Automatic gesture segmentation based on a predictive event segmentation approach

Sofija Spasojevic
http://www.pupin.rs/RnDProfile/spasojevic.html
Mihailo Pupin Institute, Robotics department
Faculty of Electrical Engineering, University of Belgrade

Rodrigo Ventura
http://users.isr.ist.utl.pt/~yoda
Institute for Systems and Robotics
Instituto Superior Técnico, Lisbon

Abstract

This paper presents preliminary work related to one approach for automatic gesture segmentation. We take a predictive event segmentation approach, according to which events are detected once sensor data departs significantly from an adaptive model-based predictor. During gesture execution, 3D positions of characteristic human joints are collected using a Kinect device. Data set consists of joints from interest, whose trajectories are modelled as Gaussian processes. Based on this data set, a Gaussian process based predictor is adapted and used to detect transitions between gestures. The preliminary results over the collected dataset are encouraging and illustrate the outcome of the approach.

1 Introduction

Recent years, research in area of human gesture recognition have become very intensive and number of publications related to this area have increased significantly. Gesture recognition systems have significant applications in many different fields such as virtual and augmented reality [1], industrial process control [2], physical rehabilitation [3], human-robot interaction [4], computer games [5] etc.

Each gesture recognition system is a complex structure which contains from several functional units. First step in every gesture recognition process is gesture acquisition. A lot of techniques for gesture acquisition, based on different type of sensors have been developed. In general, all sensors used for collecting information about human movement can be categorized in two groups – body and visual sensors. Body sensors are placed on the body parts and provide information directly each time the movement occurs. Some examples of most used body sensors are accelerometers, data gloves and special body suits that using optical or electromechanical tracker technologies. Visual sensors refers to camera systems used for recording gesture sequences, followed by image processing techniques in order to characterize each gesture with a set of visual features. Some gesture acquisition techniques include both type of sensors in order to obtain more accurate information. Attaching sensors or body markers on the body, as well as wearing special suits many people find uncomfortable. On the other hand, visual sensors can be very expensive and their performance depends a lot of illumination and background conditions. Considering above mentioned items, we decided to use Kinect device for gesture acquisition process.

Kinect is the new generation low-cost device developed by Microsoft, which consists of variable-resolution RGB camera, 3D depth sensor (infrared projector combined with infrared camera) and broadband microphone. Kinect has a corresponding user interface and compatibility to work with different software packages. It also has some embedded algorithms, one of which relates to skeleton detection and collecting 3D joints position. This algorithm presents a base for gesture acquisition process used in this work.

Another important topic when dealing with gestures is gesture segmentation, which is the pre-phase of applying gesture recognition algorithm. This phase is very important, considering the influence that inaccurate segmentation can have on process of gesture recognition. There are a lot of techniques applied in gesture segmentation problems.

Technique based on simple sliding window combined with simple moving average filter is used in [6]. Author defines content of each gesture in the following form: starting static posture, dynamic gesture part and ending static posture. In addition, to obtain more robust segmentation, author observe also the length of each analyzed sequence to eliminate appearance of static part into dynamic part of gesture. In [7] authors developed algorithm for segmentation of dance sequences. This algorithm, called *Hierarchical Activity Segmentation*, is based on division of human body onto hierarchically dependent structures. They take into account relevant motion parameters for body segments

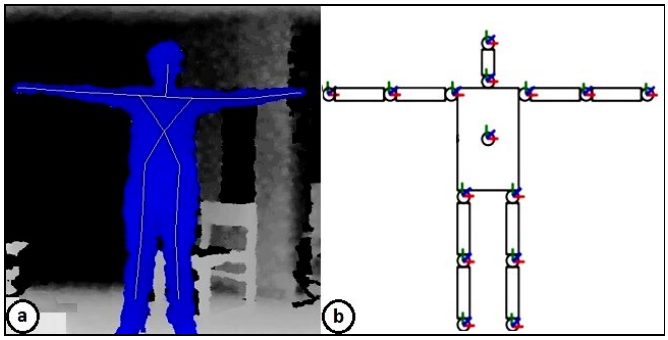

Figure 1: Skeleton tracking (a) and joints whose 3D positions are collected (b) by Kinect

(segmental force, kinetic energy and momentum) that characterize motion in the levels of defined hierarchy. In [8] the authors took a dynamical system approach for dynamic system identification, however that approach did not account for sensor noise. In this paper, we study an approach using Gaussian processes as machine learning method [9] that provides information about both, value and uncertainty. In addition, this method has shown good properties related to complexity model and processing time.

