Performance Improvement Of Data Fusion Based Real Time Multi Layered Gesture Recognition Using One-Shot Learning Approach With Kinect V2.

S.Chandrasekhar, Dr.N.N.Mhala

Abstract— Gesture recognition offers a new medium for human-computer interaction that can be both efficient and highly intuitive. This paper proposes a data fusion based multi-layered gesture recognition method using one shot learning approach with Kinect V2. The Kinect sensor, have provided new opportunities for human-computer interaction. In recent years, human action recognition has drawn increasing attention of researchers, primarily due to its potential in areas such as video surveillance, robotics, human computer interaction, user interface design, and multimedia video retrieval. Kinect V2 to recognize different human hand gestures. It is fused with many advanced visual technologies and has been widely used in various kinds of computer vision tasks, such as face recognition, scene understanding, and human gesture recognition . Kinectv2 sensor which is capable of providing Depth Image ,RGB(intensity) Image , real time provision and Low cost.

Index Terms— One-Shot-Learning, Gesture Recognition, Multi-layered classification, depth image, Kinect.

1 INTRODUCTION

Research on multi layered gesture recognition has been intensified in the last decade. [1] In a multilayered framework, which extracts features from both the segmented semantic units and the whole gesture sequence and then sequentially classifies the motion, location and shape components. [3] Gestures are both natural and intuitive for Human-Computer- Interaction (HCI) and the one-shot learning scenario is one of the real world situations in terms of gesture recognition problems. [3] data fusion-based gesture recognition model by fusing depth information and RGB data. One-shot learning is an object categorization problem, found mostly in computer vision. one-shot learning involves three main challenges Representation, Learning and recognition. [4] One-shot-learning technique training only one example of each class is considered as the unique trait of this challenge whilst using more examples per sign typically improves accuracy. [5] Recently, Microsoft's Kinect has attracted increasing interest in both industry and research communities. Kinect v2 capture both RGB and depth information of a scene. When applied to gesture recognition, the depth information recorded by Kinect can be used to accurately track the human. Due to this appealing feature, it has been widely used in human action recognition and also gesture recognition. [5] In a multi-layered framework is proposed for gesture recognition, which first segment the moving hands and then extracts features from both the segmented semantic units and the whole gesture sequence. Although, the depth information recorded by Kinect can help us to detect and track the moving hands, accurately segmenting the fingers is still very challenging since fingers have many complicated articulations and they usually occlude each other. To explore one-shot learning gesture recognition, uses only one training sample per each class. Some important challenging issues for one-shot learning gesture recognition are the Following: To extract distinctive features and select the suitable model. The above two challenging issues and propose a new approach to achieve good performance for one-shot learning gesture recognition. Our experimental results reveal that our method is competitive to the state-of-the-art methods. The key contributions of the proposed method are summarized as follows:

• The phonological model of gestures inspires us to propose a novel multi-layered gesture recognition framework, which sequentially classifies the motion, location and shape components and therefore achieves higher recognition accuracy while having low computational complexity.
• Inspired by the linguistic sequential organization of gestures the matching process between two gesture sequences is divided into two steps: their semantic units are matched first, and then the frames inside the semantic units are further registered. A novel particle- based descriptor and a weighted dynamic time warping are proposed to classify the location component.
• The spatial path warping is proposed to classify the shape component represented by unclosed shape context, which is improved from the original shape context but the computation complexity is reduced from $O(n^3)$ to $O(n^2)$.
• Performed the operation of feature extraction Initially scaling down the frame sizes, local MHI i.e. obtaining MHI for the last $N$ frames at each consecutive frames, calculating the percentage change in MHI values in comparison to the total change in MHI values at each pixel, dividing the resultant frame into blocks, calculating the percentage change in interest points (IP) at each block finally performing two dimensional Fourier transform(2D-FFT) to get the final feature. Temporal segmentation of different gestures from the gesture sequences.
• Finally features are extracted for each gesture in a manner similar to that done for training samples. The testing and classification was done

The rest of this work is organized as follows. The rest of paper is organized as follows: Section 2 reviews the
existing approaches. In Section 3, we describe the proposed approach in detail. Section 4 presents the experimental results. In Section 5, we conclude the paper.

