Temporal Hierarchical Dictionary with HMM for Fast Gesture Recognition

Haoyu Chen, Xin Liu and Guoying Zhao*
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
Email: chen.haoyu@oulu.fi, xin.liu@oulu.fi, guoying.zhao@oulu.fi

***Abstract*—In this paper, we propose a novel temporal hierarchical dictionary with hidden Markov model (HMM) for gesture recognition task. Dictionaries with spatio-temporal elements have been commonly used for gesture recognition. However, the existing spatio-temporal dictionary based methods need the whole pre-segmented gestures for inference, thus are hard to deal with non-stationary sequences. The proposed method combines HMM with Deep Belief Networks (DBN) to tackle both gesture segmentation and recognition by the inference at the frame level. Besides, we investigate the redundancy in dictionaries and introduce the relative entropy to measure the information richness of a dictionary. Furthermore, when inferring an element, a temporal hierarchy-flat dictionary will be searched entirely every time in which the temporal structure of gestures isn't utilized sufficiently. The proposed temporal hierarchical dictionary is organized in HMM states and can limit the search range to distinct states. Our framework includes three key novel properties: (1) a temporal hierarchical structure with HMM, which makes both the HMM transition and Viterbi decoding more efficient; (2) a relative entropy model to compress the dictionary with less redundancy; (3) an unsupervised hierarchical clustering algorithm to build a hierarchical dictionary automatically. Our method is evaluated on two gesture datasets and consistently achieves state-of-the-art performance. The results indicate that the dictionary redundancy has a significant impact on the performance which can be tackled by a temporal hierarchy and an entropy model.**

***Index Terms*—Hidden Markov Model; hierarchical structure; Deep Neural Network; Relative Entropy.**

I. Introduction

Human gestures are ubiquitous in visual cognition, pervading body language in all ages and cultures and tightly integrated with verbal communication [1]. Gesture recognition is actively used in applications spanning sign-language recognition, virtual manipulation to daily assistance [2]. Over the past several years, with the 3D skeletal joint coordinates of human-beings obtained from Kinects, renewed interests have been arisen in studying methods to recognize human body gestures [3]. However, there are still many challenges that deserve careful attention. First, the dimensionality of the input skeleton joint features is huge and the process is computation consuming. Secondly, the accurate segmentation of gestures from skeleton joint sequences is tough and often ignored with the assumption that pre-segmented sequences are available. At last, the variability of poses and movements in gestures is staggering, which needs a proper model for robust representation.

Many dictionary [4] or dictionary-like [5] based methods are utilized to represent the gestures all along, as it can offer clearer and stronger cues compared to the representation of entire gesture sequences. Those proposed dictionaries can be mainly divided into two categories: spatial dictionaries and spatio-temporal dictionaries. A spatial dictionary usually extracts its elements only with spatial features of gestures, while the element extraction of a spatio-temporal dictionary will consider both temporal and spatial information. For dynamic time sequence recognition, spatio-temporal dictionaries are proved to have better performance for temporal-counting attributes. A well-known spatial dictionary method is from [6]. Bourdev and Malik introduced *Poselets* (a configuration of body parts in 3D space) to gesture recognition. They set a large dictionary of frequently occurring poses and realize the recognition by comparing the poselet histogram. Ivan and Juan [7] extended this approach by a spatial hierarchical dictionary. They map k body parts to higher-level poselets with k-means, and several poselets will represent for the highest activity level. But these kinds of dictionaries don't contain temporal structures of gestures. For spatio-temporal dictionaries, the temporal information is well considered and the methods have good performances with classifiers such as support vector machine (SVM) and k-NN [8] [9] [10]. However, among most of these spatio-temporal dictionaries, there is little work on investigating the redundancy of dictionary. Besides, organization of the elements in those dictionaries are flat, which means the temporal structures of gestures are not embodied sufficiently and the search range is the entire dictionary. The method in [9] develops a temporal hierarchical dictionary by specifying the dictionary elements at each temporal segment. However, this method is limited with stationary settings: the whole pre-segmented gesture must be provided to obtain temporal elements and the total number of temporal segments is fixed. So it's hard to be adapted for recognizing a non-stationary sequence with several gestures, thus severely limiting its applications.