2 Algorithm for gesture segmentation

As previously mentioned, process of gesture acquisition is realized using Kinect device. Some functions from OpenNI software package designed for work with 3D sensors were used in order of collecting information from Kinect. During skeleton tracking, 3D positions of 15 joints (Fig. 1) for each frame are collected. In this preliminary work we concentrate on segmentation of upper body gestures, concrete gestures performed by arm (Fig. 2). Accordingly, only elbow and hand joints are important for further analysis. After normalization of coordinates with respect to torso, data set is configured.

We modelled trajectories of elbow and hand joint positions as Gaussian processes and formed three predictive Gaussian prediction model (each model for one coordinate). If we want to make predictions for following values from existing set of normally distributed variables, we can define joint distribution as:

$$\begin{bmatrix} X \\ X_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{bmatrix}\right) \quad (1)$$

where X is the known function values of the training cases and X_* is a set of function values corresponding to the test set inputs. $\mu = m(x_i)$, $i=1,\ldots,n$ for the training means and analogously for the test means μ_*; for the covariance we use Σ for training set covariances, Σ_* for training-test set covariances and Σ_{**} for test set covariances. Since the values for the training set are known, we should determine conditional probability distribution of X_* given X, which is expressed as:

$$X_* \mid X \sim N(\mu_* + \Sigma_*^T \Sigma^{-1}(X - \mu), \Sigma_*^T \Sigma^{-1} \Sigma_*) \quad (2)$$

This is the posterior distribution for a set of test cases.

Very important item when dealing with Gaussian predictive models is process of hyperparameters estimation. Depending on the form of training and testing set samples, number of hypereparameters that define mean and covariance function can vary. Estimation of hyperparameters is achieved by maximization of the log-likelihood function.

Acknowledgement: This work was partially funded by bilateral project COLBAR, between Instituto Superior Técnico, Lisbon, Portugal and Mihailo Pupin Institute, Belgrade, Serbia, and FCT [PEst-OE/EEI/LA0009/2013], III44008 and TR35003.

Figure 2: Gesture acquisition using Kinect: RGB (top) and depth (bottom) streams from Kinect

The optimization requires the computation of the derivative of log-likelihood function with respect to each of the parameters. Let n be the number of frames in gesture sequence and x_i value of x-coordinate in i-th frame. Training set consists of samples that are organized as five dimensional vectors:

$$X = ([x_1 ... x_5]; [x_2 ... x_6]; \ldots ; [x_{n-4} ... x_n]) \quad (3)$$

Testing set contains from scalar samples that represent first value that follow appropriate training sample:

$$X_* = (x_6;\ x_7; \ldots ;\ x_n) \quad (4)$$

Analogously in case of the training and testing set for y coordinate, Y and Y_*. Given this data set, corresponding mean functions of Gaussian models have per five, and covariance functions per two free parameters, that are determined in the process of hyperparameters optimization.

Predictive models are defined using training and testing set, obtained hyperparameters and selection of appropriate inference method. Models are formed for x and y trajectories of hand joins, since the gestures are performed in x-y plane. Observations of only one coordinate wouldn't be enough, given that these maybe don't include information necessary for description of all gestures. Values of z-coordinate in this case didn't give any contribution to final result, therefore they are not taken into account.

Determining of predictive values starts from the beginning of sequence and during prediction progress training set increases. Error of prediction in form of difference between real (x, y) and predicted values ($\hat{x}, \hat{y}$) and Mahalanobis distance (5) are calculated in each step. When Mahalanobis distance for several successive moments increases significantly, those small intervals are marked as event, ie potential start or end of gesture. After detecting this discontinuity, training set resets and starts to form again from first next sample onwards. Prediction process continues at the same way until end of gesture sequence.

$$MD = \sqrt{\begin{bmatrix} x-\hat{x} & y-\hat{y} \end{bmatrix} \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix}^{-1} \begin{bmatrix} x-\hat{x} \\ y-\hat{y} \end{bmatrix}} \quad (5)$$

where σ_x and σ_y are predictive variances for first and second Gaussian predictive model, respectively.

