Analysis of Choreographed Human Movements Using Depth Cameras: A Systematic Review

Danilo Ribeiro[1(✉)], João Bernardes[1], Norton Roman[1], Marcelo Antunes[2], Enrique Ortega[2], Antonio Sousa[3], Luciano Digiampietri[1], Luis Cura[4], Valdinei Silva[1], and Clodoaldo Lima[1]

[1] University of São Paulo, São Paulo, Brazil
danilo.luque@usp.br
[2] Central Kung Fu Academy, Campinas, Brazil
[3] São Paulo Faculty of Technology, São Paulo, Brazil
[4] Campo Limpo Paulista Faculty, Campo Limpo Paulista, Brazil

Abstract. The use of computer vision to analyze human movement has been growing considerably, facilitated by the increased availability of depth cameras This paper describes the results of a systematic review about the techniques used for movement tracking and recognition, focusing on metrics to compare choreographed movements using Microsoft Kinect as a sensor. Several techniques for data analysis and pattern recognition are explored for this task, particularly Dynamic Time Warping and Hidden Markov Models. Most papers we discuss used a single sensor instead of more complex setups and most took advantage of the Kinect SDK instead of alternatives. Rhythm is rarely considered in these systems due to the temporal alignment strategies used. While most systems that use the sensors for some form of interaction instead claim that this interaction is natural, very few actually perform any sort of usability or user experience analysis.

1 Introduction

With the increased availability of depth camera technology and movement analysis tools, the use of these tools, such as Microsoft's Kinect sensor and SDK that track body movements and allow their reproduction in videogames and other applications, has been more and more explored, including to aid in learning and practicing activities for which movement is essential and may even be seen as a mark of quality, personality and individuality [1]. During learning and training, it is important to have some measure of quality of performance that is as precise as possible as a form of feedback, particularly if it can detect and show where the mistakes happened and even suggest how to correct them.

The development of systems using depth sensors to aid in sports, dancing and martial arts has been gaining prominence and examples of this are the work of Chye, Connsynn and Nakajima [2], which uses Kinect to complement the training of martial arts beginners, and that of Hachaj, Ogiela and Piekarczyk [3], which uses a gesture description language to practice combat and Shorin-Ryu Karate

M. Kurosu (Ed.): HCI 2016, Part II, LNCS 9732, pp. 82–92, 2016.
DOI: 10.1007/978-3-319-39516-6_8

techniques with a reduced risk of trauma. Another application of this technology in this context is distance training for dance and martial arts practitioners using virtual avatars [4].

This paper presents a Systematic Review of the use of depth cameras for the analysis of choreographed human movements in the past years, discussing aspects such as the techniques used for analysis, equipment and setup, applications and interaction.

Kinect is a sensor developed by Microsoft, initially for its Xbox 360 videogame console and later for computers and is composed of two cameras, one RGB and a depth camera that uses an infrared projector which can measure from 0.8 to 3.5 m[5]. The widespread use of this sensor today happens mainly for two reasons, its relatively low cost and high availability and this review focuses on its use.

The paper is organized in a simple manner, as follows: Sect. 1 this introduction and the Sect. 2 describes in detail the methodology adopted for the systematic review. One problem that is often considered in the development of most of these systems is the temporal alignment of movements to facilitate the comparison of those performed by the user with those of another person or some known dataset. Two techniques used for this task are prominent in the literature, Dynamic Time Warping and Hidden Markov Models, both share similarities [6], will be discussed more frequently and, thus, are briefly introduced in the Sect. 3, along with a couple other techniques. The Sect. 4 presents and discusses the work's results and Sect. 5 one brings it to a conclusion.

2 Methodology

Systematic Review (SR) is a form of research in the literature performed in a standardized way, often performed to collect and classify the work done in a specific area or regarding a specific question and to show the state of the art in that area, providing a synthesis of the research regarding that question and its main results up to that point in time [7]. SR follows strict criteria so that its results are trustworthy, reproducible and validatable. Before the review is performed, several of its aspects must be decided and recorded, such as the research questions it must answer, control papers that it should find, databases to be searched, search strings, inclusion and exclusion criteria, what information will be extracted from each work and how it will be summarized. Below we summarize the most important of these aspects.

