Check for updates

Research Article

INTERNATIONAL JOURNAL OF ADVANCED ROBOTIC SYSTEMS

International Journal of Advanced Robotic Systems
January-February 2018: 1–8
© The Author(s) 2018
DOI: 10.1177/1729881417749482
journals.sagepub.com/home/arx
SAGE

A new action recognition method by distinguishing ambiguous postures

Zhiqiang Liu[1,2], Jianqin Yin[1], Jinping Li[2], Jun Wei[2] and Zhiquan Feng[2]

Abstract

One of the most important aspects of promoting the intelligence of home service robots is to reliably recognize human actions and accurately understand human behaviors and intentions. In the task of action recognition, there are many common ambiguous postures, which affect the recognition accuracy. To improve the reliability of the service provided by home service robots, this article presents a method of probabilistic soft-assignment recognition scheme based on Gaussian mixture models to recognize similar actions. First, we generate a representative posture dictionary based on the standard bag-of-words model; then, a Gaussian mixture model is introduced for the similar poses. Finally, combined with the Naive Bayesian principle, the method of weighted voting is used to recognize the action. The proposed scheme is verified by recognizing four types of daily actions, and the experimental results show its effectiveness.

Keywords

Robot, human action recognition, GMM, Kinect, bag of words

Date received: 30 March 2017; accepted: 20 October 2017

Topic: Special Issue – Multimodal Fusion for Robotics
Topic Editor: Marco Ceccarelli
Associate Editor: Huaping Liu

Introduction

Action recognition is a high-level computer vision task with wide practical applications in video surveillance, human–computer interaction, video retrieval, and so on.[1–3] It plays an important role in the process of promoting the intelligence of home service robots. As mentioned by Takano et al.,[4] it is a fundamental technology for robots. In the task of action recognition, there are often notably similar actions. During the execution of those actions, the postures of the person are notably similar, for example, drinking water and calling. In two or more different categories of similar actions, the proportion of identical or similar postures is notably high. Given an action test sample, for some common postures, there should be a type of fuzzy or uncertain method to discriminate them. Otherwise, it will produce relatively large errors and reduce the action recognition accuracy if a method of hard discrimination and classification is used. For ease of expression, we state that the "action" consists of several "postures" and mark the "action" and "posture" in Figure 1 to clearly understand their meaning.

In this article, based on the analysis of the semantic similarity, we study how to distinguish similar postures in different types of action. Specifically, we focus on and identify the action in which most of the postures are identical or similar. In other words, there are notably subtle

[1] Automation School, Beijing University of Posts and Telecommunications, Beijing, China
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, School of Information Science and Engineering, University of Jinan, Jinan, China

Corresponding author:
Jianqin Yin, Automation School, Beijing University of Posts and Telecommunications, Beijing, 100876, China.
Email: jqyin@bupt.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (http://www.creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

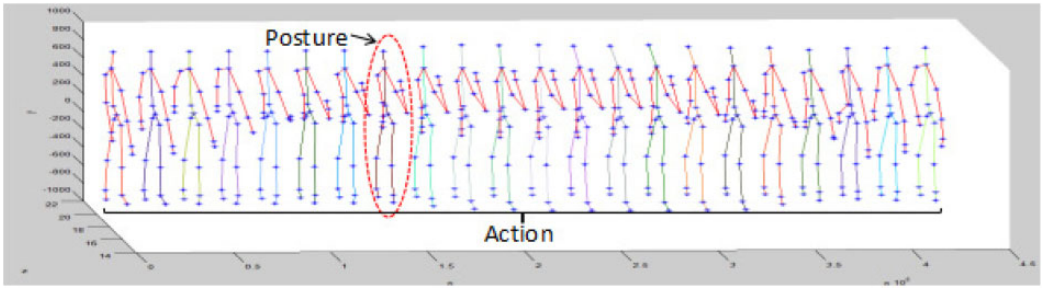

Figure 1. A class of action consisting of several postures.

differences between the few postures in different actions that we concern. Because of the similarity, in the process of discriminating similar postures, we must build a fuzzy model. Using this model, we should evaluate the scores of different categories of multiple actions according to the degree of similarity in the process of determining similar postures. On this point, there are many situations, such as slightly similar, generally similar, or notably similar. Van Gemert et al.[5] incorporated ambiguity in the codebook model by smoothly assigning continuous image features to discrete visual words and improve the classification performance. Inspired by this solution, we apply ambiguity modeling to recognizing similar actions by assigning ambiguous postures to multiple similar key poses and improve the action recognition accuracy.

