Temporal-Order Preserving Dynamic Quantization for Human Action Recognition from Multimodal Sensor Streams

Jun Ye
University of Central Florida
jye@cs.ucf.edu

Kai Li
University of Central Florida
kaili@eecs.ucf.edu

Guo-Jun Qi
University of Central Florida
guojun.qi@ucf.edu

Kien A. Hua
University of Central Florida
kienhua@eecs.ucf.edu

ABSTRACT

Recent commodity depth cameras have been widely used in the applications of video games, business, surveillance and have dramatically changed the way of human-computer interaction. They provide rich multimodal information that can be used to interpret the human-centric environment. However, it is still of great challenge to model the temporal dynamics of the human actions and great potential can be exploited to further enhance the retrieval accuracy by adequately modeling the patterns of these actions. To address this challenge, we propose a temporal-order preserving dynamic quantization method to extract the most discriminative patterns of the action sequence. We further present a multimodal feature fusion method that can be derived in this dynamic quantization framework to exploit different discriminative capability of features from multiple modalities. Experiments based on three public human action datasets show that the proposed technique has achieved state-of-the-art performance.

Categories and Subject Descriptors

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding—*3D/stereo scene analysis*

General Terms

Algorithms

Keywords

Human action recognition, temporal modeling, temporal dynamic quantization, multimodal feature fusion

1. INTRODUCTION

Recent commodity depth sensors (e.g. Kinect) provide us high-quality data that was expensive to obtain in the past. With the access to the multimodal data including RGB, depth and skeleton stream, the field of human-computer interaction has been dramatically changed in the last few years. The natural user interface (NUI) in terms of gesture recognition and speech recognition has now been applied in the field of video games, education, business and health care. Among all these applications, human action recognition plays a key role and directly determines the quality of those products and services.

Although intensive research efforts have been made to this area, human action recognition is still a very challenging problem due to the complexity of the spatio-temporal patterns of the human actions. Other factors including the noise of the sensing data and the variations of size and execution rate of individual human subjects also make the problem challenging [20].

The temporal modeling is one of the most challenging problems in human action recognition. There are many visual words and histogram based methods [6, 16, 21, 19] which can effectively discriminate human actions composed of distinct postures. However, the lack of interpretation on the temporal layout of the actions makes these methods confused by actions of similar postures but in different temporal order, for example, "sitting down" and "standing up". Other methods such as graph model-based methods [7, 3] and motion template-based methods [8, 20] emphasize on the modeling of the temporal dynamics. But they frequently suffer from the temporal misalignment due to the temporal translation and execution rate variation. The temporal pyramid [14, 10] was proposed to interpret the temporal order of a video, but their uniform partition of the temporal sequence cannot handle the variation of execution rate either.

To address the issues of the temporal misalignment and the execution rate, we formulate the temporal modeling problem from a new perspective which aims at finding an optimal quantization of the temporal sequence. We also give a solution to the optimization problem by proposing the Temporal-order Preserving Dynamic Quantizing Algorithm. The general framework of the proposed approach is illustrated in Figure 1. The video sequence is dynamically partitioned by the temporal quantization and an aggregation procedure is performed to produce a quantized vector from each partition. This partition-aggregation process is

ACM ICMR '15 Shanghai, Chian

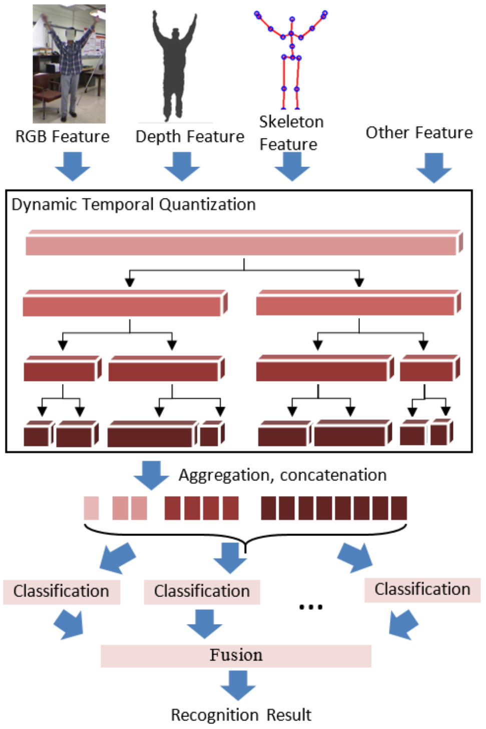

Figure 1: The general framework of the proposed approach.

recursively performed on the sequence of frames in a hierarchical way and finally produces a quantized representation of fixed size for the original video sequence. The above process is applied to multimodal features and generates their own feature-based quantization of the sequence. Supervised learning is then employed to learn the temporal model of the actions and predict the label of the new sequence upon the obtaining of the dynamic quantization vectors. The recognition rate can be further enhanced by leveraging the fusion of multimodal features in the same quantization frame.

