Sign Language Recognition with Multi-modal Features

Junfu Pu, Wengang Zhou(✉), and Houqiang Li

CAS Key Laboratory of Technology in Geo-spatial
Information Processing and Application System,
Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei 230027, China
pjh@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn

Abstract. We study the problem of recognizing sign language automatically using the RGB videos and skeleton coordinates captured by Kinect, which is of great significance in communication between the deaf and the hearing societies. In this paper, we propose a sign language recognition (SLR) system with data of two channels, including the gesture videos of the sign words and joint trajectories. In our framework, we extract two modals of features to represent the hand shape videos and hand trajectories for recognition. The variation of gesture is obtained by 3D CNN and the activations of fully connected layers are used as the representations of these sign videos. For trajectories, we use the shape context to describe each joint, and combine them all within a feature matrix. After that, a convolutional neural network is applied to generate a robust representation of these trajectories. Furthermore, we fuse these features and train a SVM classifier for recognition. We conduct some experiments on large vocabulary sign language dataset with up to 500 words and the results demonstrate the effectiveness of our proposed method.

Keywords: Sign language recognition · Joint trajectory · Gesture recognition

1 Introduction

Sign language is wildly used in communication between the deaf and hearing societies. It has attracted considerable attention thanks to the broad social impact. Sign language recognition (SLR) targets on automatically translating sign language into text or interpreting it into spoken language. Besides, SLR has great potential applications in other fields such as human-computer interaction systems [19,23] and image retrieval [17].

Sign language conveys semantic information through gestures and the movements of hands and elbows. There are many previous studies focusing on the joint trajectories or gestures. The trajectory based sign language recognition method achieved promising results [16]. Lin et al. [16] presented a curve matching method

E. Chen et al. (Eds.): PCM 2016, Part II, LNCS 9917, pp. 252–261, 2016.
DOI: 10.1007/978-3-319-48896-7_25

from the view of manifold for sign language recognition. They divided the trajectory of the sign word into several linear segments, and calculated the distances between these two sets of segments. Their method achieved good performance both on two datasets which contain 370 and 1000 sign words, respectively. There are also other trajectory based methods for sign language or action recognition such as [1,4,5,20].

The above works are commonly based on trajectory recognition. However, these methods are not able to handle the recognition for signs with gestures. Murakami et al. [18] designed a gesture recognition system for Japanese sign language. They used a recurrent neural network that employed a three-layered back propagation algorithm to recognize the finger alphabet, with the gesture and character pairs as the input of the system. Furthermore, there were also some other works which combined different kinds of features, including gestures and hand trajectories [15,27,31]. Wang et al. [27] proposed a SLR framework using trajectories, RGB videos and depth videos. They used the method introduced in their paper to select key frames. After that, HMMs were used to model each sign word with the features extracted in each key frame. This method cost less time while maintaining a high recognition rate.

Some early SLR systems achieved great successes with data gloves [18,25]. The main advantage of data gloves is that they can capture the finger joints information and hand trajectory accurately. One typical SLR framework with data gloves was proposed by Gao et al. [8]. They employed the temporal clustering algorithm to cluster a large amount of transition movements for automatic segmentation. However, the data gloves were expensive and inconvenient in real application for the signers. Hence, more and more researchers turned to sign language recognition based on Kinect [9,27,30]. Microsoft Kinect [32] can provide the RGB and depth data as well as skeleton joint coordinates in real time, which makes a great contribution to SLR. The method based on Kinect proposed in [26] achieved an average accuracy of 74.4% on a large dataset with 1000 sign words.

Our approach is based on multi-modal features extracted from RBG data and joint coordinates. We use 3D CNN to obtain the representation of RGB video captured by Kinect. At the same time, LeNet [14] is used for joint trajectory based feature extraction. After that, we use SVM to recognize each sign word with these two kinds of features.

The rest of the paper is organized as follows. In Sect. 2, we describe the framework of our sign language recognition system and present the main procedures to represent the sign words with skeleton joint coordinates and RGB data. The SVM classifier is also briefly introduced in this section. The experimental results are reported in Sect. 3. Finally, the paper ends with some conclusions in Sect. 4.

