MULTI-HMM CLASSIFICATION FOR HAND GESTURE RECOGNITION USING TWO DIFFERING MODALITY SENSORS

Kui Liu, *Chen Chen*, *Roozbeh Jafari*, *and Nasser Kehtarnavaz*
Department of Electrical Engineering, University of Texas at Dallas, Richardson TX, USA
{kxl105200; Chen.Chen12; rjafari; kehtar}@utdallas.edu

***Abstract*—This paper presents a multi-Hidden Markov Model (HMM) classification approach for hand gesture recognition by utilizing two differing modality and low-cost sensors. The sensors consist of a Kinect depth camera and a wearable inertial sensor. It is shown that the multi-HMM classification based on nine signals that are simultaneously captured by these two sensors leads to a more robust recognition compared to the situation when only a single HMM classification is used to generate the likelihood probabilities of hand gestures. This approach is applied to the hand gestures of the $1Unistroke Recognizer application and the results obtained indicate a 7% improvement in the overall classification rate over a single HMM classification under realistic conditions.**

Keywords—Multi-HMM classification; hand gesture recognition; sensor fusion; fusion of inertial and depth sensors

I. INTRODUCTION

The touchless sensing and gesture recognition market is estimated to grow to nearly $15B by 2018 [1]. Hand gesture recognition technology has already appeared in a number of consumer electronics products, e.g. [2]. Hand gesture recognition has been successfully achieved using either a depth camera sensor or an inertial body sensor. In our previous works [3] [4], we showed that the utilization of each of these sensors individually has its own shortcomings. Therefore, we developed a strategy that utilized both of these two differing modality and low-cost sensors simultaneously within a probabilistic recognition framework.

The work presented in this paper is an improvement of our previous solution described in [3] by utilizing a multi-Hidden Markov Model (HMM) classification for recognition of hand gestures. More specifically, a feature-level fusion approach was employed in [3] where the signals from a Kinect depth camera and a wearable inertial sensor were concatenated or stacked as the input to a single HMM classifier. In this work, a decision-level fusion approach is considered by using multiple HMM classifiers each having separate input signals. Then, a pooling step is performed to merge the outcomes of the classifiers. By combining the decisions of multiple HMM classifiers to form a mixture model, an improvement is achieved. Often, the data captured by one type of sensor does not capture all the variations of a hand gesture. However, by using differing modality sensors within a multi-HMM classification framework, a more robust recognition can be reached under realistic conditions.

Multi-HMM classification is not a new concept and has been previously applied to many applications, e.g. text recognition [5], handwriting recognition [6], finger-print recognition [7], speaker recognition [8]. However, this work is the first time such a classification approach is applied to hand gesture recognition based on two differing modality and low-cost sensors. More specifically, the fusion approach introduced in this paper differs from all the previous approaches not only by using a multi-HMM classification but also by using the data simultaneously from both a depth camera (Kinect) and a wearable inertial sensor. Another important aspect of this work is that the computational complexity of the recognition pipeline is kept low leading to its real-time implementation.

The rest of the paper is organized as follows: In section 2, a brief overview of multi-modality sensing is provided. The multi-HMM classification is presented in section 3. The results obtained under realistic conditions are then reported in section 4. This section also includes a comparison with our previous solution involving a single HMM classifier. Finally, the conclusion is stated in section 5.

(a) (b)

Fig. 1 (a) Kinect depth camera [9],
(b) wireless inertial body sensor [11]

II. DIFFERING MODALITY SENSING FOR HAND GESTURE RECOGNITION

The sensors in our recognition system consist of a Kinect depth camera [9] and a wearable inertial sensor as illustrated in Fig. 1. Kinect is a low-cost RGB-Depth sensor developed by Microsoft for human-computer interface applications [10-12]. The introduction of Kinect has led to successful recognition in many applications including video games, virtual reality and gesture recognition. The wearable sensor used here is the low-cost wireless inertial sensor built in the ESSP Laboratory at the

978-1-4799-5924-2/14/$31.00 ©2014 IEEE

University of Texas at Dallas [13]. It consists of: (i) a 9-axis MEMS sensor which captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength data, (ii) a 16-bit low power microcontroller, (iii) a dual mode Bluetooth low energy unit which streams data wirelessly to a laptop/PC, and (iv) a serial interface between the MEMS sensor and the microcontroller enabling control commands and data transmission. For the utilization of the magnetometer, a controlled magnetic field without any distortion is required. Therefore, only the signals consisting of the 3-axis accelerometer and the 3-axis gyroscope are used here and the signals from the magnetometer are not used.