The main contributions of the paper are:

1. We provide a quantization-based perspective for the problem of temporal modeling of human action sequences and propose a novel temporal-order preserving dynamic quantization method as the solution.

2. We present a multimodal feature fusion approach which exploits the discriminative capability of features from different data modalities; and

3. We demonstrate the performance of the proposed approach by evaluating it on three different human action datasets. Experimental results show the proposed approach has achieved state-of-the-art performance.

The remainder of the paper is organized as follows. We provide a brief review of the related work in Section 2. The Dynamic Temporal Quantization method is presented in Section 3. Multimodal feature fusion is discussed in Section 4. Experiments and performance evaluations are presented in Section 5. Finally, we conclude the paper in Section 6.

2. RELATED WORK

Many of the existing approaches for human action recognition focus on the spatio-temporal feature and local motions and do not explicitly model the temporal patterns of the action sequence. Most of these works are histogram-based. Space-time interest point (STIP) [4] and its extensions [9] were introduced to describe the local spatial-temporal features. Normally, a bag-of-word representation was then used to discriminate different types of human actions. Xia and Aggarwal [15] extended the 3D cuboid [2] and described the local depth cuboid by measuring the depth cuboid similarity. In [6], the bag-of-3D-points from the depth maps was sampled and clustered to model the dynamics of human actions. Similar ideas were proposed in [19], where the histogram of oriented gradient (HoG) was computed from the depth motion maps to classify human actions. Oreifej and Liu [10] proposed the 4D normals from the surface of the 3D point cloud and introduced the histogram of oriented 4D normals (HON4D) to achieve higher discriminative capability.

With the success of skeleton joints estimation from the depth images [11], joint-based features [12, 16, 21] are widely exploited in the human action recognition. Joint features were further quantized into code words and histogram of 3D joints (HOJ3D) [16] and histogram of visual words [21] were employed to describe the action sequences. The above methods all adopt the histogram-based representations of features. Although the distribution of the spatial-temporal feature has good discriminative capability in its feature space, it doesn't preserve the temporal layout of the individual primitive postures of the action. The missing of a holistic representation in the temporal dimension may lead to the poor performance in actions with similar postures but different temporal order. Different from the above histogram-based approaches, there are many methods focusing on the temporal order of the sequence and model the temporal dynamics in a holistic way. Motion template-based approaches [8, 20] introduced another way of modeling the temporal characteristics of human actions. In these methods, a number of motion templates indicating different action classes were trained. Dynamic Time Warping (DTW) was employed to warp the sequences of variant length and execution rate. The labels of the unknown action sequences were then determined by measuring the similarity between the unknown sequences and the motion templates.

Temporal Pyramid [14, 10] was developed to capture the temporal structure of the sequence by uniformly subdividing the sequence into partitions. Spatio-temporal feature descriptors were then applied to each partition. Since the uniform partition along the temporal axis is employed, the temporal pyramid is less flexible to handle execution rate variation. Adaptive temporal pyramid [18] was proposed to overcome the above disadvantage by subdividing the temporal sequence according to the motion energy. A Super Normal Vector is generated from the space-time partition and served as the comprehensive representation of the sequence. However, this method rely on sophisticated features such as 3D surface normals and polynormals which are inapplicable to more general problems.

Vemulapalli [13] proposed a body-part representation of the skeleton and modeled the geometric transformation between different body parts in the 3D space. The temporal dynamics in terms of the 3D transformation were captured and projected as a curved manifold in the Lie group. Classifications on the curves eventually determined the labels of the action.

Compared with the histogram-based methods, holistic temporal modeling methods achieved a comprehensive representation of the temporal dynamics and preserved the temporal order of the sequence.

3. TEMPORAL-ORDER PRESERVING DYNAMIC QUANTIZATION

The modeling of the temporal dynamics of action sequences is one of the most challenging problems in human action recognition. Many temporal modeling methods [8, 20, 14] suffer from the temporal misalignment problem due to the variations in the execution rate. A more sophisticated modeling method which considers the dynamics quantization of the sequence is desired.

3.1 Dynamic Temporal Quantization

In order to address the misalignment caused by the variations of the execution rate, the action sequences must be dynamically quantized. Ideally, the quantization needs to be in accordance with the transition between sub-actions of the sequences. Two requirements must be satisfied to achieve such quantization, 1) frames with close human postures are clustered together, 2) the temporal order of the sequence must be preserved.