2 Our Approach

In this section, we first give an overview of our SLR system. By using Kinect, we obtain 3D coordinates of 25 skeleton joints including hands, elbows, head,

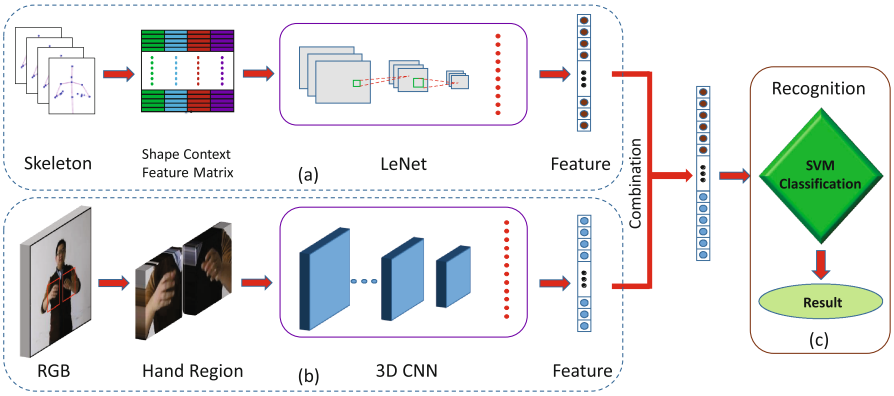


Fig. 1. The framework of our SLR system. (a) Feature extraction on skeleton trajectory from LeNet with shape context feature matrix as input. (b) Using 3D CNN to represent the hand videos formed by patches with hands from original RGB videos. (c) Classification with SVM.

and so on. Besides, the RGB videos is also captured. With the original data, we extract robust features for signs representation and the details will be introduced next. At the end of this section, we use SVM to recognize each sign word.

2.1 System Overview

Figure 1 gives the illustration for the main flowchart of the proposed method in this paper. In this method, we extract both skeleton based features and video based features for recognition. Figure 1(a) shows the procedure of feature extraction from skeletons. We first extract shape context for each point in motion trajectory, and integrate them to form a feature matrix. After that, we use these feature matrixes as the input of a convolutional neural network, and extract deep convnet features which are the intermediate response of fully connected layer from LeNet [14]. The RGB data based feature extraction process is illustrated in Fig. 1(b). Kinect [32] can help us track the skeletons of signer's hands. Hence, we can easily identify the corresponding regions of both hands and extract features from them for recognition. Thus, we get a video with lower resolution which only contains hand shape. With these videos, we can get a good representation by learning with 3D convolutional neural network [22]. These two kinds of features are concatenated for fusion. Finally, we use SVM [3] for classification of sign words.

2.2 Trajectory Representation

When we focus on a specific skeleton, we can get a trajectory formed with a set of 3D coordinates captured in each frame. Then the motion of a skeleton can be modeled by this trajectory. In this section, we will introduce how to build robust features with an effective representation of the 3D trajectory curve.

Preprocessing. In SLR, the trajectories are quite different when performed by different signers. Even for the trajectories performed by the same signer, they may be quite different in velocities. To address these problems, there are two aspects for data preprocessing, including normalization and re-sampling. The location of head and height of spine are used for normalization. In this way, the effect of various scales can be avoid. Another aspect in preprocessing is re-sampling, which will remove the noise and make the trajectory much more smooth. We use the $1 algorithm [29] for re-sampling. We use this method to sample the trajectory with a fixed number of points, making the distribution of the sampling points more uniform. In our experiments, the number of points for sampling is set to be 250.

Shape Context. Shape context [2] is a kind of feature which has been wildly used in shape recognition. It describes the distribution of other points in the neighborhood of a reference point. Denote C as a trajectory curve consisting of a set of 3D points. For a point p_i on curve C, each element of shape context feature is defined as a histogram of voting with the remaining $N-1$ points in the corresponding bins. The $log-polar^2$ coordinate system is used to make it more sensitive to nearby points. Besides, we project 3D trajectory along three orthogonal plans and obtain three 2D curves. We extract shape context feature in each coordination and concatenate them all together. There are 3 bins for log r and 12 bins for θ. Hence, we get a 108-D ($3 \times 12 \times 3$) shape context feature for each point p_i on C.

Feature Extraction from LeNet. In sign language recognition, we focus on 4 joints which are most informative to recognize a sign word, including both two hands and elbows. Hence, we get four trajectories while a signer performing a sign word. Using the method introduced in above section, the trajectory curve are sampled into a fixed points which we set as 250 in our experiments. For each point p_i on curve C, we extract a 108-D feature. Thus, we can combine all the features extracted in 4 trajectories with the method shown in Fig. 2(c). For each trajectory, we get a 250×108 feature matrix, and the 4 feature matrixes are concatenated row by row. In this way, we get a 250×432 matrix to represent a sign word.

As we know, convolutional neural network (CNN) performs very well in feature learning for a variety of tasks, such as image classification, object detection, and video tracking. Motivated by that fact, we use these feature matrix to train a CNN model. In our system, LeNet [11] is adopted for feature extraction process. For the last fully connected layer, we change it into 1000 neurons. We use the intermediate response of the last fully connected layer as a descriptor for a sign word.

