Neural Networks 80 (2016) 1–8

Contents lists available at ScienceDirect

Neural Networks

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

Generating action descriptions from statistically integrated representations of human motions and sentences

Wataru Takano *, Ikuo Kusajima, Yoshihiko Nakamura

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

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 21 November 2015
Accepted 3 March 2016
Available online 16 March 2016

Keywords:
Motion primitive
Natural language
Sentence generation

ABSTRACT

It is desirable for robots to be able to linguistically understand human actions during human–robot interactions. Previous research has developed frameworks for encoding human full body motion into model parameters and for classifying motion into specific categories. For full understanding, the motion categories need to be connected to the natural language such that the robots can interpret human motions as linguistic expressions. This paper proposes a novel framework for integrating observation of human motion with that of natural language. This framework consists of two models; the first model statistically learns the relations between motions and their relevant words, and the second statistically learns sentence structures as word n-grams. Integration of these two models allows robots to generate sentences from human motions by searching for words relevant to the motion using the first model and then arranging these words in appropriate order using the second model. This allows making sentences that are the most likely to be generated from the motion. The proposed framework was tested on human full body motion measured by an optical motion capture system. In this, descriptive sentences were manually attached to the motions, and the validity of the system was demonstrated.

1. Introduction

Symbolization and anthropomorphism are inherent outcomes of human intelligence. These modes of thought are intimately connected to the human body.

Humans have invented tools, learned how to use these tools, and lived in various environments during the process of our evolution. This environmental variety requires humans to understand many objects, actions, and events. Symbols may be an efficient cognitive system for tackling this variety to allow humans to memorize concepts (the signified) as compact forms (the signifiers), and reuse these signifiers in new situations to not only understand concepts but also transfer these concepts. According to Saussure, the relation between the signifier and the signified is arbitrary; specifically, there is no direct connection between the word and the concept to which the word refers. The arbitrary nature of the relation between the signifier and the signified contributes to the manipulability of the symbolic systems that have developed into language (Saussure, 1966).

Humans uses gestures as one communication channel. Communication and language may depend on processing by the human body. We can project the actions of others onto our own bodies. More generally, we simulate actions in our body system, and estimate sensations in others; this enables us to share the actions and understand the intent. This ability extends to understanding animals, despite their differences from humans. This understanding – anthropomorphism – may involve mirror neurons (Gallese & Goldman, 1998; Rizzolatti, Fogassi, & Gallese, 2001).

Research on representations of human motion for humanoid robots has been inspired by symbolization and anthropomorphism. Humanoid robots have a high number of degrees of mechanical freedom, and it is difficult to manually program their full body motions. Programming-by-demonstration and imitation learning (Argall, Chernova, Veloso, & Browning, 2009; Breazeal & Scassellati, 2002) have been developed as solutions for overcoming these difficulties. In these frameworks, motion trajectories are encoded into model parameters. This encoding relies on the assumption that spatio-temporal features of a continuous motion trajectory can be represented by one point in parameter space, and each such point can be regarded as a symbolic representation of the corresponding motion. Robots can then memorize motions as symbolic forms.

In robotics, the symbolization of motions has become an active area of research, drawing on theory and understanding from

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

http://dx.doi.org/10.1016/j.neunet.2016.03.001
0893-6080/

semiology, linguistics, and brain science. However, most research does not focus on acquisition of language from motions but, instead, on representations of human or robot motion. For a robot to fluidly interact with humans, it needs to acquire a linguistic representation of motions, understand human motions in the form of language, and generate language expressions signifying the motions. Otherwise, human partners cannot understand how the robot interprets actions, robots partners cannot convey requests to humans to perform specific actions. This paper proposes a novel framework for bidirectional conversion between human motions and descriptive sentences. This framework consists of two modules: a statistical model for mapping between motions and their relevant words (a motion language model) and a statistical model representing sentence structure by word n-grams (a natural language model). The humanoid robot interprets observation of human behavior as sentences and generates human-like motions from linguistic commands by integrating these two modules. We tested our framework on captured human full body motions and evaluated the sentences generated from the motions, varying the complexity of the natural language models during testing.