The problem can then be formulated as follows. Denote $S = \{s_1, s_2, \cdots, s_n\}$ as an action sequence with n frames. Each frame s_i in the sequence is represented by a k-dimensional feature vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ik})$. The length n of each sequence can vary across different videos, so we wish to dynamically quantize each video into a new sequence $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m\}$ of fixed size m. Each element of feature vector $\mathbf{v}_j$ in this new sequence represents an unknown hidden stage of a category of human action, whose temporal order is preserved. Mathematically, we use $a_i \in \{1, 2, \cdots, m\}$ to denote a frame s_i in the original sequence S is assigned to the quantized vector $\mathbf{v}_{a_i}$ in V. Obviously, for any two consecutive frames s_i and s_{i+1} in original sequences, in order to preserve their temporal order in the quantized sequence, their assignments a_i and a_{i+1} should satisfy $a_i \le a_{i+1}$.

Then, a natural way to jointly optimize the assignment $\mathbf{a} = \{a_1, a_2, \cdots, a_n\}$ and the quantized sequence V can be obtained by jointly minimizing

$$\min_{\mathbf{a}, V} \sum_{i=1}^{n} \|\mathbf{v}_{a_i} - \mathbf{x}_i\|^2, \qquad (1)$$

$$s.t.\ \forall i \in [1, n-1], \quad a_i \le a_{i+1}, a_i \in [1, m]$$

where $\|\cdot\|$ can be any distance measurement. For convenience and efficiency consideration, we use Euclidean distance in the proposed algorithm.

It is nontrivial to jointly solve the optimal assignment $\mathbf{a}$ that satisfies the constraint of preserving the temporal order in each video, along with the optimal quantized sequence V. Our idea is to break down this optimization problem iteratively in a coordinate descent fashion.

Aggregation step: given the assignment $\mathbf{a}$ is fixed and Euclidean distance is adopted, it is not difficult to show that each element $\mathbf{v}_j$ of the optimal quantized sequence is the mean vector of all the elements $\mathbf{x}_i$ assigned to v_j. In other worlds, we have

$$\mathbf{v}_i = \frac{1}{|\{a_i = j\}|} \sum_{a_i = j} \mathbf{x}_i \qquad (2)$$

where $|\cdot|$ is the set cardinality and $\{a_i = j\}$ is the set of elements in $\mathbf{a}$ whose value is j.

Assignment step: when V is fixed, the assignment $\mathbf{a}$ can be updated to minimize the above distance to these quantized vectors in V subject to the temporal-order preserving constraint. We are inspired to develop a dynamic programming approach by dynamic time warping (DTW) to solve this subproblem. Specifically, the minimal distance $D_{l+1,k+1}$ given by the best assignment from $S_{1:l+1}$ to $V_{1,k+1}$ can be induced by the following iterative equation

$$D_{l+1,k+1} = \min\{D_{l+1,k}, D_{l,k}, D_{l,k+1}\} + \|\mathbf{x}_{l+1}, \mathbf{v}_{k+1}\|^2, \quad (3)$$

where $\|\mathbf{x}_{l+1}, \mathbf{v}_{k+1}\|$ is the distance between the current video frame $\mathbf{x}_{l+1}$ and the current quantized frame $\mathbf{v}_{k+1}$. Then by starting with $D_{1,k} = \min_k \|\mathbf{x}_1 - \mathbf{v}_k\|^2$ and $D_{l,1} = \min_l \|\mathbf{x}_l - \mathbf{v}_1\|^2$, the best assignment can be found iteratively according to the above equation. The resulting assignment has the same characteristics of DTW [8]. It always measures the closest frame when choosing the next warping step, and the generated assignment path is guaranteed to be non-decreasing in the temporal order.

The assignment $\mathbf{a}$ of video frames and the update of V proceed iteratively until convergence. The above process can produce an optimal dynamic quantization of video frames.