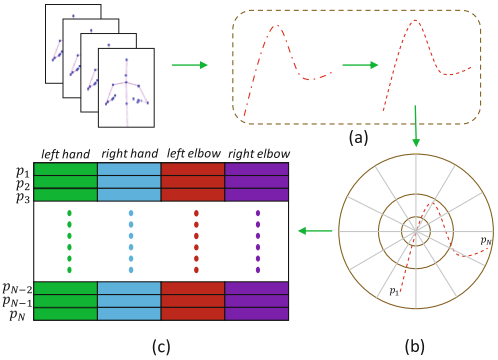


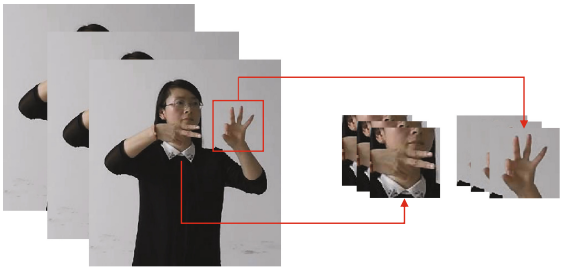
Fig. 2. Main steps for feature matrix. (a) Sampling with $1 algorithm to decrease the effect of different velocities. (b) Shape context extraction for each point. (c) The formation of feature matrix with different trajectories which include both hands and elbows.

2.3 RGB-Data Based Feature Extraction

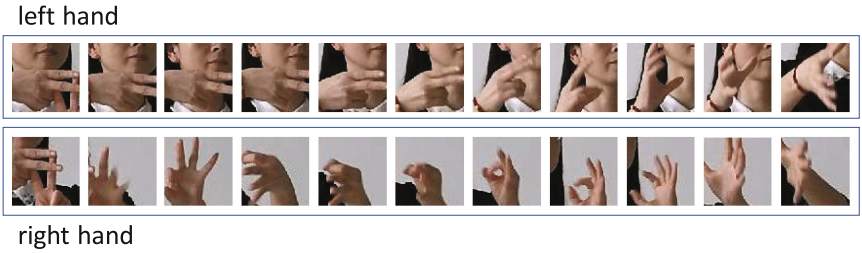
Hand Shape Segmentation. The Kinect provides us the body skeletons, including both hands. And we use the Kinect mapping function to get the corresponding location in RGB video. Hence, we can easily get an approximate region of hand. For each frame of a video, we take out a 70 by 70 image patch centered on the hand joint. We combine the two patches from both hands and get a low resolution video which only contains hands from the original video. The motion of other parts of body will be removed and we are able to focus on hand shape only. Figure 3(a) gives an illustration for fetching the patches with hand shapes, and several selected frames are shown in Fig. 3(b).

Representation with 3D CNN. To effectively extract the motion information in video analysis, [10] proposed to perform 3D convolutional layers of CNNs so that discriminative features along both spatial and temporal directions are captured. We use 3D CNN to analyze the motion of signer's hand. We follow the network architecture which is similar to Alex Net [13,22]. The main difference is that the 3D convolution kernels take the place of 2D kernels in Alex Net, and the number of kernels in each layer is also a little different. This network has 5 convolutional layers, 5 pooling layers, followed by 2 fully connected layers, and a softmax output layer. Figure 4 gives a simple illustration of network architecture.

For each video with hand shape, we split it into clips of 16 frames. SGD is adopted to train the networks with mini-batch size of 30 examples. Initial learning rate is 0.001, and divided by 5 every 20k iterations. The training stage lasts for about 310k iterations. The clips from each video are passed to the 3D convolutional neural network to extract fc6 or fc7 activations. These activations are averaged to form a 2048-dim descriptor. This representation for each video is used for sign words recognition.



(a)



(b)

Fig. 3. (a) A simple illustration for fetching the hand shapes. (b) Selected frames with left hand (first row) and corresponding right hands (second row).

Fig. 4. 3D CNN architecture. The network has 5 convolutional, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. The number of filter in each layer is denoted at the bottom of each box. Each fully connected layer has 2048 output units.