To handle the above disadvantages, we propose a novel temporal hierarchical dictionary with HMM and DBN. First, from the perspective of information encoding [11], we regard each element in the dictionary as a random independent system. The relative entropy is then utilized to evaluate and compress the elements. Then, we build the temporal hierarchy dictionary based on the HMM states. In this way, instead of traversing the entire dictionary with redundant elements,

*Corresponding author.

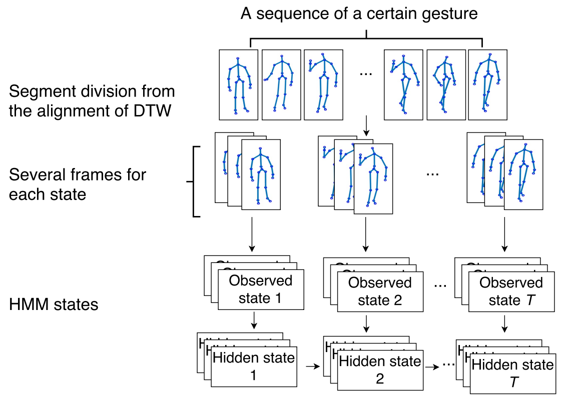

Fig. 1. The HMM model for processing a gesture. We use Dynamic Time Warping (DTW) to divide a given gesture into HMM states with T temporal segments. Note that the DTW processing is only done at training stage and the inference of testing sequence is conducted at frame level.

the search range is narrowed in the current state. For gesture recognition, we first use DBN to give an initial classification frame by frame. Due to the SoftMax output of neural networks, the elements in the dictionary can be inferred more accurately and robustly. The temporal hierarchical dictionary is then used to align elements with HMM and gives a final prediction. The major work of this paper is a temporal hierarchical dictionary with HMM including three contributions: a relative entropy model to describe and compress redundancy in the dictionary; a temporal hierarchical structure for the HMM transition; an unsupervised hierarchical clustering method to extract elements in the dictionary.

II. Methodology

Our proposed methodology is composed of four stages: HMM modeling for gesture motion, calculating relative entropies, building the temporal hierarchical dictionary with poselets and the DBN-HMM inference.

A. HMM modeling for gesture motion

We first define a related term *poselet* according to [10] as:

A configuration of body parts arranged in 3D space in an interval of time.

Since the 3D skeleton joint data captured by Kinect is superb, we can directly parameterize the configuration space of an articulated body by the 3D joint coordinates obtained from Kinect. Here an interval of time is set as one frame. Thus a poselet is the skeleton joint configuration of one frame.

Inspired by the HMM modeling techniques in [12], we model the gestures with HMM by first separating a gesture sequence into T segments. Then each segment will be assigned to a corresponding HMM state as shown in Figure 1. Since one segment may contain several frames and each frame carries one poselet, we regard all the poselets in this segment as the same in practice.

For the visible layer in HMM, we denote input features from a poselet as the observed variables x_t to an observed state X_t. One observed state X_t for $t = 1, \cdots, T$ will be connected to a corresponding hidden state variable as H_t in the hidden layer of HMM. The hidden variables are the types of the poselets and denoted as h_t. Based on the fundamental HMM, the full probability of HMM for training step is specified as:

$$\begin{aligned} p(x_1, x_2, ..., x_T, h_1, h_2, ..., h_T) = \\ p(h_1)p(x_1|h_1)\prod_{t=2}^{T} p(x_t|h_t)p(h_t|h_{t-1}) \end{aligned} \tag{1}$$

where $p(h_1)$ is the prior probability, $p(x_t|h_t)$ is the observation probability, namely, the emission probability and $p(h_t|h_{t-1})$ is the transition matrix. By pre-training DBN, the observation probability $p(x_t|h_t)$ can be obtained as:

$$p(x_t|h_t) = p(h_t|x_t)p(x_t)/p(h_t) \tag{2}$$

where $p(h_t|x_t)$ is the HMM state posterior probability estimated by the DBN model and $p(h_t)$ is the prior probability of each HMM state. For (x_t), as it does not vary with the gesture sequence and thus is always ignored.

