A Real Time Temporal Segmentation Method for Continuous Gestures Analysis

Yongmian ZHANG*, Quentin AUGE*, Haisong GU*

Abstract

The forthcoming demand of the natural use of the human hand as human–computer interaction motivates research on continuous hand gesture recognition. Gesture recognition relies on gesture segmentation to find the boundary of gestures with semantic meanings while ignoring unintentional movements. However, gesture segmentation on a stream of continuous gestures poses a challenge due to the movement ambiguity of successive gestures and unconstrained spatiotemporal variation. To address this challenge, our approach entails three major steps: the first step applies Maximum Mean Discrepancy criterion to detect the change-points over continuous gestures as the initial estimated cuts of the gesture transitions; the second step uses kinematic constraints to revise an initial estimated cut to an accurate gesture transition position; and finally a probability density estimation is used to estimate the hand motion between two cuts to eliminate unintentional movements and non-gesture segments. The proposed method is evaluated by using Chalearn benchmark datasets with 150 sequences of continuous sign gestures. The proposed method achieves the accuracy rate of 82.7 %, which outperforms the state-of-the-art approach.

1 Introduction

Recently users have come to demand a natural user interface (NUI), touchless interaction, and automatic sign language analysis. Recognizing gestures has undergone a rapid growth, starting from isolated gestures performing subsequently with a pause in-between to natural continuous gestures like sign language communication in which a subject may not pause between successive gestures. The recognition of a stream of continuous gestures is much more difficult than that of an isolated gesture because the transition point between consecutive gestures cannot be readily detected.

Previous work on the recognition of continuous gestures can be mainly divided into two categories. On one side, some works use Hidden Markov Models (HMMs) for continuous gesture recognition [1][2][3]. The HMMs are concatenated by a number of sub HMMs representing corresponding individual gestures. The entire gesture HMM is trained on the entire observation sequence for the corresponding gestures, all possible gesture boundaries are inherently considered and no additional segmentation is needed. However, the approach would risk expensive computational cost, which makes them infeasible for real-time and online applications. On the other side, many works have attempted to temporally cut a sequence of continuous gestures into segments with different semantic meanings and then each segment of gesture is used as an input to the trained models of the predetermined gesture classes [4]. In these approaches, the segmentation is a crucial step for achieving accurate gesture recognition. Despite significant research efforts over the past few decades, extracting individual gestures with semantic meanings from a sequence of continuous gesture still remains a challenge due to movement ambiguity between successive gestures and unconstrained spatiotemporal variations in gesture.

*Konica Minolta Laboratory U.S.A., Inc.

In the past few decades, several approaches have been proposed for recognizing human gestures from monocular RGB video sequences [3] [4]. Unfortunately, the RGB data is highly sensitive to various factors like illumination changes, variations in view-point and background clutter. Recently introduced cost-effective depth sensors with the real-time skeleton extraction have generated renewed interest in human gesture recognition particularly for NUI. Motivated by this, we use a stream of skeleton data of continuous gestures from a 2.5D sensor as input, as shown in Figure 1.

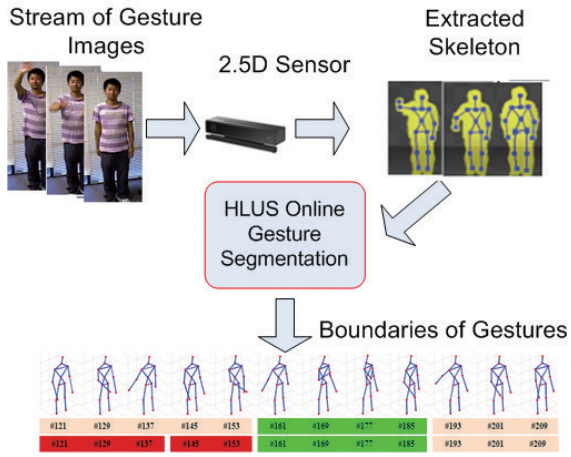

Fig. 1 System flowchart, where the input is 2:5D depth sensor data. Output is temporal segments of the gestures.

