In comparison, all features in the Zigzag task manifest a greater degree of variation, even to the extent of directly portraying the S-shape path being followed.

In most cases, the combinations of Euler angles and accelerations vary dynamically for each of the different behaviours. It is conjectured that if the trajectories for all behaviours are examined in the six-dimensional space, that these trajectories should closely correspond with the same start and end points, but possibly pass through differing areas in-between.

By repeating these behaviours many times, and on each occasion noting/labelling the skill performed, it is possible to segment all of the stream of data into a library of unique modes or clusters within the 6-D feature space. By reinstating the task-labels (or indices), post the segmentation process, it will be possible to identify which cluster/mode



(a) Grab (b) Push

(c) Zigzag

Figure 1. Example trends of all features in performing each behavioural task

is associated to which behaviour, as well as, which of these are common or unique to each manoeuvre, and also what is the generic sequence for each. In this way, we transform the 6-feature sequence data into a modecluster sequence. If the experimental results indicate promising similar output sequences for the same style of behaviours, and notable dissimilarities between different behaviours, then such sequences could be utilised as the basis for a skill model with which to recognize and classify future (or unseen) motion data.

4. Theory

There are many ways in which to partition the feature space into qualitative or purposed based regions. The simplest way could be by dividing it into a large number of cubes. Indexing each of these cubes would be a relatively easy task, but the storage manipulation and access of these may not scale satisfactory: 1) Human behaviours are restricted to the body frame; so many cubes in this hypothesised space may never be reached by those trajectories that only correspond to realistic muscular-skeletal behaviours; 2) the number of cubes will increase exponentially. Too many cubes may be required in order to achieve a reasonably sensitive system.

Fuzzy C-means (FCM) is an unsupervised classification method based on fuzzy logic, it was initially proposed by Dunn [6] and later generalized by Bezdek [3, 4]. The algorithm clusters sample data automatically according to the Euclidian distances between the data instances. Generally only the targeted number of clusters needs to be defined by the user and, the algorithm can be applied to data distributed over multiple dimensions. Seemingly, the FCM approach was easy to comprehend and implement for our purposes. However, additional (subjective) adjustments were ultimately required before satisfactory and or usable results were obtained. The data acquired from the MTx sensor units, is necessarily a stream of multivariate data types, and for our purposes these reduce to two different types of data, Euler angles and accelerations. Since FCM is based on deterministic, distance based metrics, it treats the values of Euler angles the same as accelerations. Without having to normalize these components through an additional preprocessing, this kind of disparity is suggestively resolved by adjusting a set weight for each feature. Further more, its not easy to determine a‘priori how many clusters is best when using FCM. This approach provides appropriate optimized partitions for a fixed number of clusters, but little guidance is available in determining the best number of clusters. In overcoming this, an alternative Gaussian mixture-modeling, probability based segmentation approach was considered, here Minimum Message Length (MML) encoding.

The Minimum Message Length (MML) principle [18,24, 25] of machine learning is based on information theory and statistics. The rationale behind MML is to postulate a model of the data as a series of candidate partitions, then to evaluate this by estimating the amount of code required to describe the model plus the data exceptions that fall outside of it. In considering alternative, and or successive models, those that reduce the total message length are maintained and further specialized. In short, the reduction of message length becomes the guiding metric of the data segmentation or clustering model. The MML mixture modeling programs employed addresses both the model selection and parameter estimation.

The MML principle tries to encode the data with various theories, and then evaluate the theory that maximizes the product of the prior probability of theory with the probability of the data in light of that theory.

5. Experiment and Results

Experiment is conducted where a subject repeatedly performed three predefined arm behaviours, Grab, Push and Zigzag. These behaviours are recorded using the MTx sensor in 3D accelerations and Euler angles. Each behaviour is repeated 30 times. The behaviours with similar trajectories are repeated during this section to achieve conformity.

The steps for training and classifying behaviours are as follows:

1. *Extract Euler angles and accelerations as motion features, and treat them as a 6-D data sequence.*

2. *Append all of the 3x30 behaviour data one by one as the training data set.*
3. *Apply the MML algorithm on the training data set, and record the message lengths for all the resulting mixture models.*
4. *Analyst the message length history to determine an optimal model size and number of clusters.*
5. *Re-visit the training data with MML to obtain a complete model with the selected number of clusters.*
6. *Utilize the model to recognize different behaviours and analyse differences among outputs.*
7. *Summarize the rules for recognizing behaviours from the outputs.*

By applying the MML algorithm to the training data set, the segmentation process continues to function, searching to find the best cluster set that makes the message length the shortest. Normally the cluster number can grow up to 150 clusters before the procedure stops itself, that is were there is no improvement in the message length over a significant number of preceding attempts.