2. RELATED WORK
Farhad Dadgostar et al. [7] explored a Multi-layered Hand and Face Tracking for Real-Time Gesture Recognition system. In this method consist of three layers: Adaptive skin detection and segmentation, hand and face tracking which is performed by a fuzzy-based Mean-shift algorithm, and gesture recognition. This technique can be used with a variety of frontend input systems such as vision-based input, hand and eye tracking, digital tablet, mouse, and digital glove. The real world applications are affected by its cost and adaptability to the existing technology.

Lei Zhang et al. [8] proposed a novel bag of manifold words (BoMW) based feature representation for gesture recognition. In contrast to most existing feature extraction methods for gesture recognition, the proposed BoMW does not depend on the accurate segmentation of the body or fingers and therefore is suitable for more practical scenes. In this gesture recognition method is composed of several key parts including dense covariance descriptor extracting, cookbook learning on SPD manifold, BoW histogram representation and nearest neighbour classification. To detects the hands using scene depth information and then employs Dynamic Time Warping for recognizing gestures. The effectiveness of the proposed feature extraction method is also validated on a new RGB-D action recognition dataset.

Yared Sabinas et al. [9] implemented a new approach for early recognition of full-body gestures based on dynamic time warping (DTW) that uses a single example from each category. Our method is based on the comparison between time sequences obtained from known and unknown gestures. The classifier provides a response before the unknown gesture finishes. Input sequences are compared with stored ones by using DTW, a prediction criterion is proposed to determine when the method is confident of the identity of the gesture depicted in input sequences. In this method can work under the one-shot learning framework that is, using a single example of each gesture category to be recognized. This approach is easy to implement, it has no training phase and it is very efficient.

Decebal Mocanu et al. [10] developed a Mixture of Variational Auto Encoders (MoVAE), to perform classification. Complementary to prior studies, MoVAE represents a shift of paradigm in comparison with the usual one-shot learning methods, as it does not use any prior knowledge. MoVAE is capable of successfully performing the one-shot learning task, without the need of having prior knowledge, due to its generalization learning capabilities. MoVAE achieves very good results and high accuracy on a much more complicated real-world task, fashion products recognition. Even MoVAE shows to be very successfully in one-shot learning without prior knowledge, we believe that by making use of prior knowledge, its performance can be further improved.

Tao Yu et al. [11] propose a RNN-based multilayer parallel LSTM network to recognize human activities. This method focuses on the HAR with smartphone sensors by deep learning methods. We propose an approach based on the long short-term memory (LSTM) network to recognize human activities from the time series data collected by the inertial sensors attached to smartphones. In this method automatically extract features of time dependency from the original sensor data and classifies the activities with a softmax. This approach performs better than the traditional machine-learning methods, and achieves the similar performance as that of CNN, but it has lower computation complexity than CNN-based methods.

3. MULTI LAYERED GESTURE RECOGNITION USING ONE-SHOT LEARNING APPROACH
Gesture recognition offers a new medium for human-computer interaction that can be both efficient and highly intuitive. The multilayered gesture recognition system should be able to work with a variety of input devices. As the number of sensors is reduced, the performance might degrade, but it should degrade gracefully. In order to achieve device independence, we have reconceptualized gesture recognition in terms of a multi-layer framework. This involves generating a high-level, device-independent description of the sensed object. Multi layered Gesture recognition then proceeds from this description, independent of the characteristics of any given input device.

[12] Depth imaging technology has advanced dramatically over the last few years, finally reaching a consumer price point with the launch of Kinect. Depth cameras offer several advantages over traditional intensity sensors, working in low light levels, giving a calibrated scale estimate, being color and texture invariant, and resolving silhouette ambiguities in pose.