B. Calculation of Relative Entropy

We introduce *entropy* to represent the information richness of poselets using a statistical probability model. As it is often used in data compression and encoding information, *entropy* can also reveal the ultimate compression of the poselets. Here we regard each poselet as a chaotic system [11], and the information it carries can be measured by its entropy. Obviously, one poselet by itself is isolated and unmeasurable from the perspective of statistical probability. Thus, to measure the entropy of a specific poselet, we introduce *relative entropy* to compare one poselet with all the rest poselets as [11]:

$$D(p||q) = \sum_{x} p(x) log \frac{p(x)}{q(x)} \tag{3}$$

where $D(p||q)$ is called relative entropy measured by bits, a measure of the distance between two probability mass functions p and q.

To apply relative entropy for gesture case, we use the statistical probability of the difference degree between poselets to represent the uncertainty of information or the value of entropy. Fixing one poselet as baseline chaotic system, all the rest poselets will be compared to it. Thus, at each temporal segment, if there are 20 gesture categories to be classified, then 20 different kinds of poselets from the 20 gestures will be compared. We assign each joint of the baseline poselet as a discrete random variable q^j and each joint of the comparison poselet as a variable p^j. Here j is the joint index and the total joint number is J. Then, we alter the original relative entropy for two chaotic poselet systems at temporal segment t as below:

$$D(P_t||Q_t) = log \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \frac{1}{1+e^{p_i^j - q_i^j + \delta}}}{N \times J} \tag{4}$$

where P_t and Q_t are the whole chaotic systems at temporal state t for $t = 1, \cdots, T$, i stands for the training sample index for $i = 1, \cdots, N$. We conduct probability statistics of

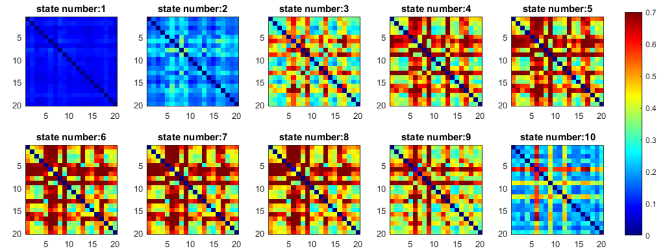

Fig. 2. The ten entropy maps of the Chalearn2014 database (corresponding to ten temporal states). Since Chalearn2014 database has 20 gesture categories, each entropy map has 20 columns and 20 rows. The cross point at column m and row n in the maps stands for the relative entropy of poselets from gesture m and gesture n.

two poselets based on a large number of training samples to obtain relative entropies. The threshold δ is used to adjust the bias of the sigmoid function when comparing two joints. For each state t, an entropy map is generated by comparing each poselet to the rest as shown in Figure 2. We can see that, at the beginning states, the overall entropies are much less than that of the following states and can be largely compressed. Taking the entropy information calculated above into account, we conclude the below criteria for constructing a temporal hierarchical dictionary in the next section:

1. Gesture configuration registered at hidden layers are mostly unique, meaning each gesture category are non-repetitive movements.
2. The uncertainty degree of one gesture at a temporal segment measured by entropy, reveals how much information it contains at that stage.
3. The information captured from the gesture will accumulate with time going, and the confidence of recognizing a gesture is based on the overall information captured.

C. Building a temporal hierarchical dictionary with poselets

In the light of the criteria above, we propose a new temporal hierarchical dictionary (THD), with which the candidate poselets can be compressed and organized by HMM time states. Following the ideas of original HMM structure, we separate a gesture motion into T temporal segments. Whereas in a temporal-flat dictionary, if there are 20 gesture categories, it will contain all 20 or even more kinds of poselets at every time states. But in our temporal hierarchical dictionary, the relative entropy is used to cluster those poselets into fewer representative poselets. In this way, not only the total number of poselets in the dictionary is reduced, but also the search ranges in each HMM state are narrowed.

For example, in THD, only two poselets might be used at the beginning state due to its low relative entropy. When it comes to the next stage, with the cues accumulating and entropy increasing, more poselets are assigned at that stage. The diagram of one temporal hierarchical dictionary structure is shown in Figure 3.

The structure has several advantages over the previous work: (1) it can largely reduce the number of HMM states in the transition phase; (2) instead of traversing all the poselets in the dictionary, it narrows down the search range at the current state; (3) it utilizes the relative entropy to capture the information change of a temporal sequence, which produces poselets with less redundancy.