To address the aforementioned difficulties in extracting the boundaries of the gestures embedded in a continuous stream of motion, we propose a novel approach which entails the following components.

1) We first use Maximum Mean Discrepancy (MMD) criterion [5] is used to capture differences among the distributions of spatiotemporal patterns of the body joints over a stream of continuous gestures. The difference provides estimated cuts of a gesture sequence.
2) Kinematic constraints are then used to measure the kinematic change around estimated cut points and then to revise them to accurate positions.
3) Finally, probability density estimation is used to estimate the movement between two cuts so that non-gestures and the unintentional movements can be eliminated.

The remainder of this paper is organized as follows: Section 2 presents an overview of the existing work and discusses their similarities and differences to our work. Section 3 introduces the details of our approach. This is then followed by the experimental results and analysis in Sections 4. The final section provides a summary of this work and the concluding remarks.

2 Related Works

Human gesture recognition has been an active area of research for the past several decades due to its forthcoming applications in human-machine interactions. Most of approaches in gesture recognition research assume the gestures be isolated each other or the gestures be already segmented. The earliest work concentrating on the issues of temporal segmentation for a stream of continuous gesture is [2]. In this work, HMMs are employed to represent the gestures and their parameters are learned from the training data. For continuous gesture recognition, the HMMs are concatenated, and each HMM is instantiated with a corresponding gesture. This large concatenated HMM can then be trained by the corresponding data. Because the entire gesture HMM is trained on the entire observation sequence for the corresponding gestures, all possible gesture boundaries are inherently considered and no additional detection is required to find gesture boundaries. However, as mentioned previously, this approach would risk expensive computational cost and a huge concatenated HMM makes this approach infeasible for real time applications.

Temporal clustering has also been proposed to perform temporal segmentations from video sequence by discovering motion primitives. Temporal clustering is to find the patterns of multidimensional time series into a set of disjoint segments that belongs to k temporal clusters [11] [12]. They generally combine kernel k-means with the generalized dynamic time alignment kernel to cluster time series data. The approach benefits from a global point of view on the data and provides cluster labels as in clustering. Unfortunately, it is not suitable for online/real-time applications such as gesture recognition for natural user interface and touchless interaction.

Change-point methods rely on various tools from signal theory and statistics [8] [5] to localize frames of abrupt change in pattern within the flow of motion. Unlike temporal clustering, the change-point approach relies on local patterns in time-series and often results in unsupervised online algorithms which can perform real-time. Although they used to be restricted to univariate series with parametric distribution assumption, the recent use of kernel methods [5] [7] [9]

released part of these assumptions, and they were successfully applied to the temporal segmentation problem [6][9][10]. Among these approaches, Kernelized Temporal Cut (KTC) algorithm [7] is the most similar to our approach. KTC models a temporal segmentation problem as a series of two-sample problems within varying-size sliding windows, and solves it by using a test statistic based on Maximum Mean Discrepancy. The algorithm is intended to segment continuous actions, given they are both long enough (few seconds) and cyclic (e.g. walking, running, boxing). However, most of gestures are non-cyclic and performed in a short time which is unknown in prior and varies with different gestures.

Besides the above works, several authors have attempted to address the issues of continuous air writing gestures particularly in writing the Arabic numbers [3][13]. In [3], the authors used orientation dynamic features obtained from spatio-temporal trajectories to train HMM. For the continuous gestures, they observed that there is a line segment between each gesture ends. As such, the problem is converted to detect the line segment to separate the gestures. The similar features are used in [13], but in this work conditional random field (CRF) is trained with one label for each gesture and an extra label is added to determine the non-gestural movements. The transition between two air writing gestures is modeled by the dependencies of non-gestural movements with gestures. Thus, a combination of gesture and non-gestural sequences are exhaustively used in the training data and gesture boundaries are inherently considered. This approach is only feasible for the 10 Arabic numbers. Our approach deviates significantly from these approaches. We are considering more general gestures in which the transition between two gestures cannot be readily distinct.