In spite of many recent successes in applying the Kinect sensor to articulated face recognition, human body tracking, and human action recognition, it is still an open problem to use Kinect for gesture recognition. The proposed multi-layered gesture recognition using for one-shot learning method contains some steps as shown in Fig.1. In the first layer, an improved principle motion is applied to model the motion component. In the second layer, a particle based descriptor is proposed to extract dynamic gesture information and then a weighted dynamic time warping is proposed for the location component classification. In the last layer, we extract unclosed shape contour from the key frame of a gesture sequence. Once the motion component classification at the first layer is accomplished, the original gesture candidates are divided into possible gesture candidates and impossible gesture candidates. The possible gesture candidates are then fed to the second layer which performs the location component classification. Compared with the original gesture candidates, classifying the possible gesture candidates is expected to reduce the computational complexity of the second layer distinctly. The possible gesture candidates are further reduced by the second layer. In the reduced possible gesture candidates, if the first two best matched candidates are difficult to be discriminated, i.e. the absolute difference of their matching scores is lower than a predefined threshold, then the reduced gesture candidates are forwarded to the third layer; otherwise the best matched gesture is output as the final recognition result. One-shot learning gesture recognition system that utilizes both RGB and depth information from Kinect sensor. We utilized depth sensor’s unique property to segment human silhouettes and perform a morphological denoising on depth images in the one-shot learning gesture recognition system, where customizable gestures can be recognized through the input of a Kinect sensor. Both the depth and color information obtained from the Kinect sensor is used for action representation, which ensures the robustness of our system to cluttered environments.
The ChaLearn Gesture challenge encouraged a few researchers to attack the problem of one-shot learning gesture recognition from RGB and depth images of the Kinect sensor provided in the ChaLearn Gesture dataset. The data are available from 2 formats A lossy compressed AVI format (5 GB) and a quasi-lossless AVI format (30 GB). It presents various features of interest.[15] The ChaLearn Gesture Dataset is designed for one-shot learning and comprises more than 50,000 gesture sequences recorded with Kinect. [16] The dataset was divided into training, validation and test sets by the challenge organizers. All three sets include data from different subjects and the gestures of one subject in validation and test sets do not appear in the training set.

3.1 MOTION COMPONENT CLASSIFICATION

During gesture recognition, gesture components in the order of importance are motion, location and shape. Understanding a gesture requires the observation of all these gesture components. None of these components can convey the complete gesture meanings independently. These gesture components complement each other. Gesture segmentation is used to segment a multi-gesture sequence into several gesture sequences. After the frames of the test and reference sequences are aligned, the next problem is how to represent the location information in a frame. In a multi-gesture sequence, each frame has the relevant movement with respect to its adjacent frame and the first frame. These movements and their statistical information are useful for inter-gesture segmentation. For a multi-gesture depth sequence \( l \), the Quantity of Movement (QOM) for frame \( t \) is defined as a two-dimensional vector

\[
QOM(l,t) = [QOM_{Local}(l,t), QOM_{Global}(l,t)]
\]

Where \( QOM_{Local}(l,t) \) and \( QOM_{Global}(l,t) \) measure the relative movement of frame \( t \) with respect to its adjacent frame and the first frame, respectively. They can be computed as

\[
QOM_{Local}(l,t) = \sum_{m,n} \sigma(l_t(m,n), l_{t-1}(m,n))\quad (2)
\]

\[
QOM_{Global}(l,t) = \sum_{m,n} \sigma(l_t(m,n), l_1(m,n))\quad (3)
\]

where \( (m,n) \) is the pixel location and the indicator function \( \sigma(x,y) \) is defined as

\[
\sigma(x,y) = \begin{cases} 1 & \text{if } |x-y| \geq \text{threshold}_{QOM} \\ 0 & \text{otherwise} \end{cases}
\]

where \( \text{threshold}_{QOM} \) is a predefined threshold, which is set to 60 empirically in this paper.

Dynamic regions in each frame contain the most meaningful location information. A simple thresholding-based foreground-background segmentation method is used to segment the user in a frame. The mask frame is then denoised by a median filter to get a denoised frame. The denoised frame is first binarized and then dilated with a flat diskshaped structuring element with radius 10 the binarized denoised frame from the dilated frame. The swing region (those white pixels in the swing frame) covers the slight swing of user’s trunk and can be used to eliminate the influence of body swing. From frame \( t \), define as

\[
\{ (m,n) \mid F_t(m,n) \land Threshold_{QOM} \}
\]

where \( F_t \) are the user masks of the first frame and frame \( t \), respectively. To compute the overlap rate between this region and the swing region. If the overlap rate is larger than \( r \), it is reasonable to think this region is mainly produced by the body swing. Therefore, it should be further removed.