Algorithm 1 Iterated Dynamic Quantizing Algorithm

```
Input length of quantized sequence $m$, feature sequence $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$,
Output $\mathbf{a} = \{a_1, a_2, \cdots, a_n\}, V = \{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$
procedure IDQA($X, m$)
    Initialize assignment a by the even partition.
    repeat
        $V = Aggregation(X, V)$
        $\mathbf{a} = DynamicAssignment(X, V)$
    until Convergence
    return $\mathbf{a}, V$
end procedure
```

The pseudo code of the proposed Iterated Quantizing Algorithm is presented in Algorithm 1. In the initialization, S is evenly split into m partitions, each of which is assigned to an element $\mathbf{v}_j$ of V. An *Aggregation* subprocedure is exploited to update the quantized sequence V according to the frame assignment results. Then, a *DynamicAssignment* subprocedure is used to warp the original feature X to the quantized sequence V. We propose a modified DTW algorithm for the warping process. The assignment $\mathbf{a}$ can be determined from the warping path. The above two steps iterate until V converges or the maximum iteration times is reached. The output is the converged quantization sequence V and the final frame assignment to it. The above iteration algorithm modifies the initial quantization result step by step by the warping and aggregation and eventually generates a result that not only considers the similarity of human postures but also preserves frame order.

The **Aggregation** subprocedure summarizes the original feature of the partition into a quantized vector. This vector must provide a highly discriminative representation of the partition. Several aggregation methods can be applied such as mean, sum, max and min. Different aggregation methods have different advantages. As an example, the mean method (Eq(2)) is robust to noise and is theoretically optimal under Euclidean distance. However, some salient human postures may be mitigated by the averaging especially when the action sequence contains many neutral postures. The salient postures normally have higher discriminative capability than the neutral postures for the action recognition. Considering these factors, we also test the max-pooling as the aggregation method in our proposed algorithm. It also provides a form of temporal translation invariance by taking the maximum values of the corresponding elements from the feature of the partition.

The **DynamicAssignment** subprocedure contains two parts. The first part computes the matrix $D_{m,n}$ storing the minimal distance between any pair of subsequences of X and V starting from the beginning. This part can be solved using Eq(3). The second part computes the warping path based on the obtained optimal D. Considering that the original feature sequence (X) is warped to a much shorter quantized sequence (V), the step size of the warping path is restricted to $(1,0)$ and $(1,1)$, which is different from the classic DTW method that also allows $(0,1)$. In other words, the iteration Eq(3) changes to a more restricted form of $D_{l+1,k+1} = \min\{D_{l,k+1}, D_{l,k}\} + \|\mathbf{x}_{l+1}, \mathbf{v}_{k+1}\|^2$. This is to prevent the warping path from going too fast on the direction of the quantized sequence that the last few quantized vectors take most of the frames. It also avoids the scenario that one frame is assigned to multiple quantized vectors at the same time.

3.2 Hierarchical Representation

Inspired by the Spatial Pyramid [5] and the Temporal Pyramid [14], we further extend the Dynamic Temporal Quantization by incorporating the hierarchical structure. Figure 2 illustrates the hierarchical architecture of the proposed dynamic quantization. The original sequence is recursively partitioned by the Iterated Dynamic Quantizing Algorithm and forms a pyramid structure. The i_{th} level has $2^i - 1$ partitions. As a result of the dynamic quantization, the length of each partition varies. A quantized vector is then aggregated from each partition and eventually the final feature vector is generated by concatenating the quantization vector of all layers. Figure 2 shows a pyramid structure with 4 levels. The proposed technique may benefit from a structure with higher levels. Nevertheless, a large number of quantizations may over-segment the action sequence and compromise the generalization capability. We study the relationship between the height of the pyramid and the performance in the experiments.

The proposed dynamic quantization approach has several benefits. First, by exploiting the dynamic temporal quantizing of the sequence, it can address the challenge of execution rate variation by achieving a dynamic temporal quantization. Second, feature vectors with highly discriminative capability can be extracted by the aggregation procedure of the quantization. Third, the hierarchical description in terms of the multi-layer pyramid achieves a comprehensive representation of the temporal dynamics of the action by capturing both the global and local temporal patterns of the action.

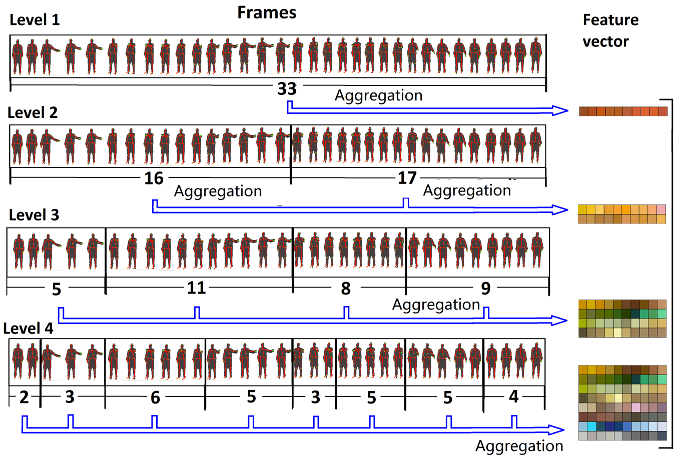

Figure 2: Illustration of the Dynamic Temporal Quantization and its hierarchical representation.