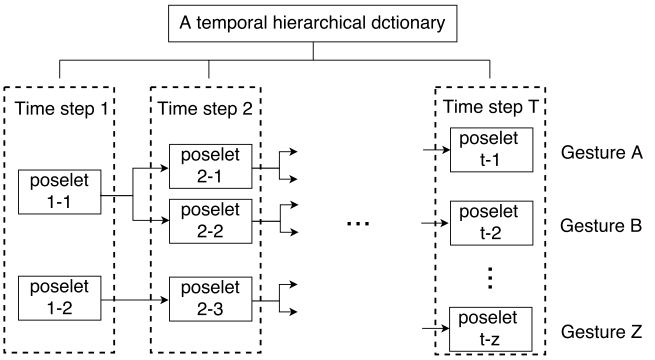

Fig. 3. One possible temporal hierarchical dictionary structure with T HMM states. The hierarchy of the structure is formed by HMM state. For each time state, all poselets in that state will be clustered based on the relative entropy.

However, another difficulty of setting a dictionary with temporal assembled poselets is labeling. When a gesture is divided into several temporal poselets, it is tricky to label the poselets in the training stage. For instance, in the time dimension, a gesture 'waving' can be assembled by poselets 'raise hand', 'waving' and 'put down hand', and another gesture 'hitting' can be assembled by poselets 'raise hand', 'hitting' and 'put down hand'. Thus in the beginning temporal segment, only one poselet 'raise hand' is needed. But how to disassemble a gesture into poselets and label those poselets with less redundancy are still problems.

With relative entropy, the problems mentioned above can be addressed with a hierarchical clustering algorithm proposed by us. It can build the temporal hierarchical dictionary and extract poselets automatically with gradually clustering. First, we choose one gesture from each category as Dynamic Time Warping (DTW) templates and interpolate them into fixed frame number (60 frames empirically). Then these templates are automatically divided into T even temporal segments. Then, based on these DTW templates, all the training gestures are divided into T segments with dynamic alignments. For each temporal segment, the poselets selected from all the gestures are regarded as candidate poselets. Then, a hierarchical clustering algorithm is used to cluster those candidate poselets based on relative entropies. The clustering details can be seen in Algorithm 1.

Here are several practical principles to implement the automatic clustering. First, the clustering is designed unidirectional and irreversible. Second, the maximum relative entropy always reaches at mid states instead of ending states, as shown in Figure 2. Thus the gesture clustering is always done at mid states and after that, the rest poselets are all kept and uncompressed.

Algorithm 1 Hierarchical clustering for THD

```
Input: $\sigma$: the threshold for relative entropies
  $T$: the total number of temporal states
  $N$: the total number of gesture categories
  $poseAll$: A dictionary of poselets with $T$ temporal states,
  containing $N$ poselets at each state
Output: $THD$: The temporal hierarchical dictionary
  for $t$ in $T$ do
    calculate $entropymap$ using Eq.3
    if $t < 2$ then
      $cluster \leftarrow poseAll(1)$
    else
      $cluster \leftarrow clusters(t-1)$
    end if
    for $c$ in $cluster$ do
      while $not(allposeletsgetclustered)$ do
        $baseline$ = $poselet$ with min relative entropy
        for $poselet$ in $c$ do
          if $entropymap(poselet, baseline) < \sigma$ then
            cluster $poselet$ to $baseline$ as $clusters(t)$
          end if
        end for
        record $clusters(t)$ in $THD$
      end while
    end for
  end for
```

D. DBN-HMM inference

In our case, a neural network is used to give a basic estimation of poselets at the frame level. This classification will be enhanced by HMM in inference phase.

DBN classification. Here we implement the Deep Belief Network with the Gaussian Restricted Boltzmann Machine (GRBM) to train the labeled poselets in THD. The architecture is shown in Figure 4. We introduce the GRBM for initialing weights for layers and generating a large number of features. The label in relation to the input is the poselet index. The final emission probability is given by:

$$p(x_i|h) = \mathcal{N}(x|\sum_j \sigma_i^2 w_{i,j} + b_i, \sigma_i^2) \tag{5}$$

where $\mathcal{N}(\mu, \sigma_i^2)$ is a Gaussian probability function with mean μ and variance σ_i^2. Here $w_{i,j}$ and b_i are the joint weights and biases of Restricted Boltzmann Machine. In the practical training, the variance σ_i^2 is set to be 1 in (5).