[17] To calculating the absolute depth difference between current frame and the start frame for each gesture segment, as the input data of deep learning network

3.2 LOCATION COMPONENT CLASSIFICATION

Location component of two aligned frames can be represented as two particle sets

\[
P = \{ P_1, P_2, \ldots, P_N \},\quad Q = \{ Q_1, Q_2, Q_N \}
\]

The matching cost between particle \( P_i \) and \( Q_j \), denoted by \( C(P_i, Q_j) \), is computed as their Euclidean distance. The distance of the location component between these two aligned gesture frames is defined by the minimal distance between \( P \) and \( Q \). Computing the minimal distance between two particle sets is indeed to find an assignment to minimize the cost summation of all particle pairs

\[
\arg \min \sum_{i=1}^{N} C(R_i, Q_j)
\]

This is a special case of the weighted bipartite graph matching and can be solved by the Edmonds method, where \( n \) is the number of particles. The distance of the location component between two aligned gesture frames is obtained

\[
dis(P,Q) = \min_{i=1}^{N} C(R_i, Q_j)
\]

The distance between the reference sequence \( R \) and the test sequence \( T \) can be computed as the sum of all distance.
between the location components of the aligned frames.
Two factors influence the choice of the parameter $c$. [18] The
first one is the number of the gesture candidates and the
other one is the type of gestures. When the number of
the gesture candidates is large or most of the gesture candidates
are the shape dominant gestures, a high threshold is
preferred.

3.3 Shape Component Classification

The shape in a hold phase is more discriminative than the
one in a movement phase. The key frame in a gesture
sequence is defined as the frame which has the minimization
$QOMLocal$. Shape component classifier classifies the shape
features extracted from the key frame of a gesture sequence
using the proposed Spatial Path Warping (SPW), which first
extracts unclosed shape context (USC) features and then
calculates the distance between the USC of the key frames
in the reference and the test gesture sequences. The test
gesture sequence is classified as the gesture whose
reference sequence has the smallest distance with the test
gesture sequence. Spatial Path Warping algorithm (SPW)
is proposed to compute the minimal distance between
two unclosed contours. The contour of a shape consists of a
2-D point set $P \equiv \{P_1, P_2, P_N\}$
Their relative positions are important for the shape recognition.
Assume $P$ and $Q$ are the point sets for the shape contours of
the two key frames, the matching cost $(p_i, q_j)$ between two points
$p_i$ and $q_j$ in $Q$ is defined as

$$d(p, q) = \frac{1}{N} \sum_{k=1}^{N} \left|h_{p_{(k)}}(k) - h_{q_{(k)}}(k)\right|^2$$

(8)

Given the set of matching costs between all pairs of points
$p_i$ and $q_j$ in $Q$, the minimal distance between $P$ and $Q$ is to find a permutation to minimize the following sum

$$\text{arg min}_{\psi} \sum_{i=1}^{N} d(p_{i}, q_{\psi(i)})$$

(9)

An unclosed contour contains valuable spatial information.
Thus, a Spatial Path Warping algorithm (SPW) is proposed
to compute the minimal distance between two unclosed contours.
Compared with the Edmonds algorithm, the time
complexity of the proposed SPW is reduced from $O(n^2)$ to
$O(n^2)$ where $n$ is the size of the point set of an unclosed
shape contour.

3.4 Feature Extraction and Training

The depth data provided in the dataset are in RGB format and
therefore need to be converted to grayscale before processing.
The proposed methods of gesture recognition are an outcome of the series of operations performed for feature extraction. Scaling down is performed in order to reduce computational complexity. In the proposed scheme, the each frame is scaled down to $s\%$ of its original size and therefore the size of each $m \times n$ frame become $m \times n = nscale \times nscale$ where $mscale = m \times s/100$ and $nscale = n \times s/100$. In order to capture the direction of flow of motion and thereby the temporal position information of the moving parts the operation of taking the motion history image (MHI) of the lexicon is proposed. [19]