Our contributions in this article can be summarized into two aspects. First, on the Kinect platform, the feature vectors extracted from 3-D skeleton data are used to represent the human body posture as indicated by Tian et al.[6] and generate some new code words using the bag-of-words model. Second, a new two-level codebook model is proposed, and the Gaussian mixture model (GMM) is introduced to express the ambiguity among the code words. All contributions are thoroughly experimentally verified on similar action data sets that we collected.

Related work

In recent years, the bag-of-words model has been prevalently applied in action recognition. On the one hand, this method is relatively easy to calculate and understand. On the other hand, this method shows notably good performance.[7] The traditional framework of the bag-of-words model was commonly used for 2-D images. The pipeline of the bag of visual words for video-based action recognition consists of five steps: (1) feature extraction; (2) feature preprocessing; (3) codebook generation; (4) feature encoding; and (5) pooling and normalization.[7] Among them, the most critical steps are feature extraction and encoding. Feature representation based on 2-D images usually includes low-level features such as interest point descriptors[7–12] and mid-level patch-based features.[13–22] Klaser et al.[11] proposed a spatiotemporal descriptor based on 3-D gradients. Wang et al.[12] sampled dense points from each frame and tracked them based on the displacement information from a dense optical flow field. These methods commonly ignore the spatial structure information of the feature points or descriptors and cause the loss of information. Taralova et al.[23] proposed a type of statistical characteristics that pool the low-level descriptor of supervoxels. Li et al.[24] used a probabilistic coding framework, which assigned the local spatiotemporal feature points to a small number of nearest visual words before extracting the patch-based features. Kovashka and Grauman[19] proposed a hierarchical bag-of-words model to express the configurations of spatial and temporal features at different scales. The main idea of these methods is to capture the structural information based on the quantized local spatiotemporal features. In some tasks of action recognition, it is often possible to determine the corresponding action category by distinguishing several key poses. The common idea is to characterize and classify each frame in an action sequence.[25–27] The contours or silhouettes of the human body are used to generate the feature vector to express the human posture. However, it is easy to be affected by factors such as illumination and occlusion.

For feature encoding, a voting-based method is more common. The usual practice is to generate the pose vocabulary or extract the key pose, match the key pose for a given test frame sequence, and count the votes of each category. An important assumption of these methods is that the pose code words are independent from one another, and following the hard-assignment method, these methods quantize the feature vector into a single pose code word. As suggested by Baysal et al.,[25] a type of learning algorithm is used to select the most representative posture of the intra-class, form the key pose vocabulary of multiple classes, and count the votes of every class. A constant weight is learned from the training data set for each key pose and score for each category.[27] These methods can satisfy the recognition accuracy requirement but lack stability because of the effect of the movement speed of execution and pause time. To avoid this problem, we should consider the many-to-one situation[28] and the structure information among different code words.[29] In recent years, the application of soft assignment is relatively

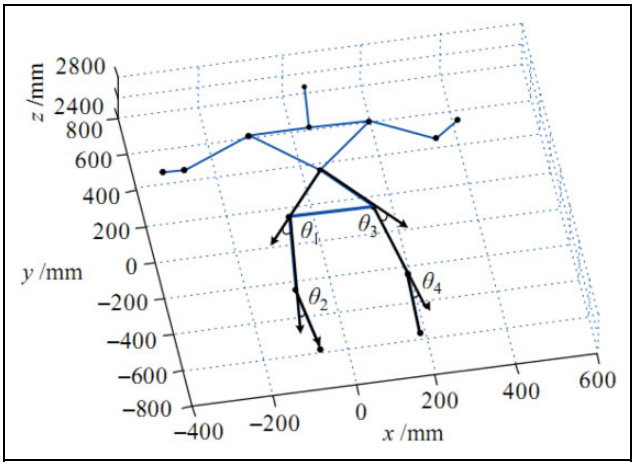

Figure 2. Angles of the lower limb.[6]

common.[30,31] Soft assignment accounts for the code word uncertainty and plausibility[7] and reduces the information loss during encoding. Liu et al.[32] considered the manifold structure of local features and used only k-nearest words to code a local feature and show that soft-assignment coding essentially estimated the posterior probability of a local feature to each visual word. Another popular encoding method is based on the super vector, which yields a notably high-dimensional representation by aggregating high-order statistics. Representative methods are Fisher vector[33–35] and vector of locally aggregated descriptors.[36–38]

Unlike these methods, in this article, we use a hierarchical soft-assignment method to identify similar actions, which enables the pose descriptor of an action sample to vote for several similar pose code words. In the process, a two-level bag-of-words model is applied, and the GMM is introduced at the second level to consider the similar-pose uncertainty; the posterior probability of each related key pose is used to score the corresponding class. Using this scheme, not only the shared information but also the different information can be captured by the hierarchical method.

