A Probabilistic Approach to Text Generation of Human Motions extracted from Kinect Videos

Mizuki Kobayashi, Ichiro Kobayashi, Hideki Asoh, and Sergio Guadarrama

***Abstract*—In this study, we propose a framework for probabilistic text generation of human motions extracted from Kinect videos. We capture human motions by a Kinect camera and extract the time-series data of the motions from the videos. The time-series data are applied by several dimension reduction procedures and then turned to be the form which can be applied to machine learning. A pair of the analyzed time-series data and its intermediate representation which corresponds to the semantics of the human motion is learned by a log-linear model. As linguistic resources to generate a text, we collected various natural language expressions for human motions and build a bi-gram model for each motion. In our framework, once the intermediate representation is decided by observing time-series data; a proper bi-gram model corresponding to the intermediate representation is chosen; and then a text is generated by solving dynamic programming of the bi-gram model. Through experiments to generate texts describing human motions, we have confirmed that our proposed framework works well.**

***Index Terms*—probabilistic text generation, bi-gram, Symbolic Aggregation approXimaion(SAX), time-series data, Kinect**

I. Introduction

IT has recently been getting easier to obtain huge amount of moving pictures. Whereas, it cannot say that we can well utilize those data for particular purposes — for example, in order to grasp the content of the videos recorded by a surveillance camera, we need to watch through all the videos, but that is considerably time-consuming work. At this point, if events happened in a video can be recognized and be described by natural language sentences, it will be easy for us to grasp the content of the videos and achieve various applications such as scene retrieval by words, etc. Considering this, in this study we propose a framework for probabilistic text generation with visual information as input information.

II. Related studies

As the studies related to text generation with multimedia information as input information, Ding et al. [1], [2] have built a system that generates textual summaries of Internet-style video clips. In the study of Tan et al. [3], a variety of visual and audio concepts in video contents are classified and the classification results are applied to simple rule-based methods to generate textual descriptions of video contents. Kobayashi et al. [4] have proposed a method to verbalize human behaviors in a room. Barbu et al. [5] have developed a system that produces sentential descriptions of short video clips. These sentences describe who did what to whom, and where and how they did it. The text generation methods adopted in these studies is basically based on templates-based generation or generation using a small set of grammar.

To flexibly generate texts, many studies taking a probabilistic approach to text generation have been so far studied. Lapata [6] has built a model that learns constraints on sentence order from a corpus of domain-specific texts and an algorithm that yields the most likely order among several alternative. Belz and Kow [7], [8] have proposed a framework that combines probabilistic generation method with a comprehensive model of the generation space, and built a system that can generate weather forecast texts. Lu et al. [9] have proposed a text generation model with hybrid tree representation in which both the meaning representation and natural language are encoded in a tree, and showed their model performs better than a previous state-of-the-art natural language generation model.

As the studies most related to our study, we can take up the studies by Liang et al.[10], Angeli et al. [11], and Konstas et al. [12], [13]. Liang et al. [10] have proposed a method to learn the correspondence between a text and its semantics with less supervision and domain independent. In their framework, the semantics corresponds to the data base records which store the information about the states of events in themselves, and the correspondence between parts of the records and the segments of natural language description of the content of the records is obtained by machine learning. Angeli et al. [11] propose a text generation model that unifies content selection and surface selection based on the model proposed by Liang et al. [10], and have introduced decision making into the generation process. Konstas et al. [12], [13] define a probabilistic context-free grammar that globally describes the inherent structure of the input. They also use database records and text describing some of them as input information as well as the studies by Liang et al. [10] and Angeli et al. [11]. They represent their grammar as a weighted hypergraph and generate a text as the task of finding the best derivation tree for a given input.

We have employed some ideas of these most related studies in our study – we have introduced a log-linear model to learn correspondences between analyzed time-series data and natural language descriptions of human motions. Although our text generation method is simple and cannot generate a complicated sentence which grammar is required to generate it, a simple but likely sentence can be easily generated with a moving picture as input information.