2.1 Research Questions

Every SR has, as a starting point, research questions that delimit the problem and act as an initial filter for the works found and that must be answered by the end of the process. The questions used in this work are:

1. What methods are used to analyze and compare choreographed human movements (mostly martial arts and dance, but not restricted to them) captured with depth cameras, particularly Microsoft's Kinect?

2. What are the techniques, if any, for temporal alignment of the movements and to analyze their rhythm?
3. What are the main applications of these systems?
4. Is the quality of interaction in these systems, if it exists, analyzed? How?

2.2 Sources and Search Strings

The papers for this review were searched in the databases of the Institute of Electric and Electronic Engineers (IEEE), the Association for Computing Machinery (ACM) and Springer, all of which bring together much of the most important work in this area and have a friendly user interface to facilitate the search process. In each of these databases, six customized searches were performed. Table 1 summarizes the strings used for these searches (which were adapted as needed for each particular engine) and the total number of papers found in each search, followed by the number of papers that were selected or rejected after the application of inclusion and exclusion criteria, the number of duplicated papers and the final number of papers that were used to extract information for this review.

Table 1. Search strings and number of papers found

Order	String	Paper				
		Total	S	R	D	E
1	(((Kinect) OR (depth camera)) AND (martial arts))	529	16	511	2	6
2	((Kinect) AND (dancing)) AND (martial arts))	18	11	5	2	9
3	((Kinect) AND (gesture description language))	3	1	2	0	1
4	“gesture description language”	6	2	3	1	0
5	“gesture recognition” AND “depth camera”	43	1	28	14	0
6	(((“rhythm”) OR (“choreography”)) and((“kinect”)OR(“depth camera”)))	309	0	2	307	0
7	((“choreography”)and((“kinect”)OR(“depth camera”)))	500	4	174	3221	4
	TOTAL	1402	34	726	648	20

S – Selected; R – Rejected; D – Duplicated; E – Extracted

Observing this table we verify that 1402 scientific papers were returned using these keywords and search strings but only 20 were finally extracted for the SR. It is interesting to notice that, due to the option for doing independent searches instead of a single search with a complex and long string, many papers, almost half of them (648) were duplicated.

2.3 Inclusion and Exclusion Criteria

Many of the works found in the initial search were excluded for, ultimately, being outside the narrow scope we selected for this review. This process of inclusion or exclusion happened through the following criteria, predetermined in the research protocol:

- Inclusion
  - Work that analyzes sequences of multiple gesture (movement) with metrics to qualify the movements and using depth cameras;
  - Work with metrics for temporal analysis of the movement sequences.
- Exclusion
  - Work analyzing independent gestures;
  - Work analyzing semi random (not predetermined or choreographed) movement patterns;
  - Work that does not use depth cameras.

2.4 Support System

A free software system called "State of the Art through Systematic Review" (Start) was used in this work to store and organize the papers found in this review. It is a rather interactive tool with features such as duplicate filtering and .bib support developed by LAPES (Research Laboratory in Software Engineering) at Federal University of So Carlos, in Brazil, and we would like to extend our thanks to its creators.

3 Brief Description of Techniques

In this section we explain in a very succinct form the main algorithms used in the works included in this review to analyze movements: Hidden Markov Models (HMM), Dynamic Time Warping (DTW), Spherical Self-Organizing Maps (SSOM) and Gesture Description Languages (GDL).

- HMM: stochastic model of temporal data series that represent the probability of the data occurring. The idea is that the process is unknown (hidden) but its results can be known. It is derived from Markovian chains and widely used in pattern recognition (including movement analysis), artificial intelligence and molecular biology [8,9].
- DTW: like HMMs, this algorithm is also based on temporal series, but it solves the problem of finding a common path between two series of different sizes but otherwise similar, without requiring initial or final points to be the same, creating a warping between the two paths and generally using euclidian distance [10].
- SSOM: clusterization techinque that creates a spherical mapping to indicate tridimensional positions, searching for the neighbour that better fits the movement and creating a link to it [11].
- GDL: used both for dynamic movements or static gestures, a script describes a movement or pose and, if recognized correctly by any other means, it is added to a heap, which may contain a chain of scripts or a single one [3,12].

