# Mining Crucial Features for Automatic Rehabilitation Coaching Systems

Norimichi Ukita
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma
Nara, Japan
ukita@is.naist.jp

Koki Eimon
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma
Nara, Japan
koki-e@is.naist.jp

Carsten Röcker
RWTH Aachen University
Campus-Boulevard 57
52074 Aachen, Germany
roecker@comm.rwth-aachen.de

## ABSTRACT

Our goal is to develop a system for coaching human motions (e.g. rehabilitation). Such a coaching system should have several function such as motion measurement, evaluation, and feedback. Among all, this paper focuses on how to modify a user's motion so that it gets closer to the good template of a target motion. To this end, it is important to efficiently advise the user to emulate the crucial features that define the good template. The proposed method automatically mines the crucial features of any kind of motions from a set of all motion features. The crucial features are mined based on feature sparsification through binary classification between the samples of good and other motions.

## Categories and Subject Descriptors

I.5.4 [**Computing Methodorogies**]: Pattern Recognition-Applications; J.3 [**Computer Applications**]: Life and Medical Sciences

## General Terms

Human Factors

## Keywords

Motion coach, Rehabilitation, Pervasive health, Ambient assisted living

## 1. INTRODUCTION

Most countries of the western hemisphere experienced a tremendous increase in the life expectancy of its citizens. While this trend can be observed for several decades now, there is currently not end in sight. Hence, it is not surprising that the World Health Organization [22] expects about 1.2 billion people over the age of 60 in 2025. Estimations by the United Nations forecast a similar trend and expect nearly 2 billion people to be 60 and older by 2050, which would equal approximately 22% of the world population [7]. At the same time, prolonged life expectancy and increasing survival of acute diseases contribute to a rising number of people suffering from chronic conditions [21]. Already today, more than 75% of elderly people in all economic, social and cultural settings are suffering from chronic diseases [20]. But not only the prevalence of chronic illnesses, also the likeliness of disabilities increases with age, which ultimately leads is an increased risk of falls due to the declining physical abilities. According to studies of the World Health Organization [22] approximately 30% of people over 65 years and 50% of people over 80 years fall each year, and about 20% to 30% of these falls result in serious injuries with long-term consequences for the patients [3]. In almost all cases, continuing physical rehabilitation and training is necessary for enabling patients an independent life after the incident.

While this is mostly done by medical personnel today, we are currently facing the problem that it gets increasingly difficult to find enough caregivers for the growing number of rehabilitation patients due to the demographic and financial constraints that most western nations are experiencing at the moment [1]. Consequently, new approaches are necessary for providing efficient and economically viable rehabilitation solutions for the growing number of people requiring care [12]. As we are moving towards a world of smart devices [11] and intelligent environments [6, 15], automated motion coaching systems, which make use of the sensing infrastructures available in such technology-enhanced home environments [13], are often cited as a promising solution for taking care of the growing number of elderly people requiring physical training [5, 23].

Addressing these challenges we aimed at developing an easy to use system for coaching human movement. Today, common motion measurement systems (e.g. motion capture systems and multi-camera systems[17, 18]) are too expensive and require users to wear binding devices. The proposed system utilizes an inexpensive depth-measurement sensor (i.e. Microsoft Kinect) in order to get high-measurement accuracy with no body-equipped devices. The system functionally consists of three modules. The first one estimates the sequence of body pose from a depth image sequence captured while a user performs a target motion. The second one evaluates the gaps between the estimated pose sequence and that of good template. The third one coaches users on how to modify their motion so that it gets closer to the good template. This paper focuses on achieving the third point. To this end, it is important to efficiently advise the

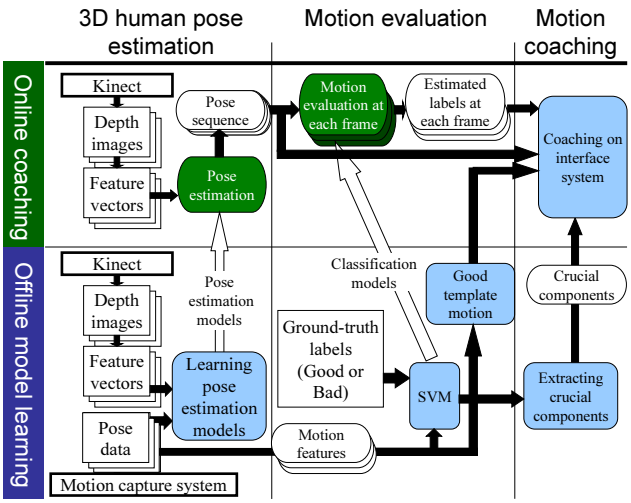



**Figure 1: Overview of the system.**

user to emulate the crucial features that define the good template. This is because many other features of the target motion might be varied among individuals, but those variations give less impacts on evaluating the target motion. The proposed method automatically mines the crucial features of any kind of motions from a set of all motion features. The crucial features are mined based on feature sparsification through binary classification between the samples of good and other motions. The following section provides a more detailed overview over the proposed system.