Skeletal data-based pose representation

Feature representation plays an important role in human action recognition. As indicated in the previous work of Tian et al.,[6] based on the cognition of human movement characteristics, Kinect is used to obtain the 3-D coordinates of a human body joint point, the structure vector is constructed, and the correlation angle and modulus ratio are calculated to generate the feature vector to express a human body posture. The partial angle of the lower limb is marked in Figure 2.

Through many experiments, this characteristic expression has been verified to satisfy the requirements of invariance in translation, scaling, and rotation. Because it is a stable description of the human body posture, it is also used in this article.

Pose vocabulary generation

In the bag-of-words model, there are many methods to generate a vocabulary. Typically, the vocabulary is constructed by applying the k-means algorithm to cluster the sample features, and the cluster center represents a code word. In this article, the typical k-means method is used to cluster the feature vectors to describe human postures. The cluster center can be considered the representative of specific postures, which is called a code word. However, if we apply the k-means algorithm to cluster the feature vectors from different action categories, the cluster center may be a type of "average pose" generated by different categories of postures and lack of representative. In addition, it is notably difficult to determine the number of clustering centers. To obtain the key postures with the class label, we apply k-means clustering on all feature vectors belonging to the same action category. Simultaneously, the number of clustering centers is determined by decomposing the action execution process. Finally, the pose vocabulary is generated in the form as follows

$$A = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ p_{\tau 1} & p_{\tau 1} & \cdots & p_{\tau k} \end{bmatrix} \quad (1)$$

where τ is the number of action classes and k is the number of cluster centers.

Because of the similarity among different action categories, the code words generated by applying k-means may also be similar. Therefore, confusion easily occurs when a test feature vector matches the code words in the vocabulary. For example, a posture that originally belongs to the category of calling may be categorized as drinking. A large amount of error will occur if we use the approach of hard assignment, which quantizes one feature vector into a single code word. Hence, a type of ambiguity modeling must be applied to express the uncertainty.

Ambiguity modeling

Hard assignment is a type of one-to-one match. However, because of the similarity among the code words, the actual situation should be one-to-one or one-to-many, that is, a posture sample may correspond to multiple candidate code words. To depict the effect of the associated soft assignment, we introduce the GMM to express the ambiguity of the code words.

The diagram of our method is shown in Figure 3. More specifically, all code words are first constructed by applying k-means clustering to different categories of action postures. In this process, three elements are calculated: the representative of specific poses in the same categories of actions, that is, μ; the number of specific poses (denoted as m); and the corresponding covariance matrix Σ. Thus, a