III. Overview of proposed framework

The overview of our proposed framework is illustrated in Figure 1. At first, the time-series data of human motions

Mizuki Kobayashi and Ichiro Kobayashi are with Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, Tokyo, Japan, e-mail : {kobayashi.mizuki,koba}@is.ocha.ac.jp.

Hideki Aso is with Intelligent Systems Research Institutey in National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, e-mail : h.asoh@aist.go.jp.

Sergio Guadarrama is with Electrical Engineering and Computer Sciences, UC Berkeley, California, USA e-mail : sguada@eecs.berkeley.edu.

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

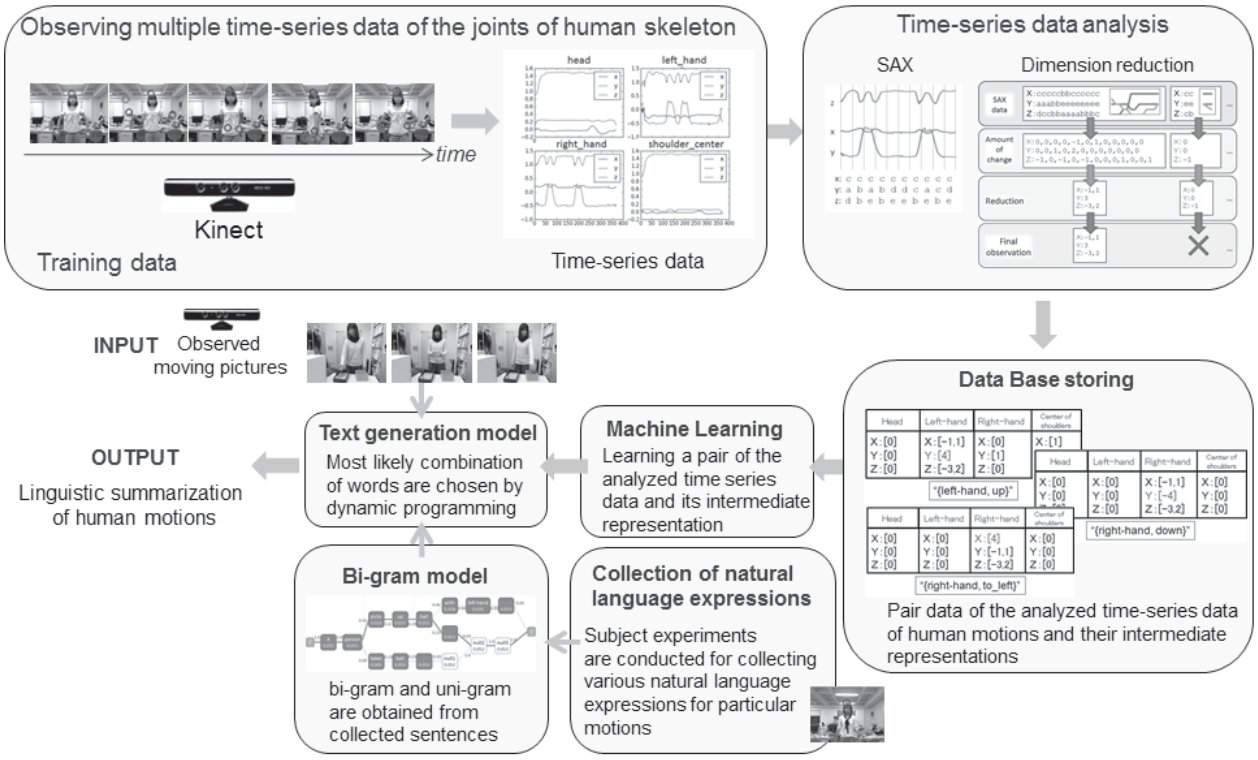

Fig. 1. Overview of probabilistic text generation of human motions

are recorded by tracking multiple joints of human skeleton with the libraries for a Kinect camera[14]. Several dimension reduction procedures are applied to the observed time-series data, and then the analyzed data are stored in a data base with the intermediate representations which correspond to the semantics of human motions and bridge the gap between time-series data and natural language sentences. After that, by adopting machine learning for the correspondence between analyzed time-series data and an intermediate representation, we build a human motion identifier with visual information as input information. To build linguistic resources used for text generation, we conduct a subject experiment to collect sentences which describe human motions and then build bi-gram models based on the collected sentences for each intermediate representation. So, once an intermediate representation is selected, a corresponding bi-gram model to the representation is selected, and then the most likely combination of words is selected as a linguistic summary for the human motion by applying dynamic programming to the selected bi-gram model.