These advantages make these templates as a suitable
candidate for motion and gait analysis. In the MHI, the
silhouette sequence is condensed into grayscale images,
while dominant motion information is preserved. Therefore, it
can represent motion sequence in compact manner. This MHI
template is also not so sensitive to silhouette noises, like holes,
shadows, and missing parts. These advantages make these
templates as a suitable candidate for motion and gait analysis.
Conventional MHI has the advantage of representing a range
of times encoded in a single frame, and this way, the MHI
spans the time scale of human gestures. In the proposed
scheme, unlike the conventional approach of taking MHI over
a complete action, a local MHI of the last $N$ frames is
considered. The local MHI is taken on every one of the
background subtracted depth image frames of the gesture samples. The advantage of performing Fourier transform is that when the depth values of the person are taken to the frequency domain, the position of the person becomes irrelevant, i.e., any slight movement of the camera vertically or horizontally would be nullified in the frequency response. The above operation is performed on all the test vectors and the outcome of the 2D Fourier transform is utilized as feature for a particular action.

3.5 Temporal Segmentation

The test samples provided in the ChaLearn Gesture Dataset
contain one or more gestures. Therefore, the first task is
temporal segmentation of different gestures from the gesture
sequences... The test samples provided in the ChaLearn
Gesture Dataset contain one or more gestures. Therefore,
the first task is to separate the different gestures from the
gesture sequences. However, in order to focus solely on the
gesture recognition task, the temporal segment positions
provided with the ChaLearn Gesture Dataset by the
developers is utilized directly for separating different actions
from a single video file.[20]

3.6 Testing and Classification

After extracting the features from the gestures in the training
data of a particular lexicon, a feature vector table is
formed. Then, for the test samples, first temporal
segmentation is done if multiple gestures exist. Then features
are extracted for each gesture in a manner similar to that
done for training samples. It is a widely used measure for
image and gesture recognition[19] One of the classifiers
used is the correlation based distance measure. The
correlation coefficient is calculated between the features
obtained from the test gesture to that similar feature obtained
from each of the training gestures. [21] The coefficient of
correlation, $r$, is a mathematical measure of how much one
value is expected to be influenced by change in another one.
It is a widely used measure for image and gesture
recognition. The correlation coefficient between two images
$A$ and $B$ is define

$$Akl = \text{the intensity of pixel } (k, l) \text{ in image } A$$

and

$$Bkl = \text{the intensity of pixel } (k, l) \text{ in image } B$$

and $A$ and $B$ are the mean intensity of all the pixels of images $A$ and $B$, respectively. A high value of the correlation
coefficient between the same features of the test sample with
that of one of the training samples indicates a higher
probability of the two gestures to be identical and vice versa
for a low value of correlation coefficient.

3.7 One Short Learning Gesture Recognition

In this section, we describe the proposed approach based on
multilayered gesture recognition features for one shot learning
technique in detail. In the gesture recognition stage, it has five
steps: temporal gesture segmentation, feature descriptor extraction, coding descriptor, coefficient histogram calculation and the recognition results[22]. The overall process is discussed in Algorithm 1.

**Algorithm 1: one-shot learning gesture recognition**

**Input:**
- Training samples (RGB-D data): \( Tr = [tr_1, ..., tr_K] \)
- A learned codebook: \( B \) (computed from training stage)
- Coefficient histograms of training samples: \( Hr = [hr_1, hr_2, ..., hr_K] \) via Equation (11) (computed from training stage)
- A test sample (RGB-D data): \( te \)

**Output:**

The recognition results: \( class \)

1: Initialization: \( class = [ ] \)
2: Temporal gesture segmentation:
\[
[te_1, te_2, ..., te_N] = DTW(Tr, te), N \geq 1
\]
3: for \( i = 1 \) to \( N \) do
4: Spatio-temporal feature extraction: \( Xte = 3D \text{EMoSIFT}(te_i) \)
5: For \( Xte \), calculate its sparse representation \( C \) over the pre-trained codebook \( B \)
\[
\min_{C} \|Xte - BC\|_2^n \quad \text{s.t.} \quad \|c_j\|_0 \leq k, \ j
\]
6: Calculate the coefficient histogram \( hte \) via Equation (11)
7: Recognition: \( \text{tmp class} = \text{nn classify } (Hr, hte) \)
8: \( class = [\text{tmp class}] \)
9: end for
10: return \( class \)

This method gives superior recognition performance than many existing approaches. Then, features from each gesture sequence are extracted based on percentage change of interest points. In a multi-layered framework is proposed for gesture recognition, which first segment the moving hands and then extracts features from both the segmented semantic units and the whole gesture sequence.