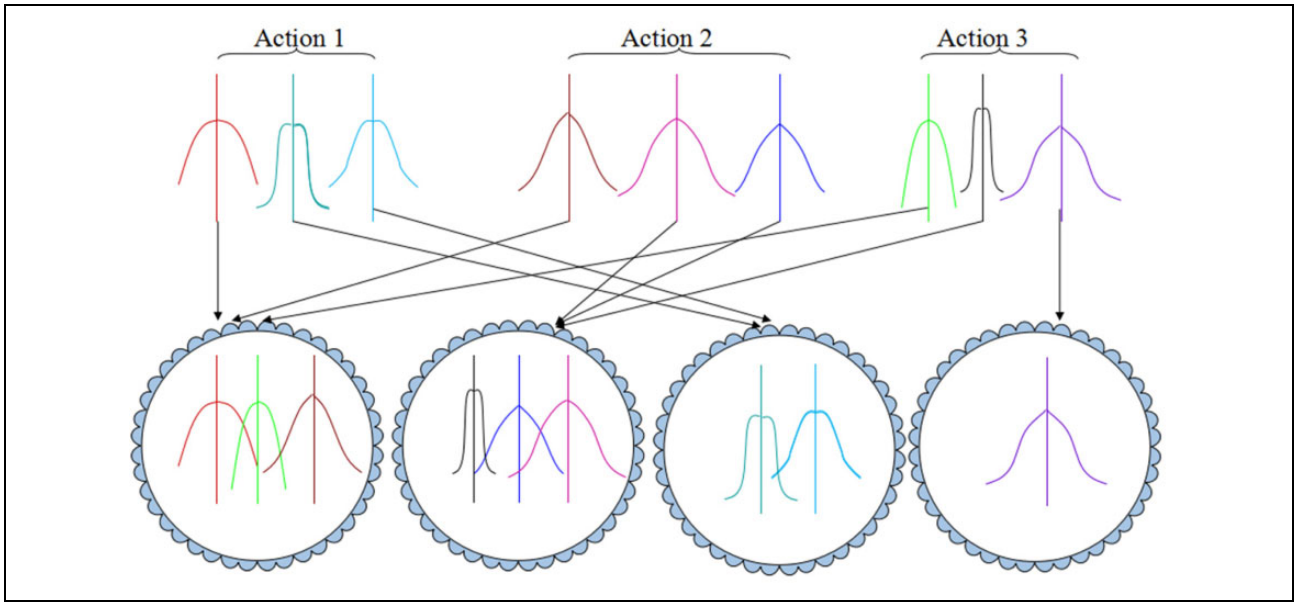

Figure 3. Process of generating a GMM. Each unit model in the top level represents the distribution of a key pose generated by *k*-means with the postures in the same action category, and each circle in the lower level represents a *k*-means-generated GMM with key poses. GMM: Gaussian mixture model.

single Gaussian model that corresponds to each code word is established as follows

$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left(\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{2}$$

In the next step, some ambiguous groups will be generated by applying *k*-means clustering to all code words. Probability $p_\mu(j)$ of each single Gaussian model j in the ambiguous group can be calculated as follows

$$p_\mu(j) = \frac{m_j}{\sum_{j=1}^{k^*} m_j} \tag{3}$$

where m_j is the number of specific poses that correspond to single Gaussian model j and k^* is the number of code words in an ambiguous group. The GMM has been generated. Each GMM corresponds to one ambiguous group.

Action recognizing

The soft assignment enables the feature vector to vote for multiple code words. In the process, the most important question is how to determine the voting value. For a given test action sequence $a = [x^{(1)}, x^{(2)}, ..., x^{(T)}]$, the probability that an element $x^{(t)}$ of the sequence that belongs to different code words in the ambiguous group is different. If code word j is one of multiple code words that correspond to sample $x^{(t)}$, the probability of j and posterior probability of $x^{(t)}$ can be obtained from the training data set. In formula (4), $f(\bullet)$ is obtained by formula (2)

$$\Pr\left(x^{(t)}|z^{(t)}=j\right) = f\left(x^{(t)};\mu_j,\sum\nolimits_j\right) \tag{4}$$

$$\Pr\left(z^{(t)}=j\right) = p_\mu(j) \tag{5}$$

where $x^{(t)}$, the posterior probability of j, is calculated by Bayesian formula

$$\Pr(z^{(t)}=j|x^{(t)}) = \frac{\Pr(x^{(t)}|z^{(t)}=j)\Pr(z^{(t)}=j)}{\sum_{j=1}^{k^*}\Pr(x^{(t)}|z^{(t)}=j)\Pr(z^{(t)}=j)} \tag{6}$$

For a given $x^{(t)}$, the denominator of formula (6) is identical, so the final voting value is

$$\text{score}^{\,j} = \Pr(x^{(t)}|z^{(t)}=j)\Pr(z^{(t)}=j) \tag{7}$$

If we use a hard assignment, the voting value is defined as

$$\text{score}^{\,j} = \begin{cases} 1, & if\ z^{(t)}=j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The diagram of our algorithm framework is illustrated in Figure 4. More specifically, for a given action sequence $a = [x^{(1)}, x^{(2)}, ..., x^{(T)}]$, code word p_{ij}^* of the nearest neighbor of sample $x^{(t)}$ is found

$$p_{ij}^* = \underset{\substack{i=1,2,\ldots,\tau;\\ j=1,2,\ldots,k}}{\arg\min}\{d_{ij}\}, d_{ij} = ||x^{(t)} - p_{ij}|| \tag{9}$$