A. Processing of time-series data

We obtain the time-series data of human motions with a Kinect camera. MicroSoft has provided a Kinect camera with standard software libraries which enable to estimate the position of each joint of human skeleton, and human position can be estimated by using 3-dimensional data by each joint. In this study, the positions of human joints are estimated with RGB information and the information observed by a depth sensor, and then the time-series data of the axioms of x,y, and z (depth) of the four positions: head, the center of shoulders, right and left hands are estimated (see, Figure 2).

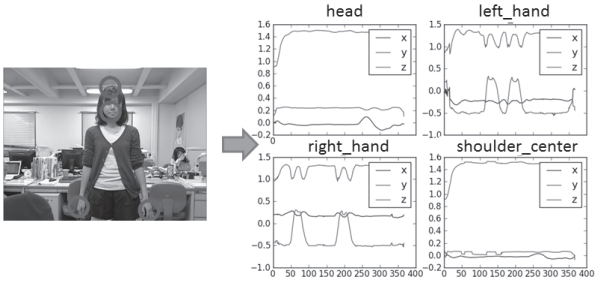

Fig. 2. Time-series data obtained by a Kinect camera

The time-series data of tracking multiple joints of human skeleton are converted into a series of letters by means of Symbolic Aggregation approXimation (SAX) [15].

The human motions are extracted from a series of letters obtained by SAX. Here, we regard that a human being did not move if the letters of all observing joints did not change from the previous states, and also regard that she or he moved if any letters change from the previous states (see, Figure 3).

In a series of letters, the parts where human motions are observed are translated into numerical values which show the amount of changes (see, Figure 4), and aggregated into more simple numerical values (see ‘Reduction’ part in Figure 5). This process enables to identify the same actions even though a series of letters is slightly different because of the different positions or speeds of human movements. Furthermore, in order to extract main movements, the movements whose amount of changes do not exceed 2 are removed as observation errors (see ‘Final observation’ in Figure 5).

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 3. An example of extracting a motion

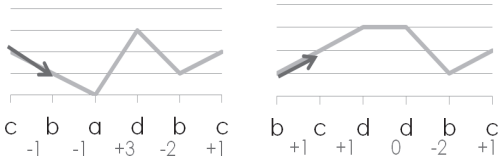

Fig. 4. Amount of changes in a series of letters

B. Intermediate representations for human motions

In our text generation framework, the intermediate representations, which bridge the gap between time-series data and natural language sentences, are required to decide which linguistic resources should be used to generate a text. We have defined a small set of the intermediate representations only enough for expressing simple human motions employed in this study (see, Table I).

TABLE I
INTERMEDIATE REPRESENTATIONS FOR HUMAN ACTIONS

action	intermediate representation	meaning
up	''{object, up}''	upward movement
down	''{object, down}''	downward movement
left	''{object, to_left}''	leftward movement
right	''{object, to_right}''	rightward movement
pass	''{object1, object2, pass}''	cooperative movement
swing	''{object, swing}''	movement of swinging

The objects in the intermediate representations are the joints of human skelton, e.g., head, left and right hands, etc.