4. MULTIMODAL FEATURE FUSION

Multimodal sensors such as Kinect can provide data from different modalities including depth image, RGB image and skeleton map. In this approach, we extract features from all three data modalities and adopt supervised learning for the classification. The final recognition result is computed by the fusion of the classification results of each individual feature.

4.1 View-invariant Features

To leverage the discriminative capability from different data modalities, we use the following features in the proposed approach.

1. **Position:** 3D coordinates of the 20 joints of the skeleton.

2. **Angle:** normalized pairwise-angle feature.

3. **Offset:** offset of the 3D joint positions between the current and the previous frame [21].

4. **Velocity:** histogram of the velocity components of the point cloud.

The position feature is computed from the skeleton map by concatenating the x, y and z coordinates of all joints. It holds the raw information of the skeleton map and no special feature extraction is needed. In the angle feature, body segments are first computed by connecting each pair of adjacent joints. The normalized cosine value of each pair of the body segments are then computed and concatenated as the angle feature.

To better leverage the multimodal data, we propose the velocity feature of the point cloud. Generally speaking, the velocity feature describes the distribution of the movement of the point cloud at the neighborhood of each joint in the 3D space. Specifically, the 3D point cloud can be constructed from the depth image. In order to track the movement of the point cloud, optical flow is first performed on the RGB image to compute the displacement of pixels between two consecutive frames. By leveraging the pixel-to-pixel correspondence between the depth image and RGB image, the displacement of each pixel of the point cloud can be obtained and is used to compute the 3D velocity. To maintain the spatial information, the point cloud is split into clusters according to the joints. A square window with the joint at the center is applied to sample the region in the neighborhood of each joint. The velocity of each 3D point is further decomposed into x, y and z components which are further quantized into three bins indicating the positive, zero and negative values, respectively. Finally, a 9-bin histogram is generated for each joint and a 180D velocity feature is produced by concatenating all the 20 joints. Different from the above features levering only the joint information, the velocity feature is extracted from the cloud point which gives more details about the micro dynamics of the human actions.

Among the above four features, the angle feature are view-independent by nature. To handle the viewpoint variation for the other features, the 3D point cloud are transformed from the world coordinate system to a human-centric coordinate system by placing the "hip-center" at the origin and aligning the "spine" with the y-axis. The PCA (Principle Component Analysis) is adopt to the above features for the dimension reduction and noise reduction purposes. The above four features cover all three modalities of the sensing data from the device and provide a rich representation of the human actions.

4.2 Frame Assignment Sharing

In the previous section, the Dynamic Temporal Quantizing algorithm is computed based on the original feature of the action sequence. All of the above four features can be used to compute their own temporal quantization of the video frames. Should the frame assignment of the quantization be computed by one feature and shared by all other features or should it be done independently? Intuitively, the Independent quantization strategy may produce the best classification result with respect to individual feature that has been used. Nevertheless, it may encounter the overfitting problem and compromise the generalization capability when multimodal features are fused. Therefore, we compute the frame assignment of the quantization based on the feature with the highest discriminative capability. This assignment is applied to all the other three features to compute their own quantization results. We performed experimental study to justify the benefit of the sharing strategy and will discuss the result in the experiment section.

4.3 Classification and Fusion

We employ the supervised learning to learn the temporal dynamics of different actions from the quantized sequence discussed in the last section. A multi-class model is learned from the training sample with the known labels for each feature. The trained model is then used to predict the type of the unknown actions. The final recognition result is the fusion of the predictions of all individual features.

To be more specific, the SVM (Support Vector Machine) [1] is employed as the classifier in our method. A multi-class SVM is trained for each feature independently. In the predicting phase, each SVM estimates the label of the action by giving a probability that the current action belongs to each action type. The label is determined according to the maximum probability. The final classification result is then fused from the estimation of the multi-class SVM of each feature.

5. EXPERIMENTS

We evaluate the performance of the proposed method on three public datasets: UTKinect-Action [17], MSR-Action3D [6] and MSR-ActionPairs [10]. We choose these datasets because they provide data from at least two modalities and satisfy our multimodal feature fusion framework. In our experiment, the levels of hierarchical quantitating structure is set to 4 and $2^4-1=15$ partitions are generated to represent the entire action sequence. The PCA dimension reduction is set to preserve 99.5% of the energy of the raw feature. The window size for the velocity feature is set to 51×51. These parameters are determined according to the performances in the experiments. We use the LibSVM [1] with the RBF kernel as the classifier in our experiments. The fusion process computes the mean probability of the prediction from each individual classifier and decides the final label of the action by the highest mean probability.