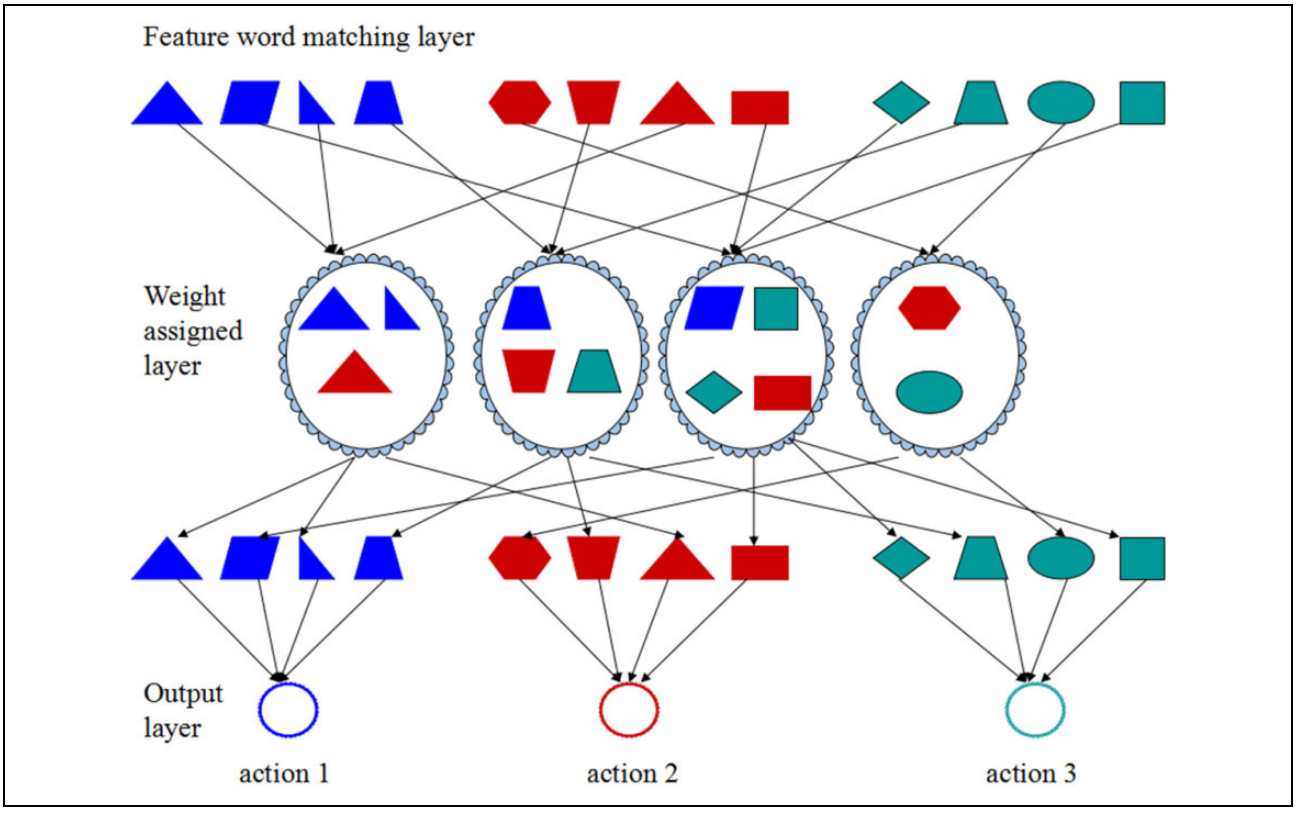

Figure 4. Framework of our action recognition algorithm, where different shapes represent different key poses, and the key poses with identical color belong to the same action category.

In the next step, the ambiguous group where code word p_{ij}^{*} is located is found. Simultaneously, all code words of this group are assigned a value according to formula (7). After all elements of the sequence have been completed, the score of each action category is computed

$$C_i = \sum_{j=1}^{n} \text{score}_i^j \tag{10}$$

where n is the number of code words belonging to the same category. Finally, the category with the highest score is determined as the label of the test sequence sample

$$C_i^{*} = \arg\max_{i=1,2,\ldots\tau} \{C_i\} \tag{11}$$

In the ambiguous group, the number of code words belonging to different categories varies, which causes an imbalanced situation among different classes. For a given action sequence, a sample posture will be assigned to one category many times and to the other categories only once. This will produce error and affect the action classification performance if too many sample postures are assigned to an ambiguous group. Therefore, we must introduce a type of adjustment weight to balance the assignment of different classes in the ambiguous group. Assuming that the number of code words belonging to class k in an ambiguous group is n_k, each code word of class k in this ambiguous group will be assigned a type of weight as follows

$$w(p_{k\bullet}) = \frac{1/n_k}{\sum_{k=1}^{k^{*}} 1/n_k} \tag{12}$$

Thus, a sample posture is assigned to different action categories in the ambiguous group with equal possibilities. To some extent, we avoid the error caused by the class imbalance. After the improvement, we adjust formula (10) as follows

$$C_i = \sum_{j=1}^{n} w(p_{ij}) \bullet \text{score}_i^j \tag{13}$$

Experiments

In this article, we focus on the recognition of similar actions with ambiguous postures. Four categories of actions were designed: calling, drinking water, using a remote controller, and pouring water. Fifty groups of data sets in each category of action were collected from different points of view, including frontal and side views. Through a cross-validation test, 40 groups of data sets in each category were randomly selected to compose the training data sets, which