2. Related work

Motions are characterized by spatio-temporal features, and these features can be encoded into parameters of dynamical systems, such as neural networks and systems of differential equations (Ijspeert, Nakanishi, & Shaal, 2003; Okada, Tatani, & Nakamura, 2002; Tani & Ito, 2003). One useful method is to tune the parameters such that an attractor for the motion pattern embedded in the dynamical system is the result. The resulting basins of attraction in the system can be helpful to robots for generating stable motion, even in the presence of a disturbance from the external environment. The system can also be used as a predictor for motion, and this function allows for motion recognition. Statistical models have been widely used for encoding motions, with a hidden Markov model (HMM) as a typical encoding (Asfour, Gyarfas, Azad, & Dillmann, 2006; Billard, Calinon, & Guenter, 2006; Inamura, Toshima, Tanie, & Nakamura, 2004). The statistical parameters can be optimized such that the likelihood of the training motion being generated by the model is maximized. The motion can be classified by finding the model with the largest likelihood of generating the observed motion, and the motion trajectory can be generated according to statistical transitions of the postures in the model. These frameworks described above make it possible to encode continuous motion trajectories into model parameters, which then form the symbolic representation of the motion. Such symbolic representations act as motion classifiers or motion synthesizers. However, this type of representation cannot be easily understood by humans since the motions are signified by an indexed set of models. As an example, suppose that the motions of "walking" and "running" are signified by models 1 and 2, respectively. The motions signified by these indices are not human understandable. As an alternative, each motion should be connected to another signifier that is understandable, and natural language is a suitable choice.

Sugita et al. proposed an approach for connecting two neural networks with bias parameters (Sugita & Tani, 2005). One network learns sensory motor data, and the other network learns word transitions. These two networks share bias parameters, and motion and language are both encoded according to these parameters. The bias parameters thus combine motions with words and allow for generating descriptive words from motions. Ogata et al. developed an algorithm for generating motions from word queries by using combined neural networks. A candidate motion is generated from the word query, and a word is then generated from that motion. Their algorithm searches for the motion that generates the query (Ogata, Murase, Tani, Komatani, & Okuno, 2007). We have also proposed a statistical framework that connects motions to their relevant words. The motions are encoded as motion primitives (HMMs in the framework). Additionally, words are manually assigned to motions. The translation model then learns the mapping between the motion primitives and words. This framework makes it possible for robots to convert motions into descriptive words and to generate motions from word queries (Takano, Hamano, & Nakamura, 2015; Takano & Nakamura, 2015a). This framework handles only verbs, and we have extended the mapping between motions onto words to handle the sentence structure (Takano & Nakamura, 2015b), where the sentence structure was simply represented by word bigram, the effect of the intrinsic structure for the mapping between the motions and the words to generation of the sentences was not evaluated according to the score popular in the natural language community. This paper adopts the word bigram and word trigram for the sentence structure, varies the complexity of mapping between the motions and words, and quantitatively evaluates the generation of descriptive sentences from the human motions.

In computer graphics, research has been conducted into synthesizing character motions from words. Arikan et al. assigned word labels to motion frames in a motion database, and developed a method for searching for smooth sequences of motions up to the frame to which the query word is attached (Arikan, Forsyth, & O'Brien, 2003). Rose et al. broke motions up into groups ("verbs", in their system) and extracted the differences among groups as parameters ("adverbs") (Rose, Bodenheimer, & Cohen, 1998). They proposed a new technique for interpolating between two motions by controlling the parameters. Their framework introduced the new concepts of verbs and adverbs in motion space, but verbs and adverbs have a clear link to natural language. Our framework establishes a connection between motion categories and the corresponding natural language sentences.

3. Connection between motions and descriptive sentences

Our framework integrates motion representations with sentence structures. Human full body motion is represented by a sequence of configurations, such as joint angles or joint positions. The sequences are encoded into HMMs, each of which is referred to as a "motion primitive". The relations between motion primitives and the corresponding words are statistically extracted as shown on the left panel in Fig. 1, and the resultant model is referred to as the "motion language model". The other model in the right panel in Fig. 1 statistically represents the transitions between words in sentences. This model is referred to as the "natural language model". The combination of the motion language model and the natural language model enables robots to not only understand human motions as sentences but also generate robot motions from a sentence command.

The motion language model consists of three layers: motion primitives λ, latent states s, and words ω, as shown in Fig. 2. The motion primitives can be derived by encoding motion data such as a sequence of joint angles into an HMM, and the word is included in the descriptive sentence for the motion. These three components are statistically connected by the probability $P(s|\lambda)$ of the latent state s being generated by the motion primitives λ, and the probability $P(\omega|s)$ of the word ω being generated by the latent state s. Given the training dataset of the motion primitive $\lambda^{(k)}$ for the kth human full body motion and sentences $\omega_1^{(k)}, \omega_2^{(k)}, \ldots, \omega_{l_k}^{(k)}$ assigned to the same motion, the probabilities $P(s|\lambda)$ and $P(\omega|s)$ are optimized by maximizing the following objective function, for which we use an expectation–maximization algorithm. Each expectation step estimates the distribution of the latent states