3 Temporal Segmentation

In this section, we provide the details of our approach for continuous gesture segmentation.

3. 1 Estimate Gesture Cuts

Let Z be an arbitrary set and $Z=(z_1, z_2, \ldots, z_{m+n}) \in Z^{m+n}$ with m, $n \in T$, MMD is a criterion for testing whether $w_1=(z_1, z_2, \ldots, z_m)$ and $w_2=(z_{m+1}, z_{m+2}, \ldots, z_{m+n})$ come from different distributions [5]. The basic idea of MMD is to measure the similarity of points w_1, $w_2 \in Z$ using a kernel $k(w_1, w_2)$. In our case, Z is the set of gesture frames (a stream of continuous gestures). As shown in Figure 1, each frame is a human body which is represented by a vector of 3D positions of N body joints, i.e., $J \in R^{3N}$. Therefore, a stream of continuous gestures is a set of $Z=R^{3\times N\times T}$.

Now let w_1, w_2 be two time sliding windows with the same length T_0 and the same moving step ΔT, where T_0, $\Delta T \in T$. Let t be an arbitrary time point at a stream of continuous gesture. The time boundaries of w_1 and w_2 are then $l_1:[t-T_0, t-1]$ and $l_2:[t, t+T_0-1]$, respectively. To test whether w_1 and w_2 come from the same distribution, we can use MMD criterion as follows:

$$g(x, l_1, l_2)=\frac{1}{T_0}[k(x, l_1, l_1)-2k(x, l_1, l_2)+k(x, l_2, l_2)], \quad (1)$$

where

$$k(x, l_1, l_1)=\sum_{i\in l_1}\sum_{j\in l_1} k(x_i, x_j),$$

$$k(x, l_1, l_2)=\sum_{i\in l_1}\sum_{j\in l_2} k(x_i, x_j),$$

$$k(x, l_2, l_2)=\sum_{i\in l_2}\sum_{j\in l_2} k(x_i, x_j),$$

MMD here quantifies the global motion of body which measures the consistency between two segments (w_1 and w_2) of gestures. For kernel $k(x, y)$, we choose a Gaussian kernel, that is,

$$k(x, y)=\exp\left(-\frac{||x-y||^2}{2\sigma^2}\right), \quad (2)$$

where ||.|| denotes the Euclidean norm and the smoothing parameter σ should be tuned to the problem at hand.

The temporal cut of a gesture can be estimated by maximizing Eqn.(1). In other words, the inconsistency of two segments occurs at the turning points of the local maxima on the MMD curve, i.e.,

$$\frac{d}{dt} g(x, l_1, l_2)=0 \quad (3)$$

Figure 2 plots a MMD curve from a continuous gesture stream containing two semantically different gestures, where the green lines denote the turning point of the local maxima of MMD, while the red lines are the ground true of gesture transition. Several observations can be made from MMD curve as shown in Figure 2:

1) Theoretically, the turning points of the local maxima are the gesture cut points. However, because the value of MMD is calculated by averaging within a sliding window, there is a shift between the turning point of the local maxima and its ground true cut.
2) A segment of gesture or non-gesture is in-between two cut points. However, whether the segment is a non-gesture or a gesture cannot be determined directly from MMD curve.
3) Because the extraction of body joint position involves errors, false turning points of the local maxima are often detected on the MMD curve.

These issues are addressed in the following subsequent sections.

