2014-05

Comparison of Two Real-Time Hand Gesture Recognition Systems Involving Stereo Cameras, Depth Camera, and Inertial Sensor

UTD AUTHOR(S): Kui Liu and Nasser Kehtarnavaz

Comparison of two real-time hand gesture recognition systems involving stereo cameras, depth camera, and inertial sensor

Kui Liu[a], Nasser Kehtarnavaz[a], Matthias Carlsohn[b]
[a]University of Texas at Dallas, USA
[b]Engineering and Consultancy for Computer Vision and Image Communication, Germany

ABSTRACT

This paper presents a comparison of two real-time hand gesture recognition systems. One system utilizes a binocular stereo camera set-up while the other system utilizes a combination of a depth camera and an inertial sensor. The latter system is a dual-modality system as it utilizes two different types of sensors. These systems have been previously developed in the Signal and Image Processing Laboratory at the University of Texas at Dallas and the details of the algorithms deployed in these systems are reported in previous papers. In this paper, a comparison is carried out between these two real-time systems in order to examine which system performs better for the same set of hand gestures under realistic conditions.

Keywords: Comparison of real-time hand gesture recognition systems, stereo image system for hand gesture recognition, dual-modality sensor fusion system for hand gesture recognition

1. INTRODUCTION

The use of hand gesture recognition in various human-computer interaction applications has been growing. The literature includes many approaches to achieve hand gesture recognition, some of these approaches are vision-based utilizing video or depth cameras, e.g. [1] [2], and some of them are non-vision-based utilizing inertial body sensors, e.g. [3] [4].

In [5], we introduced an approach by utilizing the disparity information from both the left and right images of a stereo camera system in a complementary manner to achieve a more robust hand gesture recognition compared to the situation when only a single image was used. The extensive testing performed in various backgrounds and lighting conditions indicated that stereo images can be used to complement noisy or missing information in single images. The system developed in [5] was designed in such a way that the merging or fusion of image information from a pair of stereo images was accomplished in real-time.

In [6], we introduced an alternative dual-modality sensor fusion system for the purpose of hand gesture recognition. This system involved the utilization of a depth camera (Kinect) and a wireless inertial body sensor. Noting that the signals captured from each of these sensors have limitations under realistic conditions, it was shown that the fusion of the signals from these two sensors of different modalities led to higher recognition rates compared to the situations when each sensor was used individually on its own. Similar to the system in [5], this system was designed in such a way that the fusion of sensor data and thus recognition was done in real-time.

Real-Time Image and Video Processing 2014, edited by Nasser Kehtarnavaz, Matthias F. Carlsohn, Proc. of SPIE Vol. 9139, 91390C · © 2014 SPIE
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2051342

At this point, it is worth stating that there are many different types of sensors and thus one can design many different sensor fusion systems for hand gesture recognition. The primary reason that we have considered the above two fusion systems is that both of these systems are cost-effective and can be easily deployed as there exist readily available off-the-shelf inexpensive stereo webcams as well as depth and inertial sensors.

The objective of this paper is to provide a comparison of these two real-time fusion systems for the purpose of performing hand gesture recognition. The rest of the paper is organized as follows. Section 2 provides an overview of the two fusion systems that were previously developed at the University of Texas at Dallas. The comparison of the two systems is then covered in section 3 followed by the conclusion in section 4.

2. OVERVIEW OF REAL-TIME SENSOR FUSION SYSTEMS

2.1 Stereo sensor fusion

The previously developed stereo sensor fusion system consists of four main components: online color calibration of hand color, color-based hand detection, hand tracking, and finally hand gesture recognition. The online color calibration component allows subsequent color processing to be adapted to the color characteristics of the light source under which images are captured. This online color calibration has been shown to be an effective approach to cope with unknown color characteristics of various light sources encountered in practice [7]. The calibration is done only once at the beginning when the system is turned on. It involves building a Gaussian Mixture Model (GMM) in the CrCb color space to represent the color characteristics of the hand being captured in an online manner. The calibration is performed by the user placing his or her hand in a box displayed at the image center, see Figure 1. Representative skin color pixels are collected within this box using a k-means clustering algorithm separating skin pixels from non-skin pixels via two clusters. A GMM model is then trained and used for a region growing hand color segmentation within a region-of-interest specified by a tracking module.

(a) (b)

Figure 1. Online color calibration: (a) left camera calibration box, (b) right camera calibration box

The tracking module consists of the CamShift algorithm [8], where the hue component of color is used for tracking. The histogram within a window is used to serve as the hand tracking feature together with a searching window. The center of the window provides the seed point for the so called flood fill region growing operation [9] to achieve the color segmentation in a computationally efficient manner. The color segmented areas from the left and right images are merged by aligning the left and right images. A number of merging rules are then considered to establish consistency of

information between the left and the right images. For example, since the hand motion is continuous, a large mask area difference between consecutive frames denotes an inconsistency which is utilized to improve the robustness of recognition. For recognition, the Dynamic Time Warping (DTW) method is used as this method allows obtaining a distance measure between an unknown hand gesture signal and a set of reference hand gesture signals in a computationally efficient manner while coping with different speeds of hand gestures. The flow chart of all the components involved in the previously developed stereo fusion system appears in Figure 2 with the details presented in [5].