Considering the marginal utility for reducing message length while increasing cluster number, a reasonable balance point between the cluster number and message length should be found. Figure 2 illustrates how the message length typically reduces as the number of clusters increases. The darker line of Figure 2 illustrates a typical decreasing message length corresponding to various models of increasing number of clusters. Here, the message length decreased from an initial maximum 262,669 (1 cluster) to the minimum 169,978 (83 clusters), representing a compression of 35.29%. The four red lines show the message length when decreasing extents are 80%, 85%, 90% and 95% of the total decrease. From the trend in message length in Figure 2, we notice that its curvature changes perceptibly after passing the 90% threshold line, which in turn intersects with the curve at approximately 27 clusters. This then is considered as a reasonable, minimal yet sufficiently responsive, model with which to model the skilled behaviours.

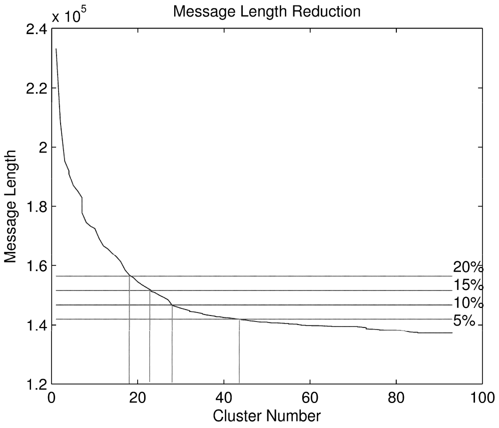


Figure 2. Typical mixture model training illustrating the reduction of the total message length

A new mixture model containing 28 clusters was subsequently trained. Each cluster (or mode) formulated within this new model accounts for a specific proportion, or abundance of the whole data. Subsequently, these clusters are re-sorted with a descending order of abundance, and re-labeled accordingly, as seen in Figure 3. One can appreciate; that the more significant clusters, those with the Figure 2. Typical mixture model training illustrating the reduction of the total message length. smaller labels, will correspond to the more common facets of the aggregated motions, where as the higher number clusters in comparison may be attributed to more refined, or unique actions. By utilizing this model, the original training data was mapped to a series of cluster sequences.

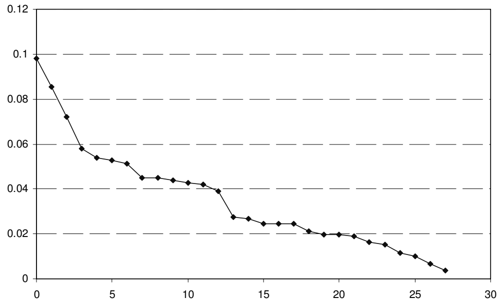
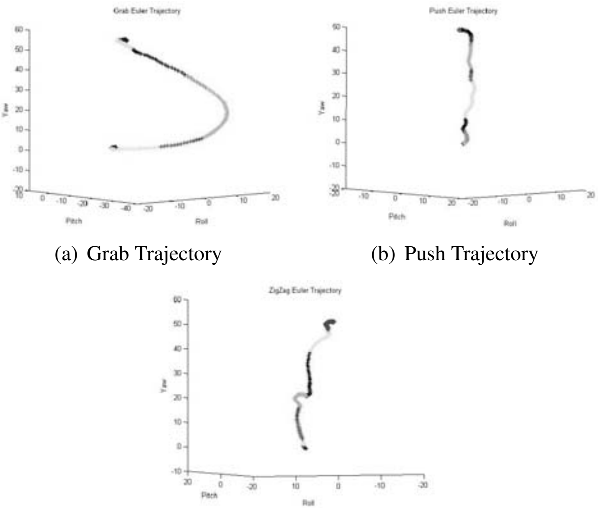


Figure 3. Decreasing abundances of modes in the final mixture model.

By representing differing clusters by different colour, we can describe the behaviours by sequences of different coloured clusters, these clusters can be treated as our motion primitives. For example, as seen in Figure 4, if we plot only the recorded angles for each behaviour within the Euler feature space, and also mark different cluster memberships with different appropriate colours, we can visually determine the temporal and multi dimensional make up or each.

The trajectories in Figure 4 represent how each behaviour is achieved in the Euler angle space, and how the primitives change as the behaviour progresses in time. The size of each cluster is not fixed; it is determined automatically by the MML training process seeking to minimise the message length for the whole model. Every cluster is a set of combinations of both Euler angles and accelerations. A more detailed perspective can be seen in Figure 5 were the behaviours are segmented by the variety and persistence of their respective clusters, within the mode or cluster sequence. In order to test how our model performs in recognizing new occurrences to these behaviours, 10 additional repetitions of each behaviour were performedł facilitating the acquisition of unseen test data. And at the same time, the final MML trained model was utilized in a classification/prediction mode to output cluster sequences.