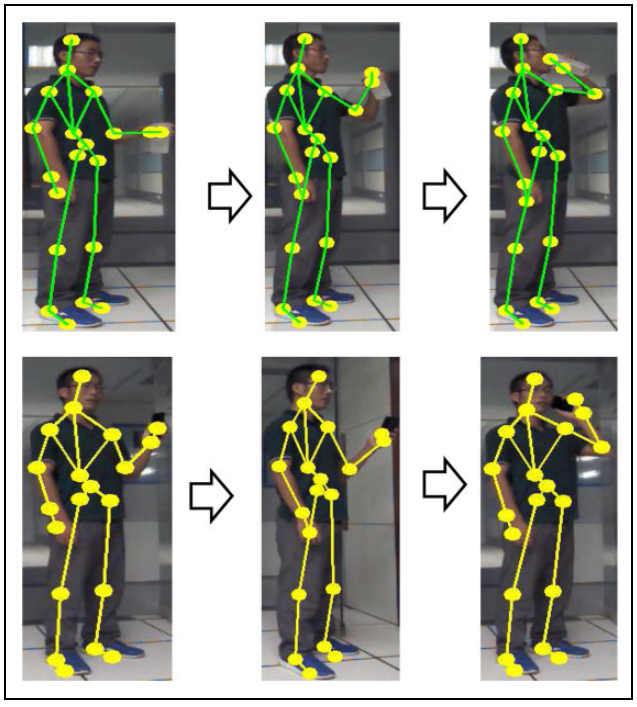

Figure 5. Similar postures of drinking and calling.

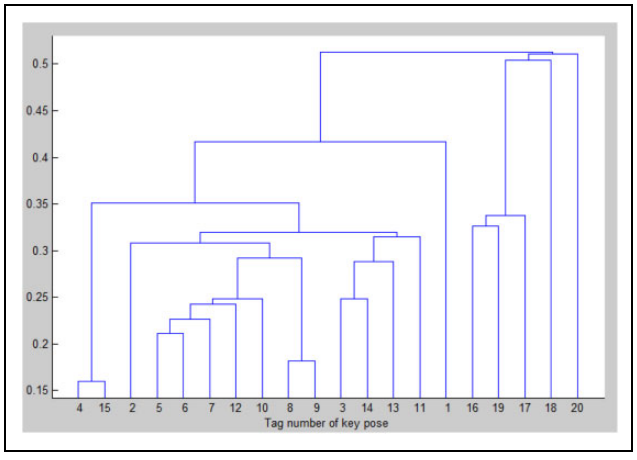

Figure 6. Hierarchical cluster tree of the pose code words.

amounted to 160 groups, and the remaining 10 groups of each category constituted 40 groups for the testing data set.

A part of the sequence of the two actions calling and drinking water is shown in Figure 5. Based on their skeleton shapes, we intuitively find that some postures in these two categories of action are notably similar.

The distribution of all code words in the feature space is visualized in Figure 6, which is constructed by applying a hierarchical clustering technique on all code words. The corresponding relationship between the tag number of key pose and the action category is shown in Table 1.

There is obvious difference between the category of pouring water and the other three categories of action. A separate cluster is generated by all code words of the category of pouring water, and a mixed cluster is formed by the remaining code words. The main reason of this situation is that there are many similar postures in these action sequences. Therefore, a fuzzy model must be established to perform the soft assignment.

Considering the randomness of the *k*-means algorithm, each experiment has been done 10 times, and the average recognition rate has been recorded. The confusion matrix of the results of different methods is shown in Tables 2 to 4. Comparing Tables 2 and 3, we observe that the classification performance in each category improved in different degrees from hard assignment to soft assignment. Comparing Tables 3 and 4, the overall recognition accuracy is also improved, which indicates that balance weights are important for the performance of the recognition algorithm. The statistical results of different methods are recorded in Table 5, which indicates that our improved method is effective and obtains good performance.

Table 1. Corresponding relationship between the tag number of the key pose and the action categories.

Tag number of key pose	Corresponding category of action
1–5	Calling
6–10	Using remote controller
11–15	Drinking water
16–20	Pouring water

Table 2. Confusion matrix for the hard assignment.

	Call	Remote	Drink	Pour
Call	0.6920	0.09	0.206	0.012
Remote	0.02	0.978	0.002	0.0
Drink	0.088	0.022	0.88	0.01
Pour	0.0	0.0	0.002	0.998

Table 3. Confusion matrix for the soft assignment.