5.1 MSR-Action3D Dataset

The MSR-Action3D dataset [6] contains 20 action classes and 10 subjects. Each subject performs each action two or three times. The 20 action types are chosen in the context of gaming. They cover a variety of movements related to arms, legs, torso, etc. The noise of the joint locations in the skeleton as well as the high intra-class variations and the inter-class similarities make the dataset very challenging. As an example, the action "Draw x" is easily confused with the action "Draw tick". We followed exactly the same experiment settings of [14], that all 20 action classes are tested in one group. Half of the subjects are used for training and the rest are for testing. We note there is another experiment setting used in the literature which splits the 20 action types into three subsets and only performs the evaluation within each subset [6]. The experiment setting we followed is more challenging than the subset one because all actions are evaluated together and the chance of confusion is much higher.

To evaluate how the height of the quantization pyramid can affect the recognition accuracy, we perform experiments on different levels of pyramid on the position feature and test their performances. Results are summarized in Table 1. In this experiment, zero to five levels of pyramid are evaluated. Zero means no hierarchical structure is employed, the entire sequence is quantized into 8 partitions by the proposed Iterated Dynamic Quantizing Algorithm. It can be regarded as the leaf-nodes of the 4-level pyramid. Two observations can be found from the results. First, as the increase of the number of levels, the recognition accuracy is also increasing. However the performance begins to drop when the levels reach five. This can be explained that the higher levels of the pyramid, the deeper hierarchical structure can be captured to describe the temporal dynamics. That's why the classification rate keeps increasing. Nevertheless, when the sequence is over quantized with too many partitions, the method becomes less generalized and the turning point occurs. Second, the accuracy of the 4-level pyramid (81.61%) is higher than that of using only the leaf-nodes (0-level, 73.56%). This has demonstrated that the hierarchical structure has contributed to the dynamic quantization of the se-

Levels	5	4	3	2	1	0
Accuracy(%)	77.39	81.61	71.26	67.82	66.28	73.56

Table 1: Experiment on different levels of quantization pyramid on the MSR-Action3D dataset.

Feature	Accuracy		
	Proposed	Deterministic quantization	w/o assignment sharing
position	81.61%	76.24%	81.23%
angle	73.95%	71.65%	72.41%
offset	73.95%	68.20%	64.75%
velocity	80.84%	72.80%	80.08%
fused result	90.42%	83.15%	88.51%

Table 2: Experiments on the effects of the dynamic temporal quantization and the frame assignment sharing strategy on the MSR-Action3D dataset.

quence. The inclusion of the upper layers has enhanced the generalization capability of the modeling method.

To also evaluate the effects of the Iterated Dynamic Quantizing Algorithm, we compare it against a deterministic quantizing method that always evenly splits the sequence. This is the same method used in the Fourier Temporal Pyramid [14]. Experiments results are summarized in Table 2. The second column and the third column show, the proposed algorithm with the dynamic quantization has a higher accuracy than the deterministic quantization method on all individual features as well as the fused result. Such performance increase demonstrates the advantage of the dynamic temporal quantization over the deterministic quantization method.

We also evaluate the performance of the frame assignment sharing strategy and compare it against the independent strategy. In the proposed method, the frame assignment of the quantization is computed with the position feature. The fourth column of Table 2 shows the performances of the strategy without the frame assignment sharing. We can see that the performance of the independent strategy is lower than the proposed method with the sharing. Other features which are less discriminative than the position feature benefit from the assignment computed based on the position feature. This has demonstrated the advantage of the frame assignment sharing strategy.

Last but not least, the fused result is significantly higher than the recognition rates of all individual features in all three columns in Table 2. One possible explanation is the multimodal features have complimentary discriminative capability which can be leveraged by the fusion and therefore yield superior performance over individual features alone. As an example, the position feature, in terms of joint coordinates, is good at describing the human pose from the global point of view. Meanwhile, the velocity feature based on the 3D point cloud is good at capturing the micro movements of the body parts. Therefore these two features from different modalities are complimentary to each other and can generate enhanced performance when fused.

We list state-of-the-art approaches in recent years in Table 3 for comparison. As can be seen, the proposed method achieves a higher recognition rate than most of state-of-the-art results. The only method achieves a higher accuracy than ours is the super normal vector method [18].