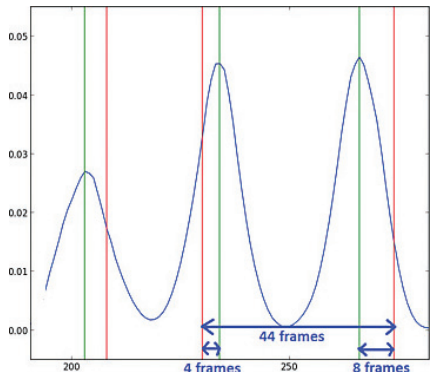

Fig. 2 A typical MMD curve (blue) calculated from a stream of continuous gestures. Green lines denote local maxima of MMD. The red lines are the ground truth of gesture transition. There is a shift (4 and 8 frames in this example) between the turning point of the local maxima and the ground true cut.

3. 2 Kinematic Constraints

We call the cut points found by using MMD through Eqn.(3) as an estimated cut point because there is a shift between the turning point of the local maxima and its ground true cut. Generally, if a person wants others to understand the meaning of individual gestures when gesturing continuously, he always starts and ends a meaningful gesture intentionally with a pause state, no matter how obvious the pause is. This has also been observed in a sentence gesture [6]. Accordingly, it implies that there must be at least a kinematic change on the transition between two meaningful gestures.

We use kinematic constraints to modify the estimated cut points to accurate positions. Additionally, the false detection of cut points can be eliminated. We use the rate of change of velocity and the rate of change of acceleration, as known as jerk to measure the kinematic change. For the rate of change of velocity, we have

$$a(t)=\frac{v(t+\Delta t)-v(t-\Delta t)}{2\Delta t+1}, \tag{4}$$

where

$$v(t)=1-\frac{1}{2\Delta t+1}\sum_{t=-\Delta t}^{\Delta t} e^{-r(\|X_t, X_{t+2\Delta t}\|^2)},$$

Here, the reason we used an exponential kernel function to model the velocity is to increase the sensitivity. The sign of acceleration at a cut point determines whether the cut is a start or an end of a gesture segment. If $a(t)>0$, the cut at t is a start point; an end point otherwise.

For the rate of change of acceleration, we use the following approximate jerk equation

$$J(t)=v(t-\Delta t)-2v(t)+v(t+\Delta t). \tag{5}$$

If jerk has a positive peak around the estimated cut at t, the cut shall shift from the estimated cut to this peak.

If a gesture is a hand gesture, the hand position which can be readily detected through the skeleton joints can be used a constraint to determine whether the cut is valid or not. For example, either the left hand or the right hand is in "up" state (the middle of gesturing), a cut shall not occur. If there are several cut candidates, the true cut shall locate the one with the lowest hand position.

Figure 3 illustrates how the above rules are applied to revise the estimated cut to a true position of a gesture transition.

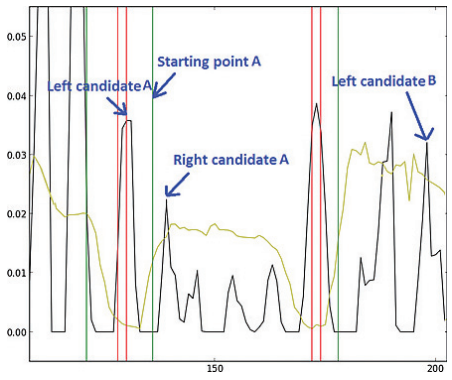

Fig. 3 The "starting point A" is an estimated cut by using MMD as in Fig. 2. The "left candidate A" and the "Right candidate A" around the estimated cut are picked according to jerk peaks (dark line). The true cut shall be one of the jerk peaks with the lower hand position (yellow line), i.e., "left candidate A". The red vertical line indicates the ground truth of the gesture transition.

3. 3 Gesture and Non-gesture Detection

Whether there is a meaningful gesture or a non-gesture segment between the two cuts cannot be known through MMD and the kinematic model. Thus, we need to detect gesture and non-gesture between the cuts. Additionally, the unintentional movements caused by skeleton detection can be eliminated.