## 2. SYSTEM OVERVIEW

Figure 1 illustrates the overview of the proposed system consisting of three modules and two steps:

**Offline model learning:** This step is achieved before users are coached by the system. In this step, two kinds of computational models required by the system are prepared. For learning the pose estimation model (i.e. "Pose estimation models" in Fig. 1) that represents the relationship between depth features and human poses, the samples of a target motion are captured by a Kinect and a motion capture system with synchronization (i.e. "Kinect" and "Motion capture system" of "Offline model learning step" in Fig. 1). The pose classification model ("Classification models" in Fig. 1) is learned by the Support Vector Machine [2] ("SVM" in Fig. 1) for evaluating whether the human pose is good or not [1]. In addition, the crucial features of the target motion (i.e. "Crucial features" in Fig. 1) are mined by a sparse coding regularization in the SVM. This mining process is a core contribution of this paper.

**Online coaching:** With the model learned beforehand, the system observes the motion of a user by a Kinect (i.e. "Kinect" of "Online coaching step" in Fig. 1), estimates the human pose at every frame (i.e. "Pose estimation" in Fig. 1), evaluates whether or not each pose is required to be modified (i.e. "Motion evaluation at each frame" in Fig. 1), and coach the user.

[1]We assume that a target motion can be classified into good and other motions. For example, any motion in rehabilitation should be as correct (i.e. good) as possible.

In the online coaching step, the three modules interact with a user as follows:

**3D human pose estimation:** A 3D human pose at each frame is estimated from a depth image captured by a Kinect. While the basic estimation method is based on [14, 4], more discriminative features are used for improving the accuracy in pose estimation instead of simple features proposed in [14, 4]. This is because the proposed coaching system requires users to wait for a feedback after his/her motion is captured (e.g. a few seconds), while real-time processing is required in some other applications (e.g. gaming interfaces). The accuracy of pose estimation is improved also by using the real pose data captured by the motion capture system instead of synthesized CG data.

**Motion evaluation:** The user's pose is evaluated whether it is good or not. If the pose is not good, it is required to be modified so that it gets closer to a good template. This evaluation is achieved by the SVM. Before using the SVM, the pose sequence of the user is temporally synchronized with that of the good template by the dynamic time warping (DTW) [10].

**Motion coaching:** At each subsequence (i.e. several sequential frames) that must be modified, the interface system [19] gives feedbacks to the user. Note that there might be a number of differences between the user's motion and the good template motion, and it is actually impossible to understand all of them simultaneously. The proposed interface system gives the feedbacks one by one. More specifically, the system gives the feedbacks from more crucial features, which define the good template motion.

## 3. MINING CRUCIAL FEATURES VIA FEATURE SPARSIFICATION

For evaluating the motion of a user (i.e. classifying the motion to good or other motions), the SVM is used in the proposed system. This classification is performed with a number of features that represent the 3D pose and its motion of a human body. Since the system should are applicable to any kinds of motions and we do not know which features of a target motion are crucial for defining the target motion, it is better to exhaustively use all features that possibly represent a body motion. In experiments, the concatenation of the following components was used as a 621D feature vector:

- 3D positions of all joints (3D $\times$ 18 joints = 54D)
- 3D velocities of all joints (3D $\times$ 18 joints = 54D)
- 3D accelerations of all joints (3D $\times$ 18 joints = 54D)
- 3D displacement between any pairs of joints (3D $\times$ 153 = 459D)

From the 621 features, the proposed method automatically mines which body parts and/or motions are crucial for improving the motion of a user. This mining is achieved by the sparse coding regularization in the SVM, as proposed in [9].

In classification, the inner product of the feature vector of a test pose (denoted by $\boldsymbol{v}$) and the weight vector $\boldsymbol{w}$is computed. If the inner product is above/below 0, the test pose is regarded as a positive/negative class (e.g. good/others). Therefore if components with a larger absolute value in $\boldsymbol{w}$ correspond to crucial motion features that give a large impact on the inner product. In learning the SVM, the $l_1$ regularized logistic regression [9] is employed so that the gap between larger and smaller absolute values of $\boldsymbol{w}$ gets much greater.