3 Results

One of analysed gesture sequences (Fig. 3) contains from three different gestures (Fig. 2), each of them performed twice. Trajectories of x and y coordinate of hand wrist for this gest sequence are shown on the Fig. 3 (top) together with predicted values. Variations of Mahalanobis distance through the time are also shown on Fig. 3 (bottom). Red points on this graph indicate an event. It can be seen that Mahalanobis distance increase every time that one of coordinates suddenly drops or increases value and form peaks on these small intervals. Note that red points are positioned at the beginning or end of the gestures.

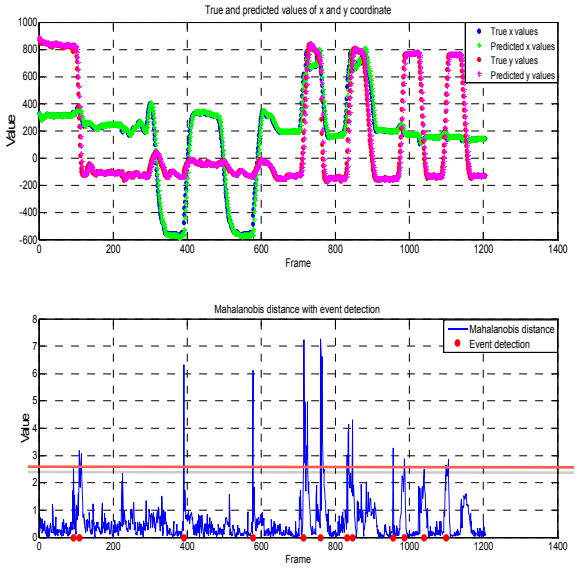

Figure 3: True and predicted values (top) and Mahalanobis distance with event detection (bottom)

4 Conclusion and future work

In this preliminary study we have analysed one approach for gesture segmentation based on predictive Gaussian model. This approach has shown very good performance according to the criteria of model complexity and time necessary for algorithm processing. For few analysed gesture sequences until now, our method has shown excellent results in the sense of correct detection of significant changes in gesture performing. Based on these results, gesture segmentation can be performed directly.

Future work will be oriented to improvement of this method and generalization in case of larger and more diverse gesture sequences.

References

[1] C. Cruz-Neira, D.J. Sandin, T.A. DeFanti, R.V. Kenyon, and J.C. Hart, "The Cave: Audio Visual Experience Automatic Virtual Environment," Comm. ACM, vol. 35, no. 6, pp. 64-72, 1992.

[2] T. Starner, B. Leibe, D. Minnen, T. Westyn, A. Hurst, and J. Weeks, "The Perceptive Workbench: Computervision-Based Gesture Tracking, Object Tracking, and 3d Reconstruction of Augmented Desks," Machine Vision and Applications, vol. 14, pp. 59-71, 2003.

[3] Alana Da Gama, Thiago Chaves, Lucas Figueiredo, Veronica Teichrieb, "Guidance and Movement Correction Based on Therapeutics Movements for Motor Rehabilitation Support Systems", 14th Symposium on Virtual and Augmented Reality, 2012.

[4] S.-W. Lee, "Automatic Gesture Recognition for Intelligent Human-Robot Interaction," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 645-650, 2006.

[5] H.S. Park, D.J. Jung, and H.J. Kim, "Vision-Based Game Interface Using Human Gesture," Advances in Image and Video Technology, pp. 662-671, Springer, 2006.

[6] Doo Young Kwon, PhD Thesis: "A Design Framework for 3D Spatial Gesture Interfaces", ETH Zurich, 2008.

[7] Kanav Kahol, Priyamvada Tripathi, Sethuraman Panchanathan, "Automated Gesture Segmentation from Dance Sequences", 6th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 883 – 888, 2004.

[8] Jus Kocian, Agathe Girard, Blaz Banko, and Roderick Murray-Smith, "Dynamic systems identification with Gaussian processes", Mathematical and Computer Modelling of Dynamical Systems, pp. 411-424, 2004.

[9] Rasmussen, Carl Edward. "Gaussian processes for machine learning." (2006).