4 Results and Discussion

In this section, we will begin by characterizing the set of papers analyzed in our review. Figure 1 illustrates the year of publication of the papers found in this review and shows that this body of work is quite recent, with most of it (60 %) from 2013. Because we focus on Kinect and how it is making this sort of research and application more easily available, this was expected, since it was released for the Xbox only in November 2010 and for Windows only in February 2012 (although even before the Kinect for Windows release there were several alternatives explored to work with the sensor on personal computers).

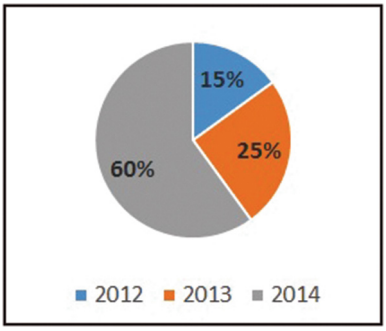

Fig. 1. Years of publication

From a geographic point of view, Fig. 2 illustrates which countries are publishing research in this particular area, showing that none of the countries is too far ahead, with each being responsible for 5 to 15 % of published papers.

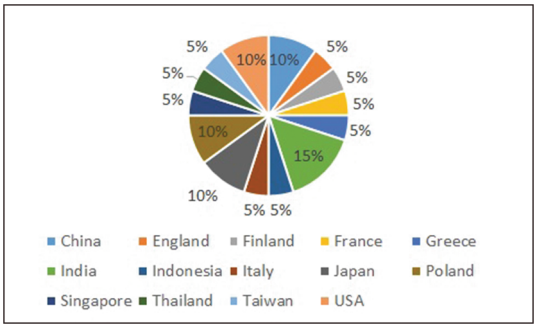

Fig. 2. Countries

If grouped by continent, however, as shown in Fig. 3, Asia pulls ahead significant (and the interest in both martial arts and computer vision in that continent is no surprise), followed respectively by Europe and America.

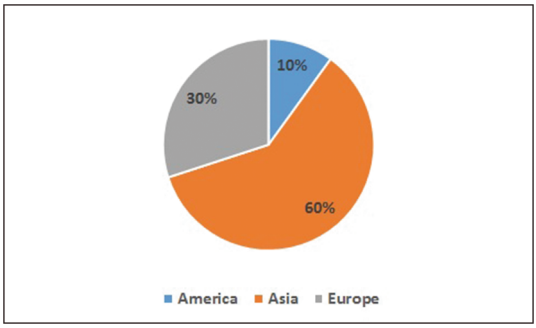

Fig. 3. Continents

Now we present the main results, according to the research questions listed previously. All studies made use of Kinect as a depth camera. Out of all of them, only two used more than one device in the experiment. Hachaj and Ogiela [13] used three sensors around a karate practitioner to aid in the learning process for martial arts techniques. It is interesting to note that the authors tested two distinct spatial configurations for the sensors. The first, less efficient, separated by an angle of $\pi/2$ and the second, more efficient, using an angle of $\pi/4$. Another work by the same group of Hachaj et al. [3] to verify the execution of karate moves compared the use of three sensors versus a single sensors, reporting an error of 13 % in movement capture with three cameras and 39 % with only one. For static poses or gestures, however, the difference between the two setups was not significant.

Different tools were used for image capture and skeleton fitting with the Kinect. Chye and Nakajima [2] use the OpenNI/NITE framework and draw a silhouette of the captured body to develop a game to aid karate practitioners in training. Other systems [14–16] also use this framework to analyze movements to give dancers a post-exercise evaluation [14], compare dance movements to the Bashir method [15] and score karate moves [16]. Microsoft's Kinect SDK was used in all other works (sometimes via a Unity wrapper), apparently being the most widely used alternative in this context. Table 2 briefly summarizes these papers.

In many cases the goal of the analysis was to compare movements between two performers (such as a novice and a master, or to measure the synchronicity of movement in a joint performance), or between one person and a pre-recorded video, using several distinct metrics. Other strategies involved recognizing specific and basic postures or gestures, for instance six basic ballet poses (as in the work of Sun et al. [11] using SSOM without much success to recognize poses or movements beyond those), or tracking user movements and mapping them to an avatar in a virtual world with virtual obstacles and such. Rhythm was often discarded in these metrics (possibly due to the temporal alignment strategies used), even in choreographed performances in which rhythm should indeed have some importance.