After the DBN-GRBM structure is trained with training samples, we can obtain a SoftMax probability of the poselets frame by frame according to $p(h_t|x_1, x_2, \cdots, x_t)$, where t stands for the current time state. To extract segment gestures from a skeleton stream, we attempt two methods. First one is to label those non-gesture frames as an extra poselet. Transitions are set between this extra poselet and other poselets [13]. The second method is to use a two-layer fully connected network to label each frame as 'motion' or 'no motion' [14]. The gesture extraction is to collect all 'motion' segments and the recognition is conducted on those segments. Both these two methods can achieve accuracies above 95% for segmentation.

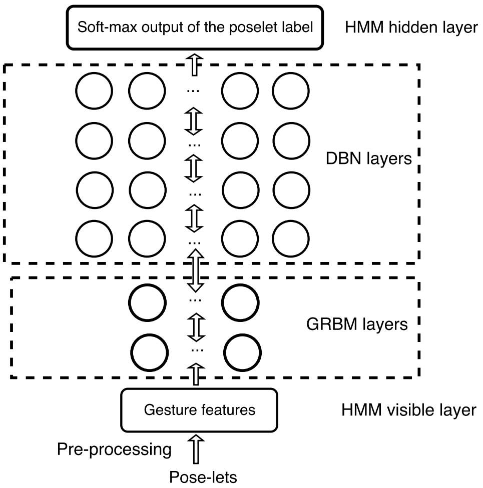

Fig. 4. The DBN-GRBM structure used for recognizing poselets.

Inference with HMM. Viterbi algorithm is a max-sum algorithm that will search the most probable path of HMM states among all the possible paths efficiently [15]. Initially, the decoded sequence $\hat{g}$ is determined as:

$$\hat{g} = arg\ \max_p\ p(x_1, x_2 \ldots x_T|g)p(g)/p(x_1, x_2 \ldots x_T) \tag{6}$$

where $p(g)$ is the prior probability of each gesture, and we are going to derivate $p(x_1, x_2, \ldots, x_T|g)$ to obtain the max value for this emission probability, then we get:

$$\begin{aligned} p(x_1, x_2, \ldots, x_T|g) &= \sum_h p(x_1, x_2 \ldots x_T, h|g)p(h|g) \\ &\cong max\pi(h_0)\prod_{t=2}^{T} p(h_t|h_{t-1})\prod_{t=1}^{T} p(x_t|h_t) \end{aligned} \tag{7}$$

With (7), we can break down the problem of solving best probability of main gesture class into solving HMM states probability with poselets $x_1, x_2, \ldots, x_T$. After determining the poselets alignment, the test sequence can be inferred.

III. EXPERIMENTS

In this section, details of experiments are illustrated.

Features. For the skeleton joint feature processing, we follow the work of [16] to extract skeleton joint properties as Eigenjoints. Eigenjoints are features that utilizing 3D position difference characters of joints to generate more spatio-temporal information. Before feature extraction, the coordinates are normalized by transforming them to a person-centric coordinate system. For time-consuming issue and gesture-orientated goal, we extract Eigenjoint features from 11 skeleton joints of upper body. The poselets are labeled according to the encoding paths in THD.

DBN-GRBM. We set our five-layer DBN-GRBM model architecture as $[N_X,\ N_h,\ N_{h1},\ N_{h2},\ N_{output}]$. Here N_X is

TABLE I
COMPARISON OF A THD-HMM AND A FLAT DICTIONARY WITH HMM ON CHALEARN 2014 DATASET.

Methods	HMM state number	Poselet number	Jacc.
Flat dictionary-HMM	5	100	0.787
	10	200	0.779
	15	300	0.716
THD-HMM	5	57	0.772
	10	88	0.813
	15	133	0.789
THD-HMM with DTW	10	81	**0.820**

the dimension of input feature observation layer, N_h is the number of hidden units in GRBM structure, the DBN inner hidden layers between N_h and N_{output} are N_{h1}, N_{h2}. N_{output} is the total poselet number. This feed-forward network are pretrained with a batch size of 100. The fine-tuning stage is set with a 0.1 learning rate and 500 epochs.