	Call	Remote	Drink	Pour
Call	0.724	0.222	0.0540	0.0
Remote	0.016	0.984	0.0	0.0
Drink	0.032	0.002	0.964	0.002
Pour	0.0	0.0	0.0	1.0

Table 4. Confusion matrix for the balance weight.

	Call	Remote	Drink	Pour
Call	0.75	0.188	0.062	0.0
Remote	0.014	0.982	0.004	0.0
Drink	0.032	0.0	0.966	0.002
Pour	0.0	0.0	0.0	1.0

Table 5. Comparison of different improved methods.

Method	Accuracy
Hard assignment	88.7%
Soft assignment	91.8%
Balance weight	92.45%

Conclusions

In this article, we present a new solution for the similar action recognition. On the Kinect platform, feature vectors are extracted from 3-D skeleton data to depict human body postures. Based on the standard bag-of-words model, this article presents a type of two-level bag-of-words model and introduces GMM to model the correlation of different code words. Using this recognition scheme, we obtain good performance on the similar action data set that we collected.

The hierarchical model is notably important for the low latency, so we will use it to study the action recognition in low latency.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipts of the following financial supports for the research, authorship, and/or publication of this article: This work was supported partly by the National Natural Science Foundation of China (grant nos 61673192, 61472163, 61375084 and 61573219), the Natural Science Foundation of China Joint Fund with Guangdong Key Project (grant no. U1201258), and the Fund for Outstanding Youth of Shandong Provincial High School (grant no. ZR2016JL023).

References

1. Rahmani H, Mian A and Shah M. Learning a deep model for human action recognition from novel viewpoints. *IEEE Trans Pattern Analy Mach Intell* 2017; DOI:10.1109/TPAMI.2017.2691768.
2. Chen C, Liu K, and Kehtarnavaz N. Real-time human action recognition based on depth motion maps. *J Real Time Image Process* 2016; 12(1): 155–163.
3. Wang H, Oneata D, Verbeek J, et al. A robust and efficient video representation for action recognition. *Int J Comput Vis* 2016; 119(3): 219–238.
4. Takano W, Obara J, and Nakamura Y. Action recognition from only somatosensory information using spectral learning in a hidden Markov model. *Robot Auton Syst* 2016; 78: 29–35.
5. Van Gemert JC, Veenman CJ, Smeulders AWM, et al. Visual word ambiguity. *IEEE Trans Pattern Analy Mach Intell* 2010; 32(7): 1271–1283.
6. Tian GH, Yin JQ, Han X, et al. Novel human activity recognition method using joint points information. *Robot* 2014; 36(3): 285–292.
7. Peng X, Wang L, Wang X, et al. Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. *Comput Vis Image Underst* 2016; 150: 109–125.
8. Wang H, Ullah MM, Klaser A, et al. Evaluation of local spatio-temporal features for action recognition. *Br Mach Vis Conf* 2009; 124: 1–11.
9. Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies. In: *IEEE conference on computer vision and pattern recognition*, Anchorage, Alaska, USA, 24–26 June 2008, pp. 1–8. IEEE Computer Society.
10. Brox T and Malik J. Object segmentation by long term analysis of point trajectories. *European conference on computer vision*, Heraklion, Crete, Greece, 5–11 September 2010, pp. 282–295. Springer-Verlag.
11. Klaser A, Marszalek M, and Schmid C. A spatio-temporal descriptor based on 3d-gradients. *Br Mach Vis Conf* 2008; 275: 1–10.
12. Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories. In: *IEEE conference on computer vision and pattern recognition*, Colorado Springs, CO, USA, 2011, pp. 3169–3176. IEEE Computer Society.
13. Jain A, Gupta A, Rodriguez M, et al. Representing videos using mid-level discriminative patches. In: *IEEE conference on computer vision and pattern recognition*, 2013, pp. 2571–2578.
14. Carreira J, Caseiro R, Batista J, et al. Semantic segmentation with second-order pooling. In: *European conference on computer vision*, Florence, Italy, 7–13 October 2012, pp. 430–443. Springer-Verlag.
15. Wang LM, Qiao Y, and Tang X. Motionlets: mid-level 3d parts for human motion recognition. In: *IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 23–28 June 2013, pp. 2674–2681. IEEE Computer Society.
16. Nguyen MH, Torresani L, De La Torre F, et al. Weakly supervised discriminative localization and classification: a joint learning process. In: *IEEE international conference on computer vision*, Miami, Florida, USA, 20–25 June 2009, pp. 1925–1932. IEEE Computer Society.
17. Everts I, Van Gemert JC, and Gevers T. Evaluation of color stips for human action recognition. In: *IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 23–28 June 2013, pp. 2850–2857. IEEE Computer Society.
18. Ryoo MS and Aggarwal JK. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: *IEEE international conference on computer vision*, Kyoto, Japan, 27 September–4 October 2009, pp. 1593–1600. IEEE Computer Society.
19. Kovashka A and Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: *IEEE conference on computer vision and pattern recognition*, San Francisco, CA, USA, 13–18 June 2010, pp. 2046–2053. IEEE Computer Society.