Table 2. Work with Kinect SDK

Author	Year	Goals
Alexiadis e Daras	2014	Compare performance of dance practicioner and expert and give feedback
Anbarsanti e Prihatmanto	2014	Analyze and score a Likok Pulo dancer
Gupta e Goel	2014	Aid in the practice of Kathak dance
Kaewplee et al	2014	Eliminate ghosting from captured images
Dancs et al	2013	Recognize ballet movements and compare with performance
Hachaj e Orgiela	2013	Identify karate moves
Holsti et al	2013	Analyze usability of a guidance system for trampoline jumps
Hachaj et al	2013	Increase movement tracking capability
Ho et al	2013	Extract and align music beat with best dance
Lin et al	2012	Synchronize dance videos from different sources
Pisharady e Saerbeck	2013	Recognize fast hand movements
Saha, Ghosh, Konar e Janarthanan	2013	Recognize Indian dance gestures
Saha, Ghosh, Konar e Nagar	2013	Recognize Indian dance gestures
Wada et al	2013	Analyze positions in a specific Kata
Keerthy	2012	Aid distance kung fu practice

Merely using euclidian distance between feature vectors of positions often did not yield very conclusive results but including velocity as a feature and still using euclidian distance showed better performance. Kaewplee et al. [17] use only Euclidian distance without temporal alignment (but using posterior movements to aid in the calculation of articulation angles) to analyze 24 basic Muay Thay movements. Chye e Nakajima [2] also use Euclidian distance and, like the previous work, also faced some difficulty to compare movements because of that, due to even slight temporal variations. Saha et al. [18] attempt to minimize the problem by defining an ideal speed for each movement and only comparing movements that did not deviate much from that speed. The same group used this approach again to recognize Indian dance moves [18]. Translating movements into a common description, such as using the Gesture Description Language [13] to create movement scripts and them comparing them showed good results, with 90 % accuracy in recognizing karate movements and comparing them to those executed by a black belt expert, using a setup with three sensors [3]. Lin et al. [19] developed an algorithm, using 103 videos from a database, that only showed significant synchronization errors when the dancer stepped outside Kinect's range.

More sophisticate algorithms for tracking and comparing temporal series were also explored, such as DTW, SSOM and HMMs. Using SSOM with articulation angles and captured body part lengths, Dancs et al. [20] mostly ignored rhythm while during training and obtained success rates of almost 90 % in leave-one-out and nearest neighbour validation and cross validation. Gupta and Goel [21]

use DTW with Euclidian distance and Earth Mover's Distance of finger positions to compare the performance of a subject and a master in Kathak. Zhu and Pun [14] used DTW to score dance practitioners comparing to the Taiji dataset and reached success rates above 80 %. Bianco and Tisato [16] also use DTW and report 96 % precision in recognizing and scoring the execution of karate movements, a similar value to that obtained by Pisharady and Saerbeck [22] using DTW to identify dynamic hand movements. Alexiadis and Daras [23] performed an experiment with and without the use of DTW, with showed a difference of over 20 % in favor of its use when comparing movements to the Huawei 3DLife/EMC Grand Challenge dataset. Keerthy [24] uses DTW in his Master's dissertation to create a Kung Fu training assistant that compares student and master movements. HMM was another technique widely explored. Anbarsanti and Prihatmanto [25] obtained promissing preliminary results in modeling the Likok Pulo dance using HMMs and classifying six individual basic dance movements and one undefined movement with almost 95 % accuracy. Masurelle et al. [15] also used HMMs to classify dance movements from a salsa database called 3DLife, comparing the results with the Bashir technique and obtaining 74 % positive matches. Figure 4 summarizes the frequency of use of these approaches to compare and classify human movements.