Based on our inspection, the distribution of a joint trajectory in the spherical coordinates (Figure 4) approximately follows a Gaussian mixture model. Parzen-window density estimation using a Normal kernel function is a generalization of the Gaussian mixture model and each single sample of the N samples is considered to be a Gaussian distribution by itself. We choose a normal function $N(0, \sigma^2)$ as kernel and then,

$$P(X_t)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi|\Sigma|)^{1/2}}e^{\frac{1}{2}(X_t-X_i)\Sigma^{-1}(X_t-X_i)} \quad (6)$$

where X is the position of a body joint in the spherical coordinate system as shown in Figure 4, and $X=[r, \theta, \varphi]^T$. Here N is the number of frames between two cuts. If we assume independence among r, θ, φ with different kernel bandwidths, then the density estimation is reduced to

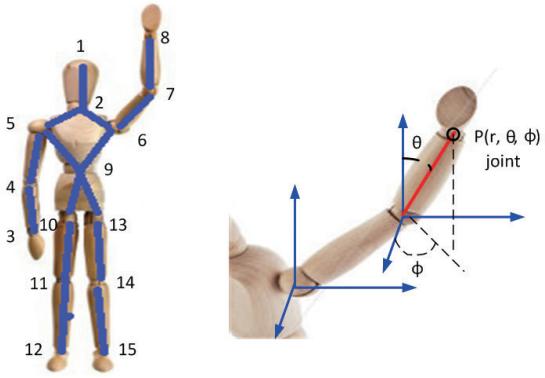

Fig. 4 Body joints in the spherical coordinate system.

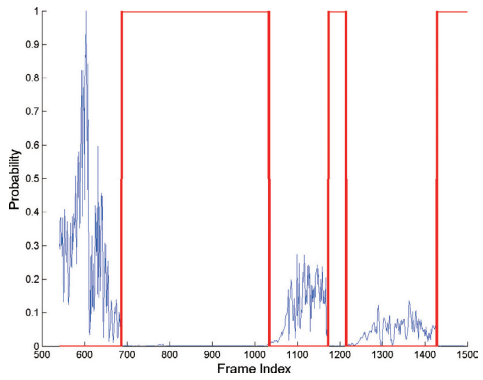

Fig. 5 A user performs three different gestures continuously. The probability density is in blue color. Non-gesture segments are superimposed by a red box.

$$P(X_t)=\frac{1}{N}\left\{\prod_{v\in\{r,\theta,\phi\}}\frac{1}{(2\pi|\Sigma|)^{1/2}}e^{-\frac{1}{2}(X_t-X_i)^T\Sigma_v^{-1}(X_t-X_i)}\right\} \quad (7)$$

where Σ is a diagonal matrix of the variance of r, θ, φ. A body joint is considered gesturing if the density is over a predetermined threshold which can be empirically determined through cross validation. $P(X_t)$ is normalized to 0 -1. The value of $1-P(X_t)$ indicates the gesture and non-gesture. Additionally, Parzen-window density estimation plays a role of smooth filter to remove the false detection caused by the errors from skeleton extraction. Figure 5 illustrates the generated gestures and non-gestures given a set of temporal cuts points. In addition, the acceleration described in the previous section may be positive in the case that, even though there is no start of gesture, a gesture ends with a slight acceleration due to movement inertia. This sometimes results in a false positive gesture even though the user does not move within the rest of the segment. The motion estimation also helps to eliminate such false positive segmentation.

4 Experiment

In this section, we report the segmentation performance by evaluating with a benchmark dataset [14]. Our approach is compared against KTC algorithm [7], the latest state-of-the-art approach. We first describe the evaluation metrics.

4. 1 Evaluation Metrics

In statistics, Jaccard Index is used for comparing the similarity and diversity of two sample sets A and B:

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|},\ 0\le J(A,B)\le 1 \quad (8)$$

where $B\neq\emptyset$. Let g be a labeled ground-truth segment and $g\subset G$, s, be a detected segment and $s\subset S$, then Mean Jaccard Index can be derived as

$$J(S,G)=J\left(\bigcup_{g\in G}g,\bigcup_{s\in S}s\right) \quad (9)$$

However, this metric is good for evaluating action recognition and it is inappropriate for evaluating the gesture segmentation. Figure 6 illustrates an example showing the deficit of the above metric for segmentation evaluation. From this figure, though Jaccard Index

scores 100%, it is still either over-segmented or under-segmented. This is because the metric does not take cuts into consideration. When applying for evaluating action recognition performance, two adjacent segments containing the same action are automatically merged, while two adjacent segments are not necessarily merged in gesture segmentation, as they are likely to contain two different actions in a stream of continuous gestures with no pause in-between.

