Robotics and Autonomous Systems 66 (2015) 35–43

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Correlated space formation for human whole-body motion primitives and descriptive word labels

Wataru Takano *, Seiya Hamano, Yoshihiko Nakamura

Mechano-Informatics, University of Tokyo, 7-3-1, Hongo, Bunkyoku, Tokyo, 113-8656, Japan

HIGHLIGHTS

- We construct correlated spaces of human motions and word labels.
- The correlated space can be applied to searching for motions from word queries.
- The motions can be retrieved, even if the queries are not assigned to the motions.
- This technology can be helpful for reusing the motion data.

ARTICLE INFO

Article history:
Received 17 April 2014
Received in revised form
22 September 2014
Accepted 1 November 2014
Available online 13 January 2015

Keywords:
Motion primitive
Word label
Canonical correlation analysis

ABSTRACT

The motion capture technology has been improved, and widely used for motion analysis and synthesis in various fields, such as robotics, animation, rehabilitation, and sports engineering. A massive amount of captured human data has already been collected. These prerecorded motion data should be reused in order to make the motion analysis and synthesis more efficient. The retrieval of a specified motion data is a fundamental technique for the reuse. Imitation learning frameworks have been developed in robotics, where motion primitive data is encoded into parameters in stochastic models or dynamical systems. We have also been making research on encoding motion primitive data into Hidden Markov Models, which are referred to as "motion symbol", and aiming at integrating the motion symbols with language. The relations between motions and words in natural language will be versatile and powerful to provide a useful interface for reusing motion data. In this paper, we construct a space of motion symbols for human whole body movements and a space of word labels assigned to those movements. Through canonical correlation analysis, these spaces are reconstructed such that a strong correlation is formed between movements and word labels. These spaces lead to a method for searching for movement data from a query of word labels. We tested our proposed approach on captured human whole body motion data, and its validity was demonstrated. Our approach serves as a fundamental technique for extracting the necessary movements from a database and reusing them.

1. Introduction

To understand real-world phenomena that are not clearly separated, humans segment those phenomena and perceive them as symbols. Rather than being based on physical properties, this segmentation is arbitrary and depends on the society to which the person belongs [1]. However, these symbols have been refined by recording the correspondence between the arbitrarily segmented world and the symbols used for denoting it, as well as the relation between the different symbols, and through the evolutionary process of the cumulative utilization of these symbols. This immense system of intricately intertwined symbols sublimated into language, allowing humans to communicate efficiently with one another and to perform high-order reasoning. It can be said without exaggeration that the high-order cognitive capabilities of humans are a product of language.

For humanoid robots to coexist with humans, they will have to be able to use the same symbols and language systems as humans. Research on robot body motion has focused on imitating learning methods that optimize the parameters of mathematical models based on various motion patterns [2,3]. In this framework, time-series data (e.g., data about the joint angles representing motions) are memorized as symbols represented discretely by parameters of a statistical model [4,5] or of a dynamical system [6–10]. This sort

* Corresponding author. Tel.: +81 3 5841 6378; fax: +81 3 3818 0835.
E-mail addresses: takano@ynl.t.u-tokyo.ac.jp (W. Takano), hamano@ynl.t.u-tokyo.ac.jp (S. Hamano), nakamura@ynl.t.u-tokyo.ac.jp (Y. Nakamura).

http://dx.doi.org/10.1016/j.robot.2014.11.020

of robot intelligence is the ability not only to understand human behavior by comparing human motions with previously memorized motions but also to generate continuous motions from memorized symbolic representations of motions to apply them to the real world. The expansion of the range of fields where humanoid robots are used is creating an increasingly strong demand for a framework to memorize many motion symbols [11].

With the improvement and spread of optical motion capture technologies, data about human body motions is getting applied not only in robotics but also in various other fields, such as animation, sports engineering and rehabilitation. A massive amount of data about human body motion has already been collected. However, when reusing memorized motion data in order to synthesize new motions for animated characters or to perform motion analysis by comparing motions with previous ones, efficient search techniques should be used to find the necessary motions in the collected data. Currently, searching for and reusing the necessary motion data is based on labels such as measurement time or motion description. An environment where it is possible to search for only labels that match the input data places a large burden on operators that reuse the motion data by requiring them to memorize the exact measurement time or motion description.