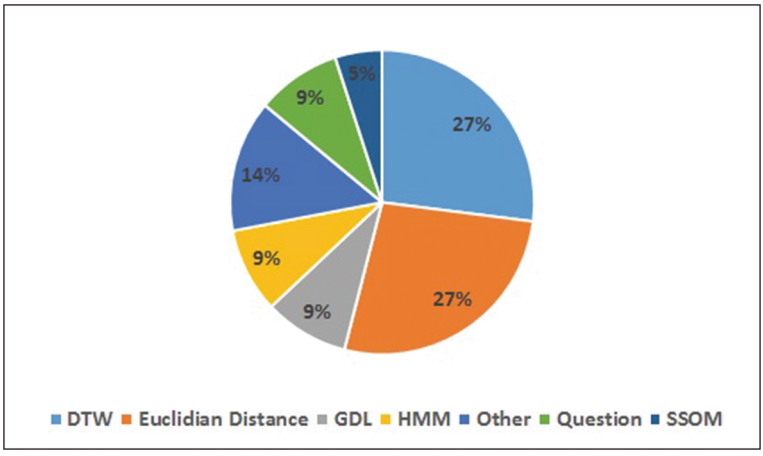

Fig. 4. Comparison approaches

Only two papers described some form of analysis of quality of interaction, Holsti et al. [26], in an application to aid in trampoline jumping, used questionnaires to evaluate their system's usability, with 90 % of users giving positive feedback despite complaining about the delay when showing the movement. Wada et al. [27] also analyzed the usability of their system to analyze kata positions with 89 % positive feedback.

5 Conclusion

The use of computer vision in several day to day applications is becoming more and more frequent, particularly with the popularization of smartphone cameras and depth cameras such as Microsoft Kinect's, which was one focus of this systematic review when applied to analyzing choreographed human movements. In this context, most papers we found take advantage of the Kinect SDK instead of alternatives, euclidian distance between feature vectors containing joint positions or angles was often used but showed poor results, often due to differences in temporal alignment of the movements being compared, but could be improved limiting the range of performance speed to be analyzed or adding speed to the feature vectors. Comparing standardized descriptions for gestures, movements and performances instead of the raw data from the sensors was another approach found. Out of the set of more sophisticate techniques to classify temporal series, DTW was the most commonly used in this context and showed good results, followed by the use of SSOM and HMMs. The quality of interaction with these systems was seldom analyzed in the papers included in this revision.

References

1. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)
2. Chye, C., Nakajima, T.: Game based approach to learn martial arts for beginners. In Embedded and Real-Time Computing Systems and Applications (RTCSA), 2012 IEEE 18th International Conference on, pp. 482–485, August 2012. doi:10.1109/RTCSA.2012.37
3. Hachaj, T., Ogiela, M.R., Piekarczyk, M.: Dependence of kinect sensors number and position on gestures recognition with gesture description language semantic classifier. In: 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 571–575, September 2013
4. Ogawa, T., Kambayashi, Y.: Physical instructional support system using virtual avatars. In: Proceedings of the 2012 International Conference on Advances in Computer-Human Interactions, pp. 262–265 (2012)
5. Ganganath, N., Leung, H.: Mobile robot localization using odometry and kinect sensor. In: 2012 IEEE International Conference on Emerging Signal Processing Applications (ESPA), pp. 91–94, January 2012. doi:10.1109/ESPA.2012.6152453
6. Fang, C.: From dynamic time warping (DTW) to hidden markov model (HMM). Univ. Cincinnati **3**, 19 (2009)
7. Rosana Ferreira Sampaio and Marisa Cotta Mancini: Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. Braz. J. Phys. Ther. (Impr.) **11**(1), 83–89 (2007)
8. Zoubin Ghahramani. Hidden markov models. chapter An Introduction to Hidden Markov Models and Bayesian Networks, pp. 9–42. World Scientific Publishing Co., Inc, River Edge (2002). ISBN: 981-02-4564-5, URL http://dl.acm.org/citation.cfm?id=505741.505743
9. Rabiner, L.R.: A tutorial on hidden markov models, selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989). doi:10.1109/5.18626. ISSN: 0018-9219