C. Motion identification from time-series data

In oder to obtain an identifier of human motions, we adopt a log-linear model to learn the correspondence between the analyzed time-series data and the intermediate representations. Here, d indicates the final observation data after processing time-series data explained in section III-A, and y indicates the intermediate representations of human motions. By using feature vector ϕ consisting of d and y, $P(y|d)$ is modeled with the log-linear model expressed in equation (1). Here, $Z_{d,\mathbf{w}}$ is a coefficient for normalization.

$$P(y|d) = \frac{1}{Z_{d,\mathbf{w}}} \exp(\mathbf{w} \cdot \phi(d, y)) \quad (1)$$

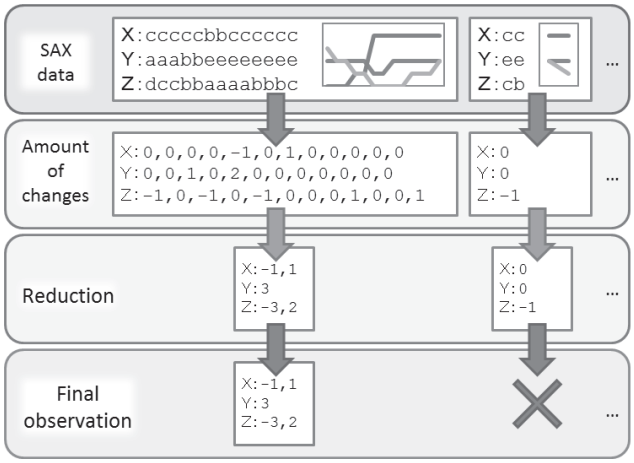

Fig. 5. The process of data compression and selection

D. Text generation based on a bi-gram model

We employ a simple text generation method based on a bi-gram model. To build a bi-gram model for each human motion, we conduct subject experiments to collect various natural language descriptions to express a particular human motion.

Once a particular intermediate representation for observed time-series data is chosen, a bi-gram model is chosen as the linguistic resources to generate a text. However, there are several ways of describing a human motion, some people might describe a motion with 10 words, the other people might describe the motion with 15 words. Considering this, we introduce "null" label into the bi-gram model so that the most likely sentence can be generated without depending on the lenght of a sentence. The "null" labels are treated as the same as words in a setence, in other words, each of them has uni-gram and bi-gram as well as the other words. To deal with the "null" labels in that way, the following pre-processing for each sentence is required before applying dynamic programming to the bi-gram model – first, we obtain the maximum and minimum length of sentences. Secondly, we obtain the value of subtracting the minimum number from the maximum number, which corresponds to the maximum number of "null" labels used in the sentence with the minimum length. Thirdly, if a sentence is not the one with the maximum length, "null" labels with a number are inserted in decsending order from the end of the sentence toward the beginning of the sentence. Figure 6 shows an image of introducing "null" labels in sentences.

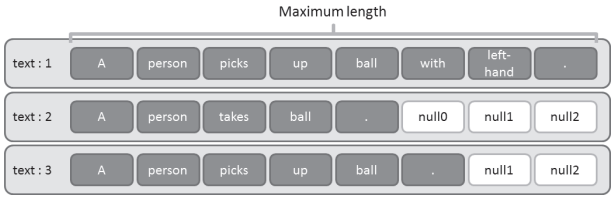

Fig. 6. An image of introducing "null" labels

By inserting "null" labels with different numbers, they can be treated as different words, in other words, each of them can be treated as a part of bi-gram model. Furtheremore, in this study when constructing a bi-gram model, we remove

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

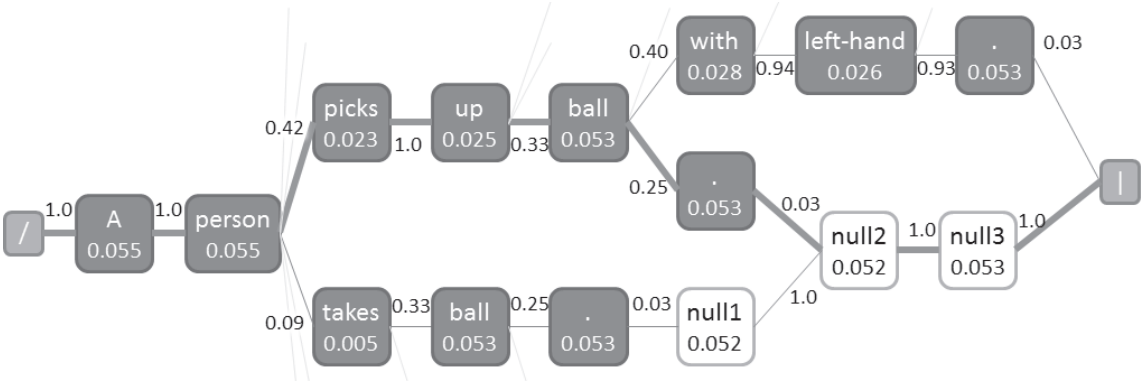

Fig. 7. Bi-gram model with null labels

definite and indefinite from the sentences, because both can be associated with many words, therefore, it will be difficult to generate a proper sentence if they are in a bi-gram model. Figure 7 illustrates a bi-gram model for describing a human motion: *pick up a ball*.