By using both of the sensors, a total of 9 signals are generated: 3 Kinect depth coordinate signals of the hand skeleton location and 6 acceleration and gyro signals from the wireless inertial body sensor worn on a subject's wrist as illustrated in Fig. 3. An example position signal from the Kinect camera and an example acceleration signal from the inertial sensor are shown in Fig. 2. All of the signals are then fed simultaneously into a multi-HMM classifier to recognize hand gestures.

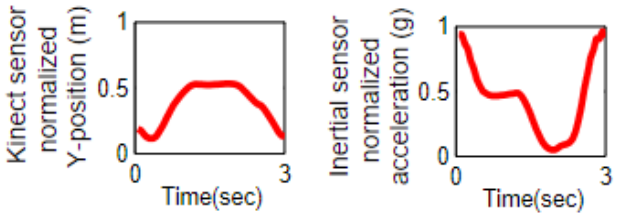

Fig. 2 Example signals from Kinect depth camera (left) and wireless inertial body sensor (right)

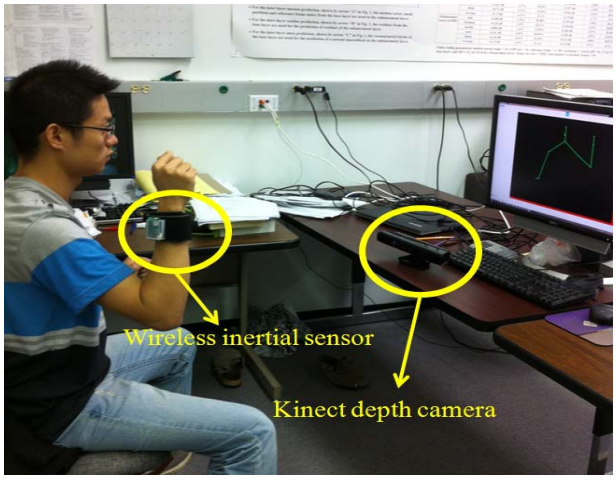

Fig. 3 Setup of two differing modality sensors used for hand gesture recognition

HMM classification is applied to process the signals. An HMM model characterizes a state transfer probability matrix A and an observation symbols probability matrix B. Given an initial state matrix π, an HMM is described by the triplet $\lambda = \{\pi, A, B\}$. In our work, a left-to-right HMM topology is adopted since hand gesture recognition involves temporal signal sequences. Let $O = \{O_1, O_2, \ldots, O_T\}$ be the observation sequence of a hand gesture, where T denotes the number of time samples. The theory of HMM is well established and the details on HMM can be found in many references, for example [14].

In our previous work [4], during operation or testing, the signals were fed into several trained HMMs each corresponding to a particular hand gesture generating a likelihood probability associated with that hand gesture. Whenever none of the probabilities was larger than a high 95% confidence level, the observation sequence was rejected and the gesture was considered to be a not-done-right gesture. If the sequence was not rejected, the gesture with the highest probability was considered to be the recognized gesture. The flowchart of this recognition process is illustrated in Fig. 4.