**4. EXPERIMENTAL RESULTS**

In this section to evaluate these 3 classifiers and determine which component is more essential to gesture recognition. The results evaluated on development batches are separately shown in Figure 2 where the integrated feature consists of the gradient-based and motion-based features. It can be seen that the performance of the gradient-based features, which are comparative to the results of the integrated feature, are much better than the performance of the motion-based features.

**Fig. 2**: Classification with detection time of samples

Fig. 3 is the graphical representation of the overall accuracy as a function of the number of original examples used to generate artificial examples to augment the training data set. A reference value for the one-shot learning was included based on the recognition accuracy of each classification algorithm. As the number of real observations grow, there is an increasing trend for the recognition accuracy in both classifiers used.

**Fig. 3**: Accuracy of gesture Recognition for One shot approach (OSA) with ML(multilayered approach) and as number of samples

The performances of the proposed method are also limited by the accuracy of separation of different actions from the videos and unexpected movements of the performer. Also in some videos, the performer changed position while performing a gesture which added some unexpected features in the model that gave a wrong match. The performance of the proposed method is compared with a number of other methods proposed by several methods.

**Fig. 4**: PERFORMANCE ACCURACY WITH PROPOSED APPROACH

Template Matching technique the average of all depth
frames for each action is used as the descriptor. The test video is split in slices estimated using the average size of actions. In recognition phase each slice of the video is compared with all the templates. The performance of the proposed method is compared with a number of other methods proposed by several researchers in Table I.

### Table 1. Performance Comparison

<table>
<thead>
<tr>
<th>Methods</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Time Wrapping</td>
<td>49.41%</td>
</tr>
<tr>
<td>Template Matching</td>
<td>62.56%</td>
</tr>
<tr>
<td>BoVW</td>
<td>72.32%</td>
</tr>
<tr>
<td>Multi layered with One shot learning</td>
<td>91.36%</td>
</tr>
</tbody>
</table>

From these comparison results, we can see that the proposed method achieves high recognition accuracy while having low computational complexity. The first layer can identify the gesture candidates at the speed. The second layer has relatively high computational complexity. If we only use the second layer for classification, the average computing time is roughly n times of the first layer. Despite with relatively high computational cost, the second layer has stronger classification ability. Compared with using only the second layer, the computational complexity of using the first two layers in the proposed method is distinctly reduced and can achieve high. To describe the gesture data set used and present the framework for one-shot learning from one example of each gesture class. In this method we extracting a set of salient points within the gesture trajectory and finding a compact representation of each gesture class. This representation is then used to augment the number of examples of each gesture class artificially, maintaining intrinsic characteristics of the gestures within that class. Then the selection of classification algorithms and the training or testing methodology is presented. The used performance metrics are described, and finally an extension to adaptive learning approach is presented with preliminary results. Results of overall accuracy are calculated as the average of gesture accuracy per class.

### 5. Conclusion

This paper proposed a novel multi-layered gesture recognition with Kinect, by one shot approach for achieve high accuracy. In the first layer, an improved standard movement is connected to demonstrate the signal movement segment. In the second layer, a molecule based descriptor is proposed to extricate dynamic signal data and after that a weighted unique time traveling is proposed to characterize the area segment. In the last layer, the spatial way twisting is additionally proposed to characterize the shape segment spoken to by unclosed shape setting, which is improved from the first shape setting however needs lower coordinating time. The proposed technique can acquire moderately superior for one-shot learning motion acknowledgment. The proposed technique gives unrivaled acknowledgment execution than many existing methodologies. We would also like to combine more than one discriminatory window for each gesture to improve performance. The hand gesture recognition technique can mimic the communications between human and involve hand gesture as a natural and intuitive way to interact with machines.

### 6. References