We apply dynamic programming to a bi-gram model to generate the most likely sentence for describing a human motion.

IV. Experiment

We conducted an experiment to express a simple human action, in which a person picks up a ball and puts it in a box (see, Figure 8), with natural language sentences.

Fig. 8. The target behavior to be described with natural language

A. Experimental settings

First, we predefine the target human behavior consists of three motions: *pick*, *pass*, and *put*. This is because it is quite difficult to decide automatically which part in human action should be described by natural language sentences. Here, we decide that natural language description of a human action is generated in each motion. We conducted an experiment in which a subject watches Kinect videos of the target human action and describes it with natural language. The number of subjects were 12. As for the features of the collected natural language sentences to explain each motion, the number of sentences and the number of words and kinds of words which appear in the sentences are shown in Table II. Based on the collected sentences, we construct a bi-gram model.

We constructed an identifier of human motions by means of a log-linear model through 5 trials by using 15 training data and 5 evaluation data which are randomly selected among all 20 data of the same human motion as the target motion to be described by natural language. The average of the accuracy of the identifier is 84 %. We used this identifier to decide the intermediate representation for text generation.

TABLE II
Features of collected sentences

Motion	Sentences	words	Kinds of words
1	33	274	47
2	18	142	28
3	36	290	28

B. Result

As a result, as for the first motion, the intermediate representation of the motion is recognized as ‘‘{left_hand, up}’’, and as for the second motion, the intermediate representation of the motion is recognized as ‘‘{left_hand, right_hand, pass}’’, as for the third motion, the intermediate representation of the motion is recognized as ‘‘{right_hand, down}’’. These results are the ones expected to be chosen. Next, we apply dynamic programming to the corresponding bi-gram models to the selected intermediate representations for generating the most likely texts to explain the motions.

As a result, Table III shows the top three generated sentences with the value of likelihood for each motion.

V. Discussions

From the result, we have confirmed that the sentences which properly describe human motions are generated. Furthermore, we see from the generated sentences in Table III that some sentences do not have the symbol of the end of a sentence, ‘|’. This is because the bi-gram model is built based on the combination of bi-grams of the words appeared in the collected sentences. Therefore, there is possibility that a generated sentence becomes a longer sentence than any collected sentences, besides we have introduced “null” labels in the bi-gram model. On the other hand, the longer a sentence is, the less likelihood of the sentence is. Therefore, under an assumption that it is likely that there is not any longer sentence than the collected sentences, we set the number of words in a generated sentence as the maximum number of words which the corrected sentences have in the collected sentences. Considering these things, we have decided the maximum length of a generated sentence is enough for the maximum length of a sentence collected by the subject experiment, if any “null” labels appear in a generated sentence.

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE III
THE GENERATED SENTENCES IN THE TOP THREE RANKING

Motion	Generated sentences	Likelihood
1	• A, person, picks, up, pink, ball, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14, null_15, |	5.68e-24
	• A, person, picks, up, ball, with, left-hand, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14, null_15	2.52e-24
	• A, person, picks, up, pink, ball, with, left-hand, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14	2.10e-24
2	• A, person, passes, ball, to, right-hand, . , null_7, null_8, null_9, null_10, null_11	6.29e-16
	• A, person, passes, red, ball, to, right-hand, . , null_7, null_8, null_9,null_11, null_10	3.08e-18
	• A, person, passes, ball, from, left, to, right-hand, . , null_7, null_8, null_9	2.05e-18
3	• A, person, puts, ball, in, box, . , null_8, null_9, null_10, |	4.90e-15
	• A, person, puts, ball, in, box, . , null_7, null_8, null_9, null_10	1.22e-15
	• A, person, puts, ball, to, another, box, . , null_7, null_8, null_9	2.16e-16