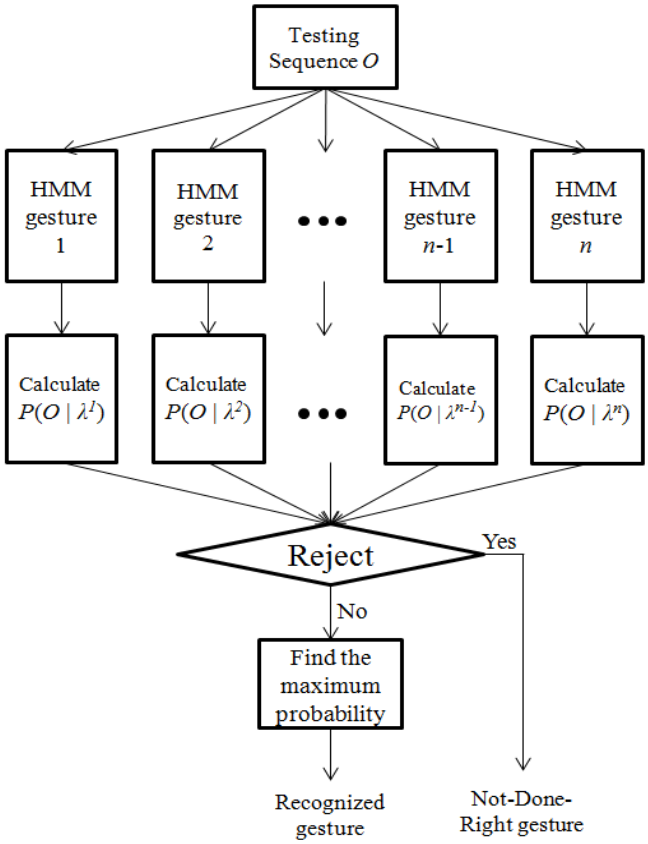

Fig. 4 Flowchart of HMM testing or recognition

III. Multi-HMM Classification

In this section, the multi-HMM classification approach involving the above two differing modality sensors is presented. For the multi-HMM classification, input signals corresponding to the hand coordinates, accelerations, and angles are grouped and fed into three HMM classifiers, each classifier generating its own likelihood probability as shown in Fig. 5. The likelihood probabilities from all the three classifiers are then multiplied by equal weights and are pooled together to generate an overall probability $P(O|\lambda)$ for the input signals.

Hand gesture data consist of 9-dimensional signals (3 dimensions for angular gyros, 3 dimensions for accelerations, and 3 dimensions for Kinect hand skeleton coordinates). The models considered are denoted by $\lambda_{g0} = \{\pi_{g0}, A_{g0}, B_{g0}\}$, $\lambda_{a0} =$

$\{\pi_{a0}, A_{a0}, B_{a0}\}$ and $\lambda_{K0} = \{\pi_{K0}, A_{K0}, B_{K0}\}$ representing gyroscope, accelerometer and Kinect HMM models, respectively. The parameters of these models are then estimated according to the Baum-Welch algorithm [12]. During operation or testing, $P(O|\lambda)$ is computed based on the three likelihood probabilities $P(O|\lambda_g)$, $P(O|\lambda_a)$ and $P(O|\lambda_K)$. Although it is possible to apply different weights to different signals, the same weight is considered for all the signals in this study. The gesture with the maximum average of the three likelihood probabilities is then considered to be the recognized gesture. As a result of using a multi-HMM in this manner, the difference of the probability likelihoods gets diminished or the discriminatory power gets increased.

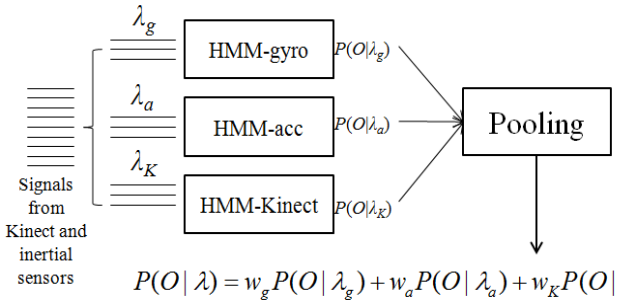

Fig. 5 Framework of the multi-HMM classification

IV. Recognition Results and Discussion

Experimentations were carried out to compare the performance when using the HMM classification in [4] and the multi-HMM classification in this paper. The code is written in C running in real-time on a PC platform with a quad core 1.7GHz processor and 4 GB memory. The input signals were captured with a Kinect camera and the inertial sensor mentioned in section 2. The inertial sensor was placed and tied to a subject's wrist.

We considered the ten single hand gestures in the \$1Unistroke Recognizer application [15]. The gestures in this application are used to manage and navigate Opera Web Browser [16]. These gestures are illustrated in Fig. 6 with the beginning of a gesture indicated by a solid dot.