10. Li, S.Z., Jain, A. (eds.): Encyclopedia of Biometrics, chapter Dynamic Time Warping (DTW), pp. 231–231. Springer, Boston (2009). ISBN: 978-0-387-73003-5, doi:10.1007/978-0-387-73003-5768
11. Sun, G., Muneesawang, P., Kyan, M., Li, H., Zhong, L., Dong, N., Elder, B., Guan, L.: An advanced computational intelligence system for training of ballet dance in a cave virtual reality environment. In: 2014 IEEE International Symposium on Multimedia (ISM), pp. 159–166, December 2014. doi:10.1109/ISM.2014.55
12. Hachaj, T., Ogiela, M.R.: Semantic description and recognition of human body poses and movement sequences with gesture description language. In: Kim, T., Kang, J.-J., Grosky, W.I., Arslan, T., Pissinou, N. (eds.) MulGraB, BSBT and IUrC 2012. CCIS, vol. 353, pp. 1–8. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35521-91
13. Hachaj, T., Ogiela, M.R.: Qualitative evaluation of full body movements with gesture description language. In: 2014 2nd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS), pp. 176–181, November 2014. doi:10.1109/AIMS.2014.32
14. Zhu, H.-M., Pun, C.-M.: Human action recognition with skeletal information from depth camera. In: 2013 IEEE International Conference on Information and Automation (ICIA), pp. 1082–1085, August 2013. doi:10.1109/ICInfA.2013.6720456
15. Masurelle, A., Essid, S., Richard, G.: Multimodal classification of dance movements using body joint trajectories and step sounds. In: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4, July 2013. doi:10.1109/WIAMIS.2013.6616151
16. Bianco, S., Tisato, F.: Karate moves recognition from skeletal motion (2013). URL http://dx.doi.org/10.1117/12.2006229
17. Kaewplee, K., Khamsemanan, N., Nattee, C.: A rule-based approach for improving kinect skeletal tracking system with an application on standard muay thai maneuvers. In: 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS). 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), pp. 281–285, December 2014. doi:10.1109/SCIS-ISIS.2014.7044763
18. Saha, S. Ghosh, S., Konar, A., Nagar, A.K.: Gesture recognition from indian classical dance using kinect sensor. In: 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), pp. 3–8, June 2013. doi:10.1109/CICSYN.2013.11
19. Lin, X., Kitanovski, V., Zhang, Q., Izquierdo, E.: Enhanced multi-view dancing videos synchronisation. In: 2012 13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4, May 2012. doi:10.1109/WIAMIS.2012.6226773
20. Dancs, J., Sivalingam, R., Somasundaram, G., Morellas, V., Papanikolopoulos, N.: Recognition of ballet micro-movements for use in choreography. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1162–1167, November 2013. doi:10.1109/IROS.2013.6696497
21. Gupta, S., Goel, S.: Pogest: A vision based tool for facilitating kathak learning. In: 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 24–29, August 2014. doi:10.1109/IC3.2014.6897142
22. Pisharady, P.K., Saerbeck, M.: Robust gesture detection and recognition using dynamic time warping and multi-class probability estimates. In: 2013 IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing (CIMSIVP), pp. 30–36, April 2013. doi:10.1109/CIMSIVP.2013.6583844.

23. Alexiadis, D.S., Daras, P.: Quaternionic signal processing techniques for automatic evaluation of dance performances from MoCap data. IEEE Trans. Multimed. **16**(5), 1391–1406 (2014). doi:10.1109/TMM.2014.2317311. ISSN: 1520-9210
24. Keerthy, N.K.: Virtual kung fu sifu with kinect (2012)
25. Anbarsanti, N., Prihatmanto, A.S.: Dance modelling, learning and recognition system of aceh traditional dance based on hidden markov model. In: 2014 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 86–92, November 2014. doi:10.1109/ICITSI.2014.7048243
26. Holsti, L., Takala, T., Martikainen, A., Kajastila, R., Hämäläinen, P.: Body-controlled trampoline training games based on computer vision. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2013, pp. 1143–1148. ACM, New York (2013). doi:10.1145/2468356.2468560 ISBN 978-1-4503-1952-2 http://doi.acm.org/10.1145/2468356.2468560
27. Wada, S., Fukase, M., Nakanishi, Y., Tatsuta, L.: In search of a usability of kinect in the training of traditional japanese. In: 2013 Second International Conference on e-Learning and e-Technologies in Education (ICEEE) (2014)