20. Bilinski P and Bremond F. Contextual statistics of space-time ordered features for human action recognition. In: *IEEE international conference on advanced video and signal-based surveillance*, Beijing, China, 18–21 September 2012, pp. 228–233. IEEE Computer Society.
21. Liu J, Yang Y, Saleemi I, et al. Learning semantic features for action recognition via diffusion maps. *Comput Vis Image Underst* 2012; 116(3): 361–377.
22. Cheema S, Eweiwi A, Thurau C, et al. Action recognition by learning discriminative key poses. In: *IEEE international conference on computer vision workshops*, Barcelona, Spain, 6–13 November 2011, pp. 1302–1309. IEEE Computer Society.
23. Taralova EH, De la Torre F, and Hebert M. Motion words for videos. In: *European conference on computer vision*, Zurich, Switzerland, 6–12 September 2014, pp. 725–740. Springer.
24. Li Y, Ye J, Wang T, et al. Augmenting bag-of-words: a robust contextual representation of spatiotemporal interest points for action recognition. *Visual Comput* 2015; 31(10): 1383–1394.
25. Baysal S, Kurt MC, and Duygulu P. Recognizing human actions using key poses. In: *IEEE international conference on pattern recognition*, Istanbul, Turkey, 23–26 August 2010, pp. 1727–1730. IEEE Computer Society.
26. Thurau C.Behavior histograms for action recognition and human detection. In: *Conference on human motion: understanding, modeling, capture and animation*, Rio de Janeiro, Brazil, 14–20 October 2007, pp. 299–312. Springer-Verlag.
27. Weinland D and Boyer E. Action recognition using exemplar-based embedding. In: *IEEE conference on computer vision and pattern recognition*, Anchorage, Alaska, USA, 2008, pp. 1–7. IEEE Computer Society.
28. Liu H, Guo D, and Sun F. Object recognition using tactile measurements: kernel sparse coding methods. *IEEE Trans Instrum Meas* 2016; 65(3): 656–665.
29. Liu H, Sun F, Guo D, et al. Structured output-associated dictionary learning for haptic understanding. *IEEE Trans Syst Man Cybern Syst* 2017; 47(7): 1564–1574.
30. Alnajar F, Shan C, Gevers T, et al. Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions. *Image Vis Comput* 2012; 30(12): 946–953.
31. Weinshall D, Hanukaev D, and Levi G. LDA topic model with soft assignment of descriptors to words. In: *International conference on machine learning*, Atlanta, GA, USA, 16–21 June 2013, pp. 711–719. The International Machine Learning Society.
32. Liu L, Wang L, and Liu X. In defense of soft-assignment coding. In: *IEEE international conference on computer vision*, Barcelona, Spain, 6–13 November 2011, pp. 2486–2493. IEEE Computer Society.
33. Liu L, Shen C, Wang L, et al. Encoding high dimensional local features by sparse coding based Fisher vectors. In: *International conference on neural information processing systems*, Montreal, Quebec, Canada, 8–13 December 2014, pp. 1143–1151. Advances in Neural Infromation Processing Systems.
34. Kantorov V and Laptev I. Efficient feature extraction, encoding and classification for action recognition. In: *IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 2593–2600. IEEE Computer Society.
35. Evangelidis G, Singh G, and Horaud R. Skeletal quads: human action recognition using joint quadruples. In: *IEEE international conference on pattern recognition*, Stockholm, Sweden, 24–29 August 2014, pp. 4513–4518. IEEE Computer Society.
36. Yue-Hei Ng J, Yang F, and Davis LS. Exploiting local features from deep networks for image retrieval. In: *IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 June 2015, pp. 53–61. IEEE Computer Society.
37. Jain M, Jegou H, and Bouthemy P. Better exploiting motion for better action recognition. In: *IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 23–28 June 2013, pp. 2555–2562. IEEE Computer Society.
38. Kantorov V and Laptev I. Efficient feature extraction, encoding and classification for action recognition. In: *IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 2593–2600. IEEE Computer Society.