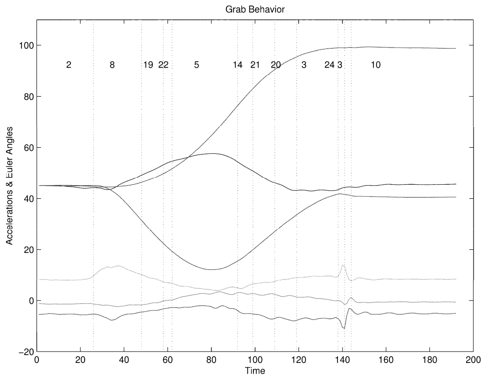
(a) Grab Trajectory (b) Push Trajectory

(c) Zigzag Trajectory

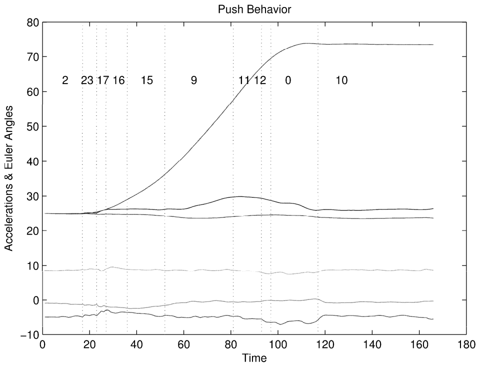
Figure 4. 3-D Euler trajectories for each behavioural task identifying common different motion modes clusters

An alternative format with which to consider these sequences is by using a form of tuples such as:
$\ldots < 5, 34 >< 14, 10 >< 21, 5 >< 20, 9 > \ldots$

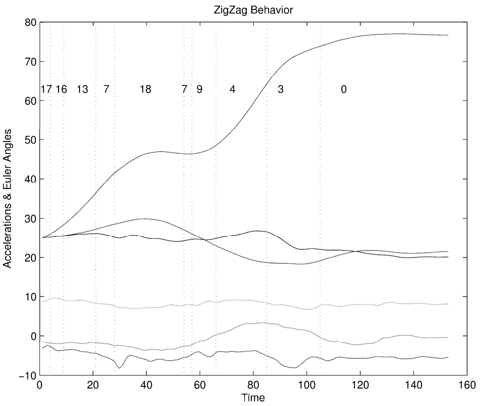
The first number in the brackets becomes the primitive index number, and followed by its mean duration, or the persistence of the primitive. By comparing and analyzing such outputs, various patterns or rules may be abstracted for the identification of similar behaviours in the future.



(a) Grab



(b) Push



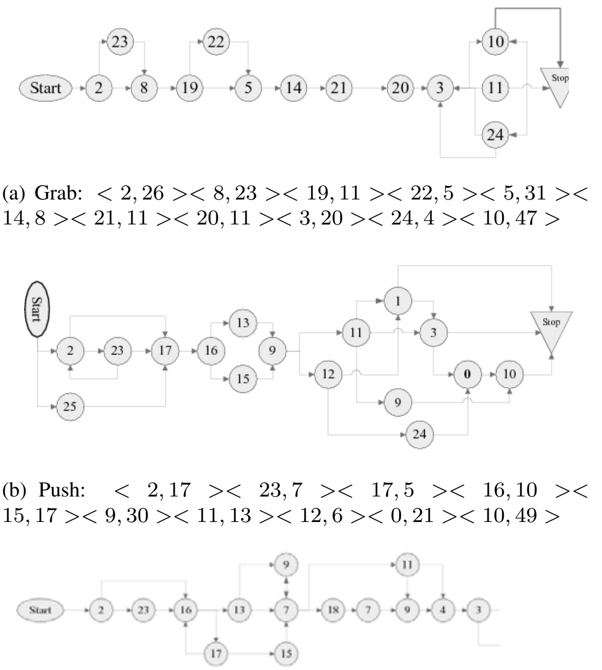
(c) Zigzag

Figure 5. Behavioural task, frame based temporal-segment decompositionsł including kinematic trajectories

6. Results analysis and discussions

A gramma in the form of a Finite State Machine (FSM) is abstracted for recognition of new data. This illustrates a strong coherence among the same behaviour, the results are as in figure 6.

Because only the Euler angles are influenced by orientation, all of these angles start from 0 in the sensor fusion step. Thus the trajectories seen in Figure 5 are actually offsets from the real Euler angle. To test whether the sensor is independent from its orientation, 10 further repeats were performed with the subject facing an orthogonal direction. By analysing the new data, the primitive sequences maintain a notable coherence within same style behaviour as before, whatever the orientation is. Several highly abundant subsequences could also be extracted as criterions for the behaviour styles. FSM (Finite State Machine) and Markov Chain [23] will be utilized in further analysis particularly the conditions for primitive transitions within behaviours.