2.4 Recognition with SVM

Support vector machine (SVM) [24] is one of the most successful statistical pattern classifiers which has recently gained popularity within visual pattern recognition [21]. In this section we provide a brief review of SVM. Consider a training data set $T = \{(\mathbf{x_i}, y_i)|i = 1, 2, 3, ..., N\}$ with N samples, where $\mathbf{x_i} \in \mathbb{R}^D$ is the feature vector and $y_i \in \{-1, +1\}$ is class label. The basic task of SVM is to separate the samples in T into two classes. Assume that the training data set is linearly separable in feature space, so that we can find at least a hyperplane $\omega \cdot \mathbf{x} + b = 0$ separating samples into two classes. The support vector method

aims at constructing a classifier of the form:

$$y(\mathbf{x}) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i \Psi(\mathbf{x}, \mathbf{x_i}) + b), \tag{1}$$

where α_k are positive real constants and b is a real constant, and Ψ is a kernel function.

We use LIBSVM [6] to classify the sign words with the features extracted in Sects. 2.2 and 2.3. It implements the one-against-one approach [12] for multiclass classification. Denote k is the number of classes, then $k(k-1)/2$ classifiers are constructed and each one trains data from two classes and a voting strategy is used in classification.

3 Experiments

In this section, we provide some experimental illustrations and relative evaluations of our method on real dataset. We use the Chinese sign language dataset built by ourselves with Kinect. Firstly, the experiments with different features are conducted to evaluate the efficiency of these features. Then we compare the performance of our method with other SLR methods, including the SLR system proposed by Lin et al. [16] and the improved Dense Trajectories (iDTs) [28].

3.1 Dataset

The dataset in our experiments is collected by different sign language signers with Kinect. It contains 500 isolate Chinese sign language words in our daily life. There are 50 signers with different ages taking part in the data collection, which will make this dataset more various and challenging. For each sign word, it is performed by these 50 signers for 5 times. Hence, the dataset consists of 125k ($500 \times 50 \times 5$) samples. To evaluate the effectiveness of our method for signer independent, we choose 200 samples of each word performed by 40 signers for training, and the rest samples performed by the other 10 signers for testing.

3.2 Evaluation of Different Features

Different kinds of features have different discriminative abilities. We represent a sign word by hand trajectory and gestures. They both play important roles in SLR, since the hand trajectory describe the dynamic motion and gesture describe the static appearance. We conduct some experiments with gesture based features and trajectory based features, respectively. As introduced in Sect. 2.3, we split a video into several clips with every 16 frames, then calculate the average activations of fc6 (or fc7) from 3D CNN with all these clips as the representation of a sign word video within hands. Both fc6 and fc7 features are 2048D vectors. With the same idea for trajectory, the activations of the last fully connected layer of LeNet is used for recognition.

Table 1. Recognition rates with different features

Feature		Dimension	Accuracy
Video (Gesture)	clip-softmax	-	0.361
	fc6.ave-knn	2048	0.529
	fc7.ave-knn	2048	0.550
Skeleton (Trajectory)	fc-softmax	1000	0.773

Table 2. Performance comparison

Method		Top1	Top5	Top10
iDTs [28] (RGB)		0.685	0.888	0.927
CM_VoM [16] (Trajectory)		0.546	0.820	0.897
Ours	fc6-3DCNN+fc-LeNet	0.847	0.970	0.987
	fc7-3DCNN+fc-LeNet	0.858	0.973	0.988

The experimental results with different features are shown in Table 1. For sign words with videos, the accuracy in clip level is about 36.1%. The we classify the videos with the average fc6 (or fc7) activations from the network using KNN [7] classifier. The recognition rates increase to 52.9% for fc6 and 55.0% for fc7, which is higher than accuracy in clip level. The accuracy with trajectory features is 77.3%. For better performance, we combine the trajectory features and hand shape features for recognition.

3.3 Comparison with Other Methods

In this part, the baseline improved Dense Trajectories (iDTs) [28] and curve matching from the view of manifold for SLR (CM_VoM) [16] are evaluated on our dataset for comparisons. The recognition rates for different methods are shown in Table 2. The iDTs method is proposed for action recognition, and it also works for SLR task with a top1 accuracy of 68.5%. Another method CM_VoM is based on 3D trajectories, and it reaches the accuracy of 54.6%. In our method, we fuse the trajectory features and hand shape features, and use SVM for classification. The recognition rate is improved significantly with our method. The best recognition rates can reach up to 85.8%, 97.3% and 98.8% in top 1, top 5 and top 10, respectively. Among these results, our method achieves much higher accuracy and outperforms the baselines by a large margin on our SLR dataset.

4 Conclusions

This paper presents a framework for analyzing Chinese sign language using gestures and trajectories of skeletons. We implement the shape context for trajectory representation and extract features from LeNet with the feature matrix.