Ten subjects were asked to perform the ten gestures 30 times with different speeds in different backgrounds and lighting conditions. Our experimentations revealed that 8-12 HMM states were able to capture the transitions of the actions examined. Therefore, 8 HMM states were considered for both the single HMM and multi-HMM classification for computational efficiency purposes. The 3-axis accelerometer and the 3-axis gyroscope signals from the wireless inertial sensor and the 3-axis {X, Y, Z} coordinates signals from the Kinect camera were captured in real-time and simultaneously to form the observation sequence O = {O1, O2, …, OT}. For the training of the original single HMM classifier, the 9-dimensional signals were used to train one HMM classifier for each hand gesture. While in the multi-HMM approach, three HMM classifiers were trained, where each classifier was trained for 3-dimensional signals. We repeated the recognition process 10 times, each time choosing a different set of 9 training subjects. The recognition rates obtained were then averaged to remove any bias to any particular subject. For the "Not-done-right (N)" gesture category, 100 gestures were performed with 50 of them done in an incomplete way and with the other 50 done totally differently which included gestures done backwards.

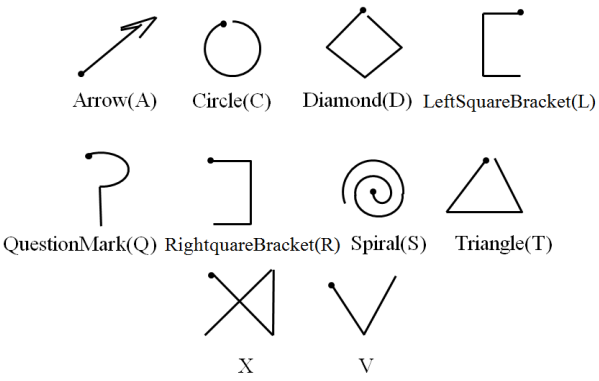

Fig. 6 Hand gestures in the \$1 Gesture Recognizer application

The recognition rates obtained are shown in the form of confusion matrices in Tables I and II for the ten studied gestures of "Arrow (A)", "Circle (C)", "Diamond (D)", "Left square bracket (L)", "Question mark (Q)", "Right square bracket (R)", "Spiral (S)", "Triangle (T)", "X", "V" and the additional class of "Not-done-right (N) ". The capital letters in the tables represent the corresponding initials of the gestures. Table I corresponds to the situation when using the original HMM training and testing approach, and Table II when using the multiple HMM training and testing approach. As can be seen from Table I, misclassifications occurred among these gestures: "Circle", "Diamond", "Question mark", "Right square bracket" and "Triangle". These were due to the variance of the likelihood probabilities not being discriminatory enough to distinguish these gestures from each other. From Table II, it is seen that the multi-HMM classification led to lower misclassifications among these gestures leading to a higher overall recognition rate of 91% under realistic operating conditions compared to the overall recognition rate of 84% for the original single HMM classification. Basically, this increase in the overall recognition rate was due to the enhanced discriminatory power of using the multi-HMM classification, in particular for situations involving unreliable signals from the inertial sensor or Kinect camera. At the end, it is worth stating that our entire recognition pipeline runs in real-time at 27 frames per second. A video clip of the developed system in action can be viewed at http://www.utdallas.edu/~kehtar/multiHMMdemo.wmv.

V. Conclusion

In this paper, two differing modality and low-cost sensors consisting of a Kinect depth camera and a wearable inertial sensor were used simultaneously as part of a multi-HMM classification framework to recognize hand gestures. It was shown that for the ten hand gestures in the \$1Unistroke Recognizer application, an overall recognition rate of 91% was obtained under realistic conditions which included different backgrounds and lighting conditions as well as different hand speeds while the overall recognition rate of a single HMM

classification was only 84%. The recognition framework in this work based on the utilization of two differing modality sensors is general purpose in the sense that it is applicable to other applications such as recognition of human body movements.