A. Datasets

We validate the proposed approach on two heterogeneous gesture datasets and compare it with two most common evaluation methods.

First is ChaLearn dataset [17] from a challenge of "multiple instances, user independent learning". There are 20 gesture categories in the dataset. It contains 470 labeled sequences for training, 230 sequences for validating and the rest 240 sequences for testing.

The second is MSRC-12 gesture dataset [18]. It includes 594 sequences and 719,359 frames from 30 people performing 12 gestures with 20 joints estimated by Kinect. We conduct our experiments following the same leave-subjects-out protocol as [18]. For each time, we leave the sequences of selected three subjects by order of subject number for testing and use the rest sequences for training. To train the network, we encode a window of 90 frames centered on the action point from the ground truth, with the beginning 20 frames of each sequence encoded as 'no-motion' poselets.

B. Results and discussion

We follow the standard validation protocols for each dataset to evaluate the proposed method.

For ChaLearn dataset, we strictly follow the evaluation protocol in [14] and report Jaccard index (Jacc.) as the measure. The Jaccard index measures similarity between samples, and is defined as the size of the intersection divided by the size of the union of the sample sets. First, we realize two THD-HMM structures with different HMM state numbers. Besides, as a sanity check of THD-HMM, we implemented flat-dictionaries with HMM. We compare our THD-HMM structures with the flat dictionary-HMM on ChaLearn dataset and the experimental results are shown in Table I.

TABLE II
THE VITERBI DECODING PERFORMANCES WITH 10 HMM STATES ON CHALEARN DATASET WITH A SINGLE GPU: NVIDIA TESLA K80 (RAM: 12 GB)

Viterbi decoding	Flat dictionary	THD-HMM
Decoding time/second	14-20s	2-3s
Decoding speed/frame per second	85-90/fps	400-600/fps

TABLE III
COMPARISON OF THE THD-HMM METHOD WITH THE STATE-OF-THE-ART METHODS ON CHALEARN DATASET.

Method	Results (Jacc.)
THD-HMM (our method)	**0.820**
ModDrop [14]	0.808
HOG, MRF [19]	0.792
HOG, Boosted classifier [20]	0.789
DNN-HMM [13]	0.787
Fisher Vector [21]	0.747

In the previous temporal flat dictionaries, the search ranges are the entire dictionaries with possible redundant elements. When taking the time consumption into account, this structural redundancy becomes a serious limitation. Instead, switching to our THD-HMM structure is helpful for the HMM-distributed temporal structure. For each temporal state of HMM, we filter redundant poselets with relative entropy. The results in Table I show the THD-HMM generally yields better performance and leads to a substantial decrease in the number of poselets. This is apparent in the last row of Table I: better performance (0.820) is obtained with two times fewer poselets (81 vs. 200). A large redundancy will cause a significant loss of accuracy which can be seen in the 15-state flat dictionary (0.716). In the case of five HMM states, the performance of THD-HMM is not better than that of the flat dictionary. The accuracy decrease of THD-HMM with five states might be caused by the information loss in data compression. At its root, THD-HMM is a kind of sparse coding which encodes the gesture with poselets along the time states. In this way, a good THD-HMM structure should make a balance between data loss and statistical redundancy.

Besides, we compare the Viterbi decoding speeds of these two kind of dictionary structures and the THD-HMM is proved to be more computationally efficient as shown in Table II. Note that the decoding time is measured with a whole test sequence in Chalearn dataset. One test sequence contains several gestures and has 1200-2400 frames. It can be seen that the THD-HMM method can considerably reduce the Viterbi decoding time compared to the flat one. Performances of other state-of-the-art techniques on ChaLearn are given in Table III. Our THD-HMM shows the best result using single skeletal modality.

In the MSRC-12 gesture dataset, F-score is always used for testing. It is the harmonic mean of *recall* and *precision* in a tolerated latency. On this dataset, the latency is set as $333ms$, the same as [18] and the same testing protocol and set as [22] are used. The results of means and standard deviations is given by leave-subjects-out with ten runs and compared with the state-of-the-art methods in Table IV. The proposed method

TABLE IV
COMPARISON OF THE THD-HMM METHOD WITH THE STATE-OF-THE-ART METHODS ON MSRC-12 GESTURE DATASET.