VI. CONCLUSION

We have proposed a framework for probabilistic text generation of human motions extracted from Kinect videos. The human motions extracted from Kinect videos are observed as time-series data; the data are applied by several dimension reduction methods; and then turned to be a proper form for machine learning. We applied dynamic programming to a bi-gram model, built based on the collected natural language sentences through a subject experiment, for the most likely combination of words to express an observed human motion. In addition, by introducing the "null" labels with a number in the bi-gram model, we could generate the most likely natural language sentence without limitation of the number of words in a generated sentence. Furthermore, unlike template-based text generation, our approach is to generate the most likely texts based on probabilistic model – which means that various natural language expressions can be generated as linguistic resouces, i.e., collected sentences, increase.

On the other hand, we have not yet introduced the knowledge of syntactics and the knowledge about objects in the world into our framework. So, as future work, we are going to introduce those kinds of knowledge and then improve our framework so that it can generate texts to more precisely explain obserbed phenomenon. We also like to achieve more flexible correspondence between the intermediate representations and bi-gram models, and tackle the problem of how to divide human behaviors into the motions to be described by a natural language sentence.

REFERENCES

[1] Ding, Duo and Metze, Florian and Rawat, Shourabh and Schulam, Peter F. and Burger, Susanne, Generating natural language summaries for multimedia, Proceedings of the Seventh International Natural Language Generation Conference, INLG '12, Utica, Illinois, pp.128–130 2012.

[2] Ding, Duo and Metze, Florian and Rawat, Shourabh and Schulam, Peter Franz and Burger, Susanne and Younessian, Ehsan and Bao, Lei and Christel, Michael G. and Hauptmann, Alexander, Beyond audio and video retrieval: towards multimedia summarization Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, No.2, pp.1–8 2012.

[3] Tan, Chun Chet and Jiang, Yu-Gang and Ngo, Chong-Wah, Towards textually describing complex video contents with audio-visual concept classifiers, Proceedings of the 19th ACM international conference on Multimedia, pp.655–658 2011.

[4] Ichiro Kobayashi, Mami Noumi, and Atsuko Hiyama, A Study on Verbalization of Human Behaviors in a Room FUZZ-IEEE 2010 Barcelona, Spain, 18-23 July 2010.

[5] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. *Video In Sentences Out*, Conference on Uncertainty in Artificial Intelligence (UAI), 2012.

[6] Mirella Lapata, Probabilistic Text Structuring: Experiments with Sentence Ordering, In Proc. of the Annual meeting of the Association for Computational Linguistics pp.545–552 2003.

[7] Anja Belz, Probabilistic Generation of Weather Forecast Texts, Proceedings of NAACL HLT 2007, pp.164-171 2007.

[8] Anja Belz and Eric Kow, System building cost vs. output quality in data-to-text generation, In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09, pp.16-24, Athens, Greece 2009.

[9] Wei Lu and Hwee Tou Ng and Wee Sun Lee, Natural language generation with tree conditional random fields, EMNLP, pp.400–409 2009.

[10] Percy Liang, Michael I. Jordan, Dan Klein Learning Semantic Correspondences with Less Supervision, ACL-IJCNLP 2009.

[11] Angeli, Gabor and Liang, Percy and Klein, Dan, A simple domain-independent probabilistic approach to generation, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 502–512,Cambridge, Massachusetts 2010.

[12] Konstas, Ioannis and Lapata, Mirella, Unsupervised concept-to-text generation with hypergraphs, Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada, pp.752–761 2012.

[13] Konstas, Ioannis and Lapata, Mirella, Concept-to-text generation via discriminative reranking, booktitle = Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, pp. 369–378, Jeju Island, Korea 2012.

[14] MicroSoft Kinect : http://www.microsoft.com/en-us/kinectforwindows/

[15] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms DMKD' 03 2003.

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)