TABLE I. HAND GESTURE RECOGNITION RATES (%) CONFUSION MATRIX WHEN USING THE SINGLE HMM CLASSIFICATION

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	2	6	7	0
C	0	82	5	0	0	0	1	10	0	0	2
D	0	8	79	1	0	0	2	7	0	0	3
L	0	7	0	90	0	0	1	2	0	0	0
Q	0	0	0	0	82	14	1	0	2	0	1
R	1	0	0	0	13	81	1	0	2	0	2
S	0	10	4	0	0	0	84	2	0	0	0
T	0	3	12	0	0	0	1	82	1	0	1
X	0	2	0	0	0	0	0	2	87	7	2
V	7	0	1	0	0	0	0	0	5	86	1
N	0	2	7	2	0	0	0	3	2	0	84

TABLE II. HAND GESTURE RECOGNITION RATES (%) CONFUSION MATRIX WHEN USING THE MULTI-HMM CLASSIFICATION

	A	C	D	L	Q	R	S	T	X	V	N
A	88	0	0	0	0	0	0	1	4	7	0
C	0	90	5	0	0	0	2	2	0	0	1
D	0	4	86	0	0	0	1	3	2	0	4
L	0	5	0	90	0	1	1	2	0	0	1
Q	0	0	0	0	91	6	0	0	1	0	2
R	0	0	0	0	7	92	0	0	1	0	0
S	0	5	1	0	0	0	93	1	0	0	0
T	0	4	3	0	0	0	1	90	0	0	2
X	0	1	0	0	0	0	0	3	91	5	0
V	2	0	0	0	0	0	0	0	3	95	0
N	0	2	1	0	1	0	0	2	3	1	90

REFERENCES

[1] M&M, "Gesture Recognition & Touchless Sensing Market," http://www.marketsandmarkets.com/Market-Reports/touchless-sensing-gesturing-market-369.html, 2013.

[2] S. Kim and T. Kim, "Hand gesture recognition input system and method for a mobile phone," U.S. Patent No. 8064704, 2011.

[3] K. Liu, C. Chen, R. Jafari and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," *IEEE Sensors Journal*, vol. 14, no. 6, pp. 1898-1903, June 2014.

[4] K. Liu and N. Kehtarnavaz, "Comparison of two real-time hand gesture recognition systems involving stereo cameras, depth camera, and inertial sensor," *Proceedings of SPIE Conference on Real-Time Image and Video Processing,* paper no. 91390C, Brussels, Belgium, April 2014.

[5] C. Liu and H. Fujisawa, "Classification and learning methods for character recognition: advances and remaining problems," *Machine Learning in Document Analysis and Recognition Studies in Computational Intelligence*, vol. 90, pp. 139-161, 2008.

[6] S. Connell, "Online handwriting recognition using multiple pattern class models," PhD Dissertation, Department of Computer Science, Michigan State University, May 2000.

[7] Y. Yao, P. Frasconi and M. Pontil, "Fingerprint classification with combinations of support vector machines," *Audio- and video-based biometric person authentication lectures notes in computer science*, vol. 2091, pp. 253–258, 2001.

[8] N. Karam and W. Campbell, "A multiple-class MLLR kernel for SVM speaker recognition," *Proceedings of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pp. 4117–4120, Las Vegas, NV, Mar. 2008.

[9] Microsoft Kinect, http://www.microsoft.com/en-us/kinectforwindows/, 2013.

[10] C. Chen, K. Liu and N. Kehtarnavaz, "Real-time human action recognition based depth motion maps," *Journal of Real-Time Image Processing*, August 2013, doi: 10.1007/s11554-013-0370-1, print to appear in 2014.

[11] C. Chen**,** K. Liu, R. Jafari, and N. Kehtarnavaz, "Home-based Senior Fitness Test Measurement System Using Collaborative Inertial and Depth Sensors," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'14)*, Chicago, IL, August 2014, pp. 4135-4138.

[12] C. Chen, N. Kehtarnavaz, and R. Jafari, "A Medication Adherence Monitoring System for Pill Bottles Based on a Wearable Inertial Sensor," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'14)*, Chicago, IL, August 2014, pp. 4983-4986.

[13] M. Bidmeshki and R. Jafari, "Low power programmable architecture for periodic activity monitoring," *Proceedings of ACM/IEEE Int. Conf. on Cyber-Physical Systems*, pp.81-88, Philadelphia, PA, April 2013.

[14] L. Rabiner, "A tutorial on hidden Markov model and selected application in speech recognition," *Proceedings of IEEE*, vol.77, no.2, pp. 257-286, Feb.1989.

[15] J. Wobbrock, A. Wilson, and Y. Li, "Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes," *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 159-168, Newport, RI, Oct. 2007.

[16] Opera Mediaworks, http://www.opera.com/, 2013.