Method	Accuracy
Actionlet Ensemble [14]	88.2%
HON4D [10]	88.89%
DCSF [15]	89.3%
Lie Group [13]	89.48%
Super Normal Vector [18]	93.09%
Proposed approach	90.42%

Table 3: Comparison with state-of-the-art results.

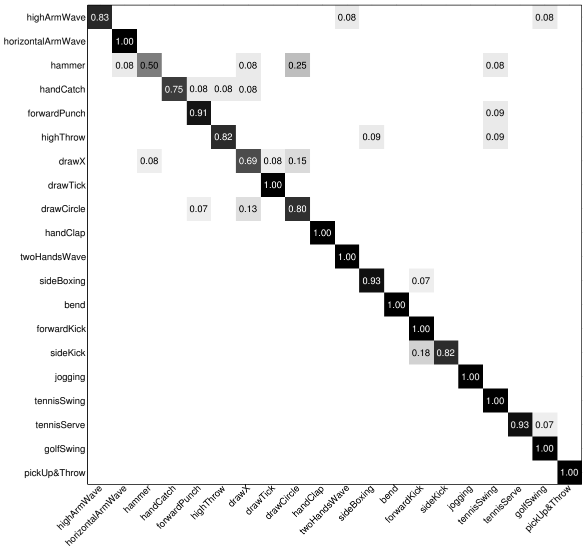

Figure 3: Confusion Matrix on the MSR-Action3D dataset.

Figure 3 shows the confusion matrix on the MSRAction-3D dataset. Most of the confusions between similar actions have been correctly addressed. The remaining confusions are between actions containing very similar primitive postures such as “Side kick” and “Forward kick” and “Draw x” and “Draw circle”.

5.2 UTKinect-Action Dataset

The UTKinect-Action dataset [17] consists of 199 action sequences in total. These sequences have 10 action types performed by 10 subjects. All subjects perform each action two times. Different from the previous dataset, subjects perform actions at varied locations in the scene. The huge viewpoint variation and intra-class variance make the dataset very challenging. We follow the same cross-subject test setting as in [13]. Half of the subjects are used for training and the rest are for testing.

We follow the same strategy in the MSR-Action3D dataset that the frame assignment of the quantization is computed based on the position feature and is applied to all other features. The performances of individual features and the fused result are summarized in Table 4. Similar results can

Feature	Accuracy		
	Proposed	Deterministic quantization	w/o assignment sharing
position	94.95%	88.89%	94.95%
angle	91.92%	87.88%	94.95%
offset	80.81%	74.75%	74.75%
velocity	79.80%	81.82%	78.79%
fused result	100%	96.97%	97.98%

Table 4: Experiments on the effects of the dynamic temporal quantization and the frame assignment sharing on the UTKinect-Action dataset.

Method	Accuracy
Histogram of 3D joints [17]	90.92%
Combined features with random forest [21]	91.9%
Lie Group [13]	97.08%
Proposed approach	100%

Table 5: Comparison with state-of-the-art results on the UTKinect-Action dataset.

be found that the proposed dynamic quantization algorithm achieved higher performance than the deterministic quantization on most of the features as well as the fused result. This has again demonstrated the advantage of the dynamic temporal quantizing algorithm. To our surprise, the fused result achieved the 100% accuracy. All samples in the testing set are correctly classified, which strongly demonstrates the performance of the proposed methods.

Table 5 shows the comparison of the classification accuracy between the proposed algorithm and state-of-the-art methods. Although the Lie Group [13] method achieves very good performance on this dataset (97.08%), the proposed algorithm still outperformed state-of-the-art results by achieving the 100% accuracy.

5.3 MSR-ActionPairs Dataset

The MSR-ActionPairs dataset [10] is composed of 12 actions performed by 10 subjects. Each subject performs all actions three times. Therefore, the dataset contains 360 action sequences in total. Different from the other two datasets, this dataset contains 6 pairs of actions. Each pair of actions has exactly the same primitive postures but the reversed temporal orders. As an example, “Pick up” and “Put down”, “Put on a hat” and “Take off a hat”. This dataset is collected to investigate the effects of temporal order on the recognition of the actions. The huge within-pair similarity makes the dataset very challenging. Therefore, histogram-based temporal modeling methods relying on the bag-of-words or visual codes may perform poorly on this dataset if the dynamic patterns are not properly interpreted. We follow the same test setting of [10] that the first 5 subjects are used for testing and the last 5 subjects are used for training. Experimental results are summarized in Table 6.