To overcome the above problem, we propose a new metric by combining Jaccard Index and both precision and recall in F measure as

$$F = 2 \times \frac{R \times P}{P + R} \tag{10}$$

where P, R denote precision and recall and they can be calculated by the following equations:

$$R = \frac{1}{|G|} \sum_{g \in G} \max_{s \in S} J(g, s), \qquad P = \frac{1}{S} \sum_{s \in S} \max_{g \in G} J(s, g). \tag{11}$$

A recall measures how much ground-truth segmentation matches a generated segmentation while a precision measures how much a detected segmentation matches ground-truth segmentation.

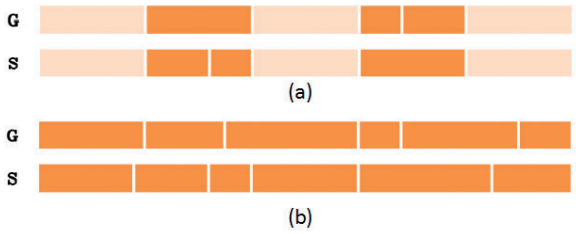

Fig. 6 Explanation of the shortcoming using Eqn. (9) as evaluation metric for gesture segmentation. G and S denote the ground-truth and the detected segments, respectively. A gap (white color) between two segments is a cut position. (a),(b) shows that the result is either over-segmented or under-segmented though the value of Jaccard Index is 100%.

4. 2 Result

We use a benchmark dataset namely Chalearn to evaluate the performance of our approach. The datasets consists of more than 14,000 gestures in 940 sequences collected particularly for computer vision challenge contest in gesture recognition [14]. The gestures are drawn from a vocabulary of 20 Italian sign gesture categories performed by several different users, with the aim of performing user independent continuous gesture spotting. For most of these gestures, their start frame and end frame are annotated, but some of them are not. We choose full-annotated 150 sequences for evaluating the segmentation performance of our approach. We compared the performance of our method against the KTC algorithm [7]. Table 1 is the comparison result showing that our approach achieves 82.7% accuracy rate, which significantly outperforms KTC. Figure 7 presents an arbitrary result from 150 sequences for visually inspecting the quality of segmentation by our method.

Table 1 Comparison result.

	Recall	Precision	F1 Score
KTC	61.70%	69.01%	65.01%
Our Method	86.47%	79.80%	82.74%

Fig. 7 An example of segmentation result takes from 150 sequences. The generated segments match the ground-truth very well, where green indicates the match while red means mismatch.

From the experiment, we found that the acceleration is more sensitive to the error of skeleton detection, which makes a wrong decision of the boundary of a gesture. This contributes the major false positive in the result as shown in Table 1. This will be our future work to address this issue.

Figure 8 illustrates another type of gestures where a Karate Master performs continuously the 10 different moves of Karate martial art. The sequence consists of a thousand of frames. We want to separate these 10 moves by cutting at the final pose of each move. Figure 9 shows the segmentation result using our method against the ground truth. Again, the result shows that the generated segments agree the labeled segments very well though few segments have about 5-8 frames error compared with the ground-truth.