(a) Grab: $< 2,26 >< 8,23 >< 19,11 >< 22,5 >< 5,31 >< 14,8 >< 21,11 >< 20,11 >< 3,20 >< 24,4 >< 10,47 >$

(b) Push: $< 2,17 >< 23,7 >< 17,5 >< 16,10 >< 15,17 >< 9,30 >< 11,13 >< 12,6 >< 0,21 >< 10,49 >$

(c) Zigzag: $< 17,4 >< 16,6 >< 13,13 >< 7,8 >< 18,27 >< 7,4 >< 9,10 >, < 4,20 >< 3,21 >< 0,48 >$

Figure 6. Brief FSM models for 3 behaviours

7. Future Work

The work mentioned in this paper shows how the human arm behaviour can be modelled and distinguished with a single MTx sensor. In future investigations, up to ten additional sensor units will be utilized, installing these on various parts of a subjects body, in order to capture and study more general contexts of coordinated multifaceted of human behaviours. As the number of sensors increases, the model will also increase in its complexity. Identifying strategies to overcome such issues, and improving the perception of the models, is a motivating challenge for the immediate future.

References

[1] R. Amit and M. Matari. Learning movement sequences from demonstration. pages 203–208, 2002.

[2] K. Babu, R.V.; Ramakrishnan. Compressed domain human motion recognition using motion history information. *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference*, 3:321– 324, 2003.

[3] J. Bezdek. Pattern recognition with fuzzy objective function algorithms. 1981.

[4] J. Bezdek and R. Hathaway. Recent convergence results for the fuzzy c-means clustering algorithms. *Journal of Classification*, 5:237–247, 1988.

[5] X. T. B.V. Mti and mtx user manual and tech. doc.. 2005.

[6] J. Dunn. A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[7] O. Fuentes and R. C. Nelson. Learning dextrous manipulation skills using multisensory information. pages 342–348, 1996.

[8] K. Y. Y. Haga, T.; Sumi. Human detection in outdoor scene using spatio-temporal motion analysis. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference*, 4:331 – 334, 2004.

[9] R. Haibing and X. Guangyou. Human action recognition in smart classroom. pages 399–404, 2002.

[10] R. Haibing and X. Guangyou. Human action recognition with primitive-based coupled-hmm. 2:494–498 vol.2, 2002.

[11] S. Kumar, D. K. Kumar, A. Sharma, and N. McLachlan. Classification of hand movements using motion templates and geometrical based moments. pages 299–304, 2004.

[12] D. Matsui, T. Minato, K. Macdorman, and H. Ishiguro. Generating natual motion in an android by mapping human motion. 2005.

[13] J. D. Morrow. *Sensorimotor Primitives for Programming Robotic Assembly Skills*. PhD thesis, Carnegie Mellon University, 1997.

[14] T. Mukai, T. Mori, and M. Ishikawa. A sensor fusion system using mapping learning method. 1:391–396 vol.1, 1993.

[15] A. Nakazawa, S. Nakaoka, and K. Ikeuchi. Synthesize stylistic human motion from examples. 3:3899–3904 vol.3, 2003.

[16] A. Nakazawa, S. Nakaoka, T. Shiratori, and K. Ikeuchi. Analysis and synthesis of human dance motions. pages 83–88, 2003.

[17] S. Nascimento, B. Mirkin, and F. Moura-Pires. A fuzzy clustering model of data and fuzzy c-means. 2000.

[18] J. Oliver, T. Roush, P. Gazis, W. Buntine, R. Baxter, and S. Waterhouse. *Analysing Rock Samples for the Mars Lander, In: Knowledge Discovery and Data Mining*, pages 299–303. AAAI press, 1998.

[19] R. Palm. Kinematic modeling of the human operator. page 6 pp., 2003.

[20] G. Reybet-Degat and B. Dubuisson. Multisensor fusion with a pattern recognition approach: parametric case. 2:1386–1391 vol.2, 1995.

[21] K. Smith and H. S. W. *Perception and Motor*. W.B. Saunders, 1962.

[22] G. Sukthankar and K. Sycara. A cost minimization approach to human behavior recognition. *AAMAS?05, July*, 2005.

[23] K. S. Trivedi. *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[24] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[25] C. Wallace and D. Dowe. Intrinsic classification by mml - the snob program. *Proc. 7th Australian Joint Conf. on Aritificial Intelligence*, 1994.

[26] W. Xinyu, O. Yongsheng, Q. Huihuan, and X. Yangsheng. A detection system for human abnormal behavior. pages 1204–1208, 2005.

[27] Z. Zhongfei. Mining surveillance video for independent motion detection. pages 741–744, 2002.