Feature	Accuracy		
	Proposed	Deterministic quantization	w/o assignment sharing
position	86.28%	86.85%	86.28%
angle	82.86%	82.86%	82.86%
offset	81.71%	81.14%	70.86%
velocity	89.71%	88.57%	90.28%
fused result	93.71%	91.43%	93.14%

Table 6: Experiments on the effects of the dynamic temporal quantization and the frame assignment sharing strategy on the MSR-ActionPairs dataset.

Method	Accuracy
Skeleton + LOP + Pyramid [14]	82.22%
HON4D [10]	93.33%
HON4D + D_{disc} [10]	96.67%
Super Normal Vector [18]	98.89%
Proposed approach	93.71%

Table 7: Comparison with state-of-the-art results on the MSR-ActionPairs dataset.

Similar to the results from the other two dataset, the position feature and the velocity feature show higher discriminative capability than the other features in all three columns of Table 6. The fused results all achieve higher performance than any individual feature alone by leveraging the fusion of multimodal data. One interesting result is, the strategy without the assignment sharing has equal or higher accuracy than the proposed method on all feature except the offset feature. However, fused result is still lower than that of the proposed method enabled with the sharing strategy. Such results justify our hypothesis that the dynamic quantization may get overfitted on some individual features and the overall generalization is compromised when multimodal results are fused.

We further compare the performance of the proposed method with state-of-the-art approaches and report the results in Table 7. It can be seen that the proposed method achieves a comparable accuracy to the methods listed. The only methods having a higher classification rate than ours are the histogram of 4D normals with discriminative projection [10] and the super normal vector [18]. These two methods rely on sophisticated features such as 4D surface normals and polynormals which are inapplicable to more general problems. Nevertheless, the proposed method provides a generic solution to the optimal temporal quantization problem and is independent from any features which makes the direct comparison unfair.

Figure 4 shows the confusion matrix of the dataset. Although the actions within each pair are highly confusing, the proposed method still achieves very good performance by discriminating the actions in each pair. It can be seen that the within-pair confusion only occurs between “Pick up box” and “Put down box”. Most of the actions with similar postures and temporal variations are correctly distinguished by the proposed method. Such performance has demonstrated the advantages of the proposed Dynamic Temporal Quantization method.

6. CONCLUSIONS

In this paper, we address the challenge of temporal modeling for human action recognition in multimodal streams. We study the problem of optimal temporal quantization of the

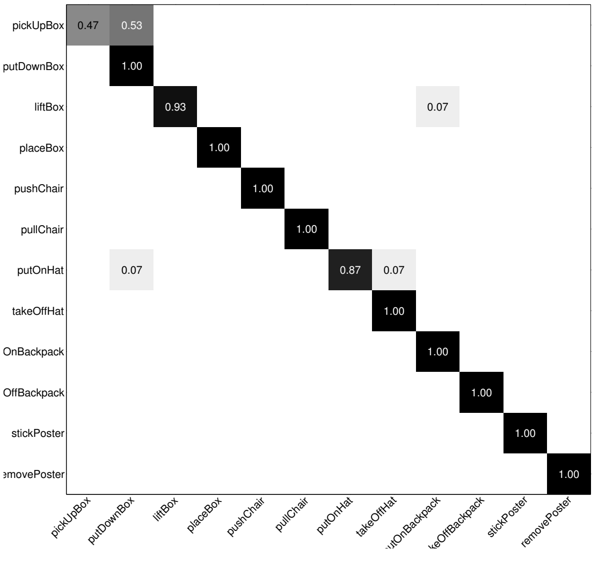

Figure 4: Confusion Matrix on the MSR-ActionPairs dataset.

video sequences and provide a solution of dynamic temporal quantizing. We further present a fusion method under this quantization framework to leverage the complementary discriminative capability of multimodal features. Experimental results on three public action datasets show the proposed approach has achieved state-of-the-art performance.

7. REFERENCES

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72.

[3] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5):836–849, 2010.

[4] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006.

[6] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on*, pages 9–14.

[7] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Computer Vision–ECCV 2006*, pages 359–372.

[8] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH*, pages 137–146.

[9] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[10] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition, 2013 IEEE Conference on*, pages 716–723.

[11] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[12] L. Sun and K. Aizawa. Action recognition using invariant features under unexampled viewing conditions. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 389–392, 2013.

[13] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2014.

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition, 2012 IEEE Conference on*, pages 1290–1297.

[15] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Computer Vision and Pattern Recognition, 2013 IEEE Conference on*, pages 2834–2841.

[16] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on*, pages 20–27.

[17] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on*, pages 20–27.

[18] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2014.

[19] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.

[20] 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*, pages 23–32, 2013.

[21] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *Computer Vision and Pattern Recognition workshops, 2013 IEEE Conference on.*