The ability of a robot to perform intelligent information processing by encoding and categorizing large amounts of body motion data and linking that data to linguistic representations forms the basis of the robot's comprehension of language and body motion. Also, this is closely related to technology for searching and presenting motion data related to simple linguistic input. This ability would substantially improve the reusability of motion data in motion analysis or motion generation for CG characters. Frameworks proposed thus far have been based on arrays of motion symbols representing body motion data learned through a Hidden Markov Model (HMM) [12] and arrays of verb labels attached to those motions. In those frameworks, emphasis is placed on learning the correspondence between motions and verbs by restricting the linguistic representations to verbs and considering the context of the symbol and verb arrays, however, without taking into account the interrelation between motions or other linguistic units, such as nouns or adverbs [13]. One proposed method for expressing the interrelation between motion symbols involves calculating the distances between individual motion symbols and constructing a multidimensional motion symbol space and assigning motion symbols to points such that those distances are preserved [14]. Furthermore, there are language processing techniques in which sentences composed of verbs, nouns and other elements are represented as points in a vector space based on the presence, absence or frequency of the constituent words [15–17]. Thus, it may be possible to construct a computational model connecting motions to word labels by using a common representation of both motions and word labels as points in some spaces. In this paper, we construct a space of motion symbols learned by applying an HMM to body motion data and a word label space consisting of verbs, nouns and other words assigned to those motions. Next, through canonical correlation analysis [18], these spaces are reconstructed such that a strong correlation is formed between motion symbols and word labels. Using these spaces, we propose a method for searching for motion data based on word labels. This serves as a fundamental technique for extracting the necessary motions from a database and reusing them.

2. Mapping between motions and word labels

Research on intelligent robots through conversion of bodily senses or movements into symbols is being conducted in robotics. These approaches encode the continuous spatio-temporal data of motions into the low dimensional parameters of motion primitives, and these parameters allow robots to classify the motions into the motion primitives. However, the motion primitives represented in the parameters cannot be intuitively understood by humans. Humans have acquired language through the process of evolution. We can understand motions in same expression that others can do by using the language. The mapping between the motion parameters and words is crucial to establishing communication between robots and humans. This section describes an approach to extract the mapping between motions and word labels. The motion data is encoded into a Hidden Markov Model (HMM), which is referred to as "motion symbol". The motion data is also given word labels by human annotators. Relation between the motions and the word labels is extracted from the training pairs of the motion symbols and the word labels as shown in Fig. 1.

2.1. Extracting correlation between motions and words

Fig. 2 shows the overview of mapping between human whole body motions and word labels. The human motion primitive data are encoded in Hidden Markov Models (HMMs). Each HMM is referred to as "motion symbol" since it represents spatio-temporal features of its corresponding motion primitive. Dissimilarity between each motion symbol can be calculated by using the Kullback–Leibler information.

$$d(\lambda_i, \lambda_j) = \sum_{\hat{\mathbf{o}}_i^{(k)}:\, k=1,2,3,\ldots,N} \frac{1}{N} \left\{ \ln P(\hat{\mathbf{o}}_i^{(k)}|\lambda_i) - \ln P(\hat{\mathbf{o}}_i^{(k)}|\lambda_j) \right\} \quad (1)$$

$d(\lambda_i, \lambda_j)$ is the Kullback–Leibler information from motion symbol λ_i to motion symbol λ_j. $\hat{\mathbf{o}}_i^{(k)}$ is the kth motion data that the motion symbol λ_i generates by the Monte Carlo method. $P(\hat{\mathbf{o}}_i^{(k)}|\lambda_j)$ is the likelihood that motion symbol λ_j generates the motion data $\hat{\mathbf{o}}_i^{(k)}$. The Kullback–Leibler information does not necessarily satisfy the symmetry. In Eq. (2), $d(\lambda_i, \lambda_j)$ and $d(\lambda_j, \lambda_i)$ are summed to obtain $D(\lambda_i \parallel \lambda_j)$, which satisfies the symmetry.

$$D(\lambda_i \parallel \lambda_j) = \frac{d(\lambda_i, \lambda_j) + d(\lambda_j, \lambda_i)}{2}. \quad (2)$$

This is defined as the distance between motion symbol λ_i and motion symbol λ_j. All of the motion symbols are arranged as points on a multidimensional space such that the distance between all the motion symbols is satisfied. The coordinates of the point in the multidimensional space corresponding to motion symbol λ_i are taken as $\boldsymbol{x}_i$, and this position is found such that the following error function is minimized.

$$T = \sum_{\forall i,j} \frac{\left(D(\lambda_i \parallel \lambda_j)^2 - d_{ij}^2\right)^2}{4D(\lambda_i \parallel \lambda_j)^2} \quad (3)$$

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \quad (4)$$

Here, the multidimensional scaling proposed by Takane et al. [19] is used. The error function T is represented by a fourth-order polynomial in coordinate $\boldsymbol{x}_i$ of motion symbol λ_i. The optimal position of the motion symbol, which minimizes the error function T, can be found by the Newton–Raphson method. This process constructs the motion symbol space based on dissimilarity between the motion symbols by the multidimensional scaling.

Multiple word labels are manually assigned to the same motion primitive data that the HMM encodes into the motion symbol. The word labels are descriptive of the motion primitive. A set of the word labels is represented by a feature vector with binary elements taking value 1 if the corresponding word label is present in the set