This sparsification can be regarded as dimensionality reduction because the dimensions having smaller values can be neglected. For dimensionality reduction, many other techniques have been proposed (e.g. PCA, Isomap [16], GPLVM [8]). Those techniques, however, cannot provide a user intuitive feedbacks for understanding how to modify the motion. This is because these techniques project a vector from a high-dimensional space to a low-dimensional space defined by arbitrary subspace in the high-dimensional space. That is, each axis in the low-dimensional space might corresponds to multiple axes in the original high-dimensional space. As a result, even if a motion feature corresponding to only one axis is selected for motion coaching, the user might be required to move the body as follows: "you should move the right hand, the right elbow, the left toe, and the hip so that ...". On the other hand, in the proposed method, only one motion feature (e.g. "the right hand" or "the right elbow") is selected from the low-dimensional space generated by the sparse coding regularization.

## 4. EXPERIMENTS

Experiments were conducted with baseball pitching motions[2] captured from 34 people; 13 good (i.e expert) players and 21 beginners. From 34 people, 445 sequences were captured in total. Both of pose estimation and classification models were trained by the data of 33 people, and the data of the remaining one person was used for testing. Note that all 621 features were normalized.

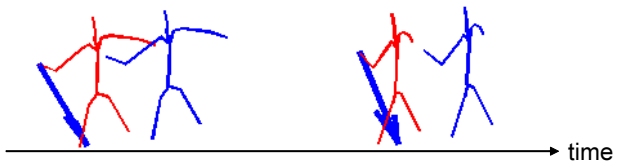
The following two ways were tested for selecting motion features that should be modified by a user in first:

**Naivë selection:** The distance between features of a user's pose and a good template is computed at each feature component (e.g. 3D position of the right hand); the distance at $i$-th feature is denoted by $d_f$. Features having the larger distance are regarded as crucial features.

**Selection with the sparsification:** $d_f$ is multiplied with the weight of $f$-th feature (i.e. $f$-th component of $\boldsymbol{w}$). Features having the larger product are regarded as crucial features.

[2]For validating the proposed system, a sport motion is a good example because its exercise is important for skill proficiency of beginners as well as rehabilitation of experts.


(a) Crucial feature of naivë selection (i.e. left-toe velocity)




(b) Crucial feature selected by the proposed method (i.e. right-hand velocity)

**Figure 2: Visual feedbacks illustrating the difference between the user's motion (shown with red) and the good template (shown with blue) in pitching motions. Two examples in each of (a) and (b) show the selected crucial motions at different two frames.**

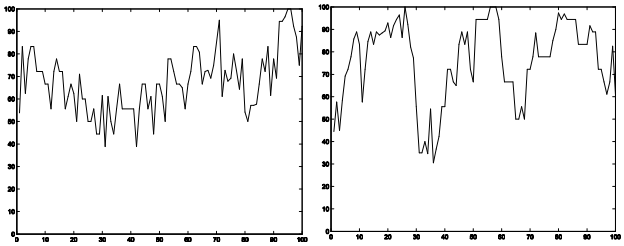
Motions selected by the above two systems were checked by expert players. In examples shown in Fig. 2, naivë selection recommended the left-tow velocity as the most crucial motion (i.e. (a) in Fig. 2), while the right-hand was selected by the proposed method (i.e. (b) in Fig. 2). It is natural that the motion of the hand having a ball is more important for pitching. Actually the experts also mentioned the validity of the selection of the proposed method.

We also demonstrate the effectiveness of the sparsification in motion evaluation. The results are shown in Fig. 3. Each graph shows the accuracy of motion classification (i.e. good or not) at each frame. The mean classification rate of all 445 sequences, each of which was evaluated by leave-one-out cross-validation, is shown at each frame.

The means of the classifications rates in all frames were 67 % and 76 % in (a) and (b), respectively. These results demonstrate the effectiveness of the sparsification also in motion classification. This effect is gained because, rather than a high-dimensional feature space, a low-dimensional feature space allows us to improving the generalizing capability of classifiers such as the SVM, which was implemented with LIBSVM [2] in the proposed system.

## 5. CONCLUSIONS

This paper proposes how to mine the crucial features in any kind of motions. The crucial motions are mined by the sparse coding regularization during training of the SVM that classifies target motions into good or not. The weight vector of the SVM shows which features are crucial for classifying whether a user's motion is good or not. In particular, the sparse coding regularization allows us to enhance the difference between crucial and non-crucial features.