Meanwhile, we fetch out the hand region with a bounding box of 70×70 pixels in the original RGB videos, and then use 3D CNN to extract features for representing the sign words. Furthermore, these two kinds of features are fused for classification and demonstrate state-of-the-art results. We conduct some experiments with a large isolate Chinese sign language dataset. From the experimental results, it can be seen that the features extracted with neural networks demonstrate strong discriminative capability. Comparing to the baselines, the framework proposed in this paper achieves superior performance. In our future work, we will explore more robust feature fusion strategies among RGB data, depth data, and the skeleton coordinate to improve the performance of our sign language recognition system.

Acknowledgement. This work is supported in part to Prof. Houqiang Li by the 973 Program under Contract 2015CB351803 and the National Natural Science Foundation of China (NSFC) under Contract 61390514 and Contract 61325009, and in part to Dr. Wengang Zhou by NSFC under Contract 61472378, the Natural Science Foundation of Anhui Province under Contract 1508085MF109, and the Fundamental Research Funds for the Central Universities.

References

1. Amor, B.B., Su, J., Srivastava, A.: Action recognition using rate-invariant analysis of skeletal shape trajectories. TPAMI **38**(1), 1–13 (2016)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. TPAMI **24**(4), 509–522 (2002)
3. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**(2), 121–167 (1998)
4. Cai, X., Zhou, W., Li, H.: An effective representation for action recognition with human skeleton joints. In: SPIE/COS Photonics Asia, 92731R. International Society for Optics and Photonics (2014)
5. Cai, X., Zhou, W., Wu, L., Luo, J., Li, H.: Effective active skeleton representation for low latency human action recognition. TMM **18**(2), 141–154 (2016)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
7. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
8. Gao, W., Fang, G., Zhao, D., Chen, Y.: Transition movement models for large vocabulary continuous sign language recognition. In: FG, pp. 553–558 (2004)
9. Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using 3D convolutional neural networks. In: ICME (2015)
10. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. TPAMI **35**(1), 221–231 (2013)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678 (2014)
12. Knerr, S., Personnaz, L., Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Neurocomputing, pp. 41–50 (1990)

13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
15. Lee, G.C., Yeh, F.H., Hsiao, Y.H.: Kinect-based taiwanese sign-language recognition system. Multimedia Tools Appl. **75**(1), 261–279 (2016)
16. Lin, Y., Chai, X., Zhou, Y., Chen, X.: Curve matching from the view of manifold for sign language recognition. In: Jawahar, C.V., Shan, S. (eds.) ACCV 2014. LNCS, vol. 9010, pp. 233–246. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16634-6_18
17. Liu, Z., Li, H., Zhou, W., Hong, R., Tian, Q.: Uniting keypoints: local visual information fusion for large-scale image search. TMM **17**(4), 538–548 (2015)
18. Murakami, K., Taguchi, H.: Gesture recognition using recurrent neural networks. In: SIGCHI Conference on Human Factors in Computing Systems (1991)
19. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: a review. TPAMI **19**(7), 677–695 (1997)
20. Pu, J., Zhou, W., Zhang, J., Li, H.: Sign language recognition based on trajectory modeling with HMMs. In: Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X. (eds.) MMM 2016. LNCS, vol. 9516, pp. 686–697. Springer, Heidelberg (2016). doi:10.1007/978-3-319-27671-7_58
21. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR, vol. 3, pp. 32–36 (2004)
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV, pp. 4489–4497 (2015)
23. Ueoka, R., Hirose, M., Kuma, K., Sone, M., Kohiyama, K., Kawamura, T., Hiroto, K.: Wearable computer application for open air exhibition in expo 2005. In: PCM, pp. 8–15 (2001)
24. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
25. Wang, C., Gao, W., Xuan, Z.: A real-time large vocabulary continuous recognition system for Chinese sign language. In: Shum, H.-Y., Liao, M., Chang, S.-F. (eds.) PCM 2001. LNCS, vol. 2195, pp. 150–157. Springer, Heidelberg (2001). doi:10.1007/3-540-45453-5_20
26. Wang, H., Chai, X., Chen, X.: Sparse observation (so) alignment for sign language recognition. Neurocomputing **175**, 674–685 (2016)
27. Wang, H., Chai, X., Zhou, Y., Chen, X.: Fast sign language recognition benefited from low rank approximation. In: FG, vol. 1, pp. 1–6 (2015)
28. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR, pp. 3169–3176 (2011)
29. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Annual ACM Symposium on User Interface Software and Technology, pp. 159–168 (2007)
30. Zhang, J., Zhou, W., Li, H.: A threshold-based HMM-DTW approach for continuous sign language recognition. In: ICIMCS, p. 237 (2014)
31. Zhang, J., Zhou, W., Li, H.: Chinese sign language recognition with adaptive HMM. In: ICME (2016)
32. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)