Methods	**Results (F-score)**
THD-HMM (our method)	**0.762±0.053**
DBN-ES-HMM [22]	0.7243
Structured Streaming Skeletons [23]	0.718±0.159
Randomized Forestg [18]	0.62±0.04

attributes to less redundant poselets with the best result.

C. Computational complexity

The training time of the DBN structures is around 12 hours for over 500,000 samples (poselets) in Chalearn dataset and 10 hours 400,000 samples (poselets) in MSRC dataset. The training platform is Theano with a single GPU: NVidia Tesla K80 (RAM: 12 GB). Theoretically, the single feed-forward neural network incurs linear computational time $O(T * |S|)$ for both our approach and a flat dictionary method, where T is the number of frames and S is the number of poselets. But the computational complexity of the Viterbi algorithm is $O(T * |S|)$ in our approach, which is apparently less than that of a flat dictionary method $O(T * |S|^2)$.

IV. CONCLUSION

In this paper, for 3D gesture recognition, we introduce the relative entropy to investigate the redundancy in gesture dictionaries and compress them efficiently. We propose a temporal hierarchical dictionary with HMM (THD-HMM), which has strong capability in narrowing down the search range in the poselet dictionary. An unsupervised hierarchical clustering algorithm is further proposed for the THD-HMM structure construction. The experimental results on two gesture datasets show the effectiveness of proposed method with state-of-the-art performances. Future research is to improve supervised poselet path encoding to unsupervised learning and explore more complementary representations from heterogeneous inputs such as RGB and audio data.

ACKNOWLEDGMENT

This work was supported by the Academy of Finland, Tekes Fidipro program (Grant No. 1849/31/2015) and Business Finland project (Grant No. 3116/31/2017), Infotech Oulu, and the National Natural Science Foundation of China (Grants No. 61601362 and 61772419). Haoyu Chen is supported by China Scholarship Council. As well, the authors wish to acknowledge Nokia visting professor grant, and CSC IT Center for Science, Finland, for computational resources. Special thanks to Jiawei and Henglin for their selfless help.

REFERENCES

[1] S. D. Kelly, S. M. Manning, and S. Rodak, "Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education," *Language and Linguistics Compass*, vol. 2, pp. 569–588, 2008.

[2] H. Cheng, L. Yang, and Z. Liu, "Survey on 3d hand gesture recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1659–1673, Sept 2016.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, June 2011, pp. 1297–1304.

[4] L. Zhang, S. Zhang, F. Jiang, Y. Qi, J. Zhang, Y. Guo, and H. Zhou, "Bomw: Bag of manifold words for one-shot learning gesture recognition from kinect," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[5] A. Richard, H. Kuehne, and J. Gall, "Weakly supervised action learning with rnn based fine-to-coarse modeling," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1273–1282, 2017.

[6] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 1365–1372.

[7] I. Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos," *Image Vision Comput.*, vol. 59, no. C, pp. 63–75, Mar. 2017.

[8] V. Bettadapura, G. Schindler, T. Ploetz, and I. Essa, "Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2619–2626.

[9] A. Alfaro, D. Mery, and A. Soto, *Human Action Recognition from Inter-temporal Dictionaries of Key-Sequences*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 419–430.

[10] K. Xu, X. Jiang, and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 567–576, March 2017.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[12] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 20–27.

[13] D. Wu, L. Pigou, P. J. Kindermans, N. D. H. Le, L. Shao, J. Dambre, and J. M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, Aug 2016.

[14] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, Aug 2016.

[15] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[16] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 14–19.

[17] S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-Lópes, H. J. Escalante, J. Shotton, and I. Guyon, *ChaLearn Looking at People Challenge 2014: Dataset and Results*. Cham: Springer International Publishing, 2015, pp. 459–473.

[18] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *CHI*, 2012.

[19] J. Y. Chang, *Nonparametric Gesture Labeling from Multi-modal Data*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015.

[20] C. Monnier, S. German, and A. Ost, *A Multi-scale Boosted Detector for Efficient and Robust Gesture Recognition*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015.

[21] G. D. Evangelidis, G. Singh, and R. Horaud, *Continuous Gesture Recognition from Articulated Poses*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015.

[22] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[23] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, "Online human gesture recognition from motion data streams," in *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2013, pp. 23–32.