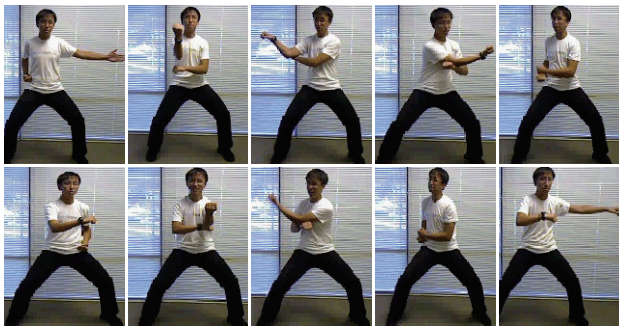

Fig. 8 A Karate Master performs continuously the 10 moves of Karate martial art. This figure only shows the final pose of each move.

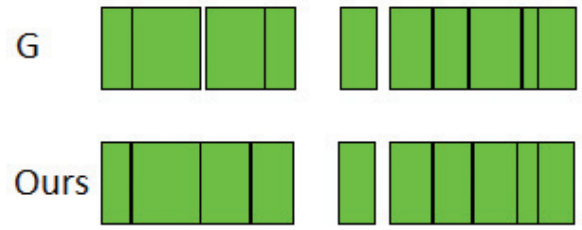

Fig. 9 The top segments are the ground-truth and the bottom ones are the generated segments by our method.

5 Conclusion

The recognition of a continue gesture relies on gesture segmentation to find a boundary of a meaningful gesture. However, the transition between two consecutive gestures cannot be readily distinct. To address this difficulty, we proposed a novel approach which entails three major steps. The first step applies Maximum Mean Discrepancy criterion to detect the change-points over continuous gestures as the initial estimated cuts of the gesture transitions; the second step uses kinematic constraints to revise the initial estimated cuts to an accurate gesture transition position; and finally, probability density estimation is used to estimate the hand motion between two cuts to eliminate unintentional movements and segments with no-gestures. Our approach outperforms the state-of-the-art approach in term of segmentation accuracy. Additionally, the algorithm runs online and real-time, which can be applied to the applications such as a natural user interface, touchless interaction, and automatic sign language analysis.

Reference

[1] Lee H. K. Lee and J. H. Kim, An HMM-based threshold model approach for gesture recognition. IEEE Trans Pattern Anal Mach Intell. 21(2): 961–972, 1999

[2] J. Yang and Y. Xu, Hidden Markov Model for gesture recognition, CMU-RI-TR-94-10, 1994

[3] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory, International Conference on Pattern Recognition, 2008

[4] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. PAMI, 23(3):257–267, 2001

[5] A. Gretton, K. M. Borgwardt, M. Rasch, B. Scholkopf and A.J. Smola (2007a). A Kernel method for the two-sample-problem. Proc. Of Advances in Neural Information Processing Systems, 2006.

[6] E. J, Davis Gestural Cues for Sentence Segmentation. Technical report R2005, MIT AI Memo, pp 1–14

[7] D. Gong, G. Medioni, S. Zhu, and X. Zhao. Kernelized temporal cut for online temporal segmentation and recognition. European Conference on Computer Vision, 2012

[8] E. G. Carlstein, H. G. Muller and D. Siegmund. Change point problems. IMS, 1994.

[9] M. Tanzini, P. Tripicchio, E. Ruffaldi, and G. Galgani, Human gesture segmentation based on change point model for efficient gesture interface, in IEEE International Symposium on Robot and Human Interactive Communication, 2013

[10] K. Kahol, P. Tripathi, and S. Panchanathan. Automated gesture segmentation from dance sequences. In IEEE International Conference of Automatic Face and Gesture Recognition, 2004

[11] F. Zhou, F Torre, and J. K Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 3(35), p582-596, 2013.

[12] P. Turaga, A. Veeraraghavan, and R. Chellappa Unsupervised view and rate invariant clustering of video sequences, Computer Vision and Image Understanding, 113(3):353 {371, 2009

[13] M. K. Bhuyan, D. A. Kumar, K. F. MacDorman and Y. Iwahori, A novel set of features for continuous hand gesture recognition, Journal of Multimodal User Interfaces, 02 July 2014.

[14] http://gesture.chalearn.org/