(a) all features (b) features selected by sparse coding regularization

**Figure 3: Temporal histories of the mean classification rates. After each test sequence is virtually synchronized with a good template motion by the DTW, the mean classification rate was computed at each frame. In (b), only four features were selected in decreasing order of the weight in $w$.**

Experimental results demonstrated that 1) the proposed method could extract intuitively-correct crucial features and 2) the extracted features could improve the accuracy in motion classification.

A part of this research is supported by SCOPE (132307013) of Ministry of Internal Affairs and Communication Japan.

# 6. REFERENCES

[1] Adam, S., Mukasa, K. S., Breiner, K., and Trapp, M. An apartment-based metaphor for intuitive interaction with ambient assisted living applications. In *BCS HCI (1)* (2008), pp. 67–75.

[2] Chang, C.-C., and Lin, C.-J. Libsvm: A library for support vector machines. *ACM TIST 2*, 3 (2011), 27.

[3] de Ruyter, B. E. R., and Pelgrim, E. Ambient assisted-living research in carelab. *Interactions 14*, 4 (2007), 30–33.

[4] Girshick, R. B., Shotton, J., Kohli, P., Criminisi, A., and Fitzgibbon, A. W. Efficient regression of general-activity human poses from depth images. In *ICCV* (2011), pp. 415–422.

[5] Holzinger, A., Ziefle, M., and Röcker, C. Human-computer interaction and usability engineering for elderly (hci4aging): Introduction to the special thematic session. In *ICCHP (2)* (2010), pp. 556–559.

[6] Kasugai, K., Ziefle, M., Röcker, C., and Russel, P. Creating spatio-temporal contiguities between real and virtual rooms in an assistive living environment. In *Create 10, the conference for innovative interaction design* (2010), pp. 62–67.

[7] Laleci, G. B., Dogac, A., Olduz, M., Tasyurt, I., Yuksel, M., and Okcan, A. Spahire: A multi-agent system for remote healthcare monitoring through computerized clinical guidelines. In *Agent Technology and E-Health* (2007), pp. 25–44.

[8] Lawrence, N. D. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research 6* (2005), 1783–1816.

[9] Li, L.-J., Su, H., Xing, E. P., and Li, F.-F. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS* (2010), pp. 1378–1386.

[10] Myers, C. S., and Rabiner, L. R. Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition. *The Bell System Technical Journal 60*, 7 (1981).

[11] Röcker, C., and Etter, R. Social radio: a music-based approach to emotional awareness mediation. In *IUI* (2007), pp. 286–289.

[12] Röcker, C., Ziefle, M., and Holzinger, A. Social inclusion in aal environments: Home automation and convenience services for elderly users. In *ICAI* (2011), pp. 55–59.

[13] Röcker, C., Ziefle, M., and Holzinger, A. From computer innovation to human integration: Current trends and challenges for pervasive health technologies. In *Pervasive Health - State-of-the-Art and Beyond*. Springer, 2014.

[14] Shotton, J., Fitzgibbon, A. W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. Real-time human pose recognition in parts from single depth images. In *CVPR* (2011), pp. 1297–1304.

[15] Streitz, N. A., Prante, T., Röcker, C., van Alphen, D., Stenzel, R., Magerkurth, C., Lahlou, S., Nosulenko, V., Jegou, F., Sonder, F., and Plewe, D. A. Smart artefacts as affordances for awareness in distributed teams. In *The Disappearing Computer* (2007), pp. 3–29.

[16] Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (2000), 2319–2323.

[17] Ukita, N., Hirai, M., and Kidode, M. Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *ICCV* (2009).

[18] Ukita, N., and Kanade, T. Gaussian process motion graph models for smooth transitions among multiple actions. *Computer Vision and Image Understanding 116*, 4 (2012), 500–509.

[19] Ukita, N., Kaulen, D., and Röcker, C. Towards an automatic motion coaching system: Feedback techniques for different types of motion errors. In *International Conference on Physiological Computing Systems* (2014), pp. 167–172.

[20] Vergados, D. J., Alevizos, A., Mariolis, A., and Caragiozidis, M. Intelligent services for assisting independent living of elderly people at home. In *PETRA* (2008), p. 79.

[21] Villar, A., Federici, A., and Annicchiarico, R. K4care: Knowledge-based homecare eservices for an ageing europe. In *Agent Technology and E-Health* (2007), pp. 141–148.

[22] World Health Organization. *Active Aging: A Policy Framework*, 2012.

[23] Ziefle, M., Röcker, C., Wilkowska, W., Kasugai, K., Klack, L., Mollering, C., and Beul, S. A multi-disciplinary approach to ambient assisted living. In *E-Health, Assistive Technologies and Applications for Assisted Living: Challenges and Solutions*. 2014, pp. 76–93.