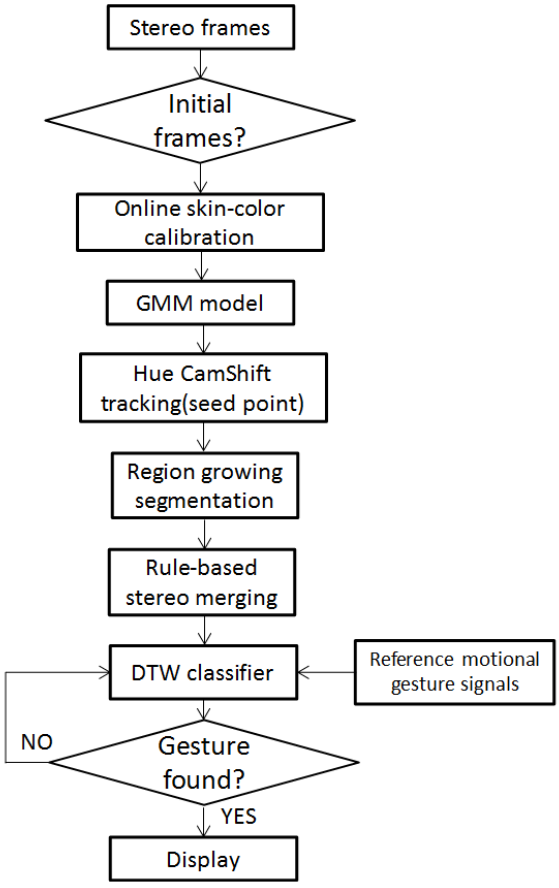


Figure 2. Flowchart of previously developed stereo fusion system for hand gesture recognition

2.2 Dual-modality sensor fusion

In this fusion system, the three coordinate signals from a Kinect depth camera corresponding to the hand centroid and the six acceleration and angle signals from an inertial body sensor worn on a subject's wrist are used to achieve hand gesture recognition. An example of a position signal from the Kinect camera and an example of an acceleration signal from the inertial sensor used are shown in Figure 3. These signals are then fed simultaneously into a Hidden Markov Model (HMM) classifier [10] to identify hand gestures.

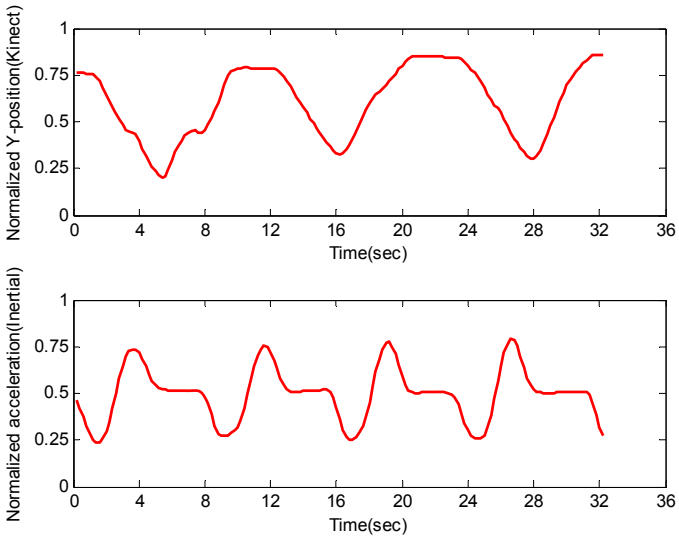


Figure 3. Example signals from Kinect camera and inertial sensor

Given an initial state matrix π, an HMM is represented by the triplet $\lambda = \{\pi, A, B\}$, where A denotes a states probability matrix and B denotes an observation transition probability matrix. Suppose a random sequence $O = \{O_1, O_2, \ldots, O_T\}$ is observed; let $V = \{v_1, v_2, \ldots, v_T\}$ denote all possible outcomes and $S = \{S_1, S_2, \ldots, S_M\}$ denote all HMM states with q_t representing the state at time t, where T indicates the number of time samples. Then, the HMM probability matrices are given by

$$\pi = \{p_i = P(Q_1 = S_i)\},\ 1 \le i \le M; \tag{1}$$

$$A = \{a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)\},\ 1 \le i, j \le M; \tag{2}$$

$$B = \{b_j(k) = P(O_t = v_k \mid q_t = S_j)\},\ 1 \le j \le M,\ 1 \le k \le T; \tag{3}$$

$$\text{where} \sum_{i=1}^{M} \pi_i = 1,\ \sum_{j=1}^{M} a_{ij} = 1 \ \text{ and } \sum_{k=1}^{T} b_j(k) = 1 \tag{4}$$

For training, the probability of the observation sequence O is obtained by $P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda)$. The Baum-Welch algorithm [10] is used to compute the initial probability $P(O|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} a_{q_3 q_4} \cdots a_{q_{T-1} q_T}$ towards updating λ. Noting that $P(O, Q|\lambda)=P(O|Q, \lambda)P(Q, \lambda)$, the following probability is then computed:

$$P(O|\lambda) = \sum_Q P(O|Q,\lambda)\mathrm{P}(Q,\lambda) = \sum_{q_1,q_2,\ldots,q_T} \pi_{q_1}\, b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{5}$$

For testing, a hand gesture signal sequence is fed into n trained HMM models, each corresponding to a hand gesture in order to calculate the probabilities. A high confidence interval (95%) is applied to the probabilities to classify the sequence. Let μ and σ represent the mean and variance of the probabilities. For the 95% confidence interval, if none of the n probabilities is larger than $\mu + 1.96\frac{\sigma}{\sqrt{n}}$, the gesture sequence is rejected and the gesture is considered to be a Not-Done-Right gesture. If it is not rejected, the gesture with the maximum probability is considered to be the recognized gesture. The testing of the HMM sensor fusion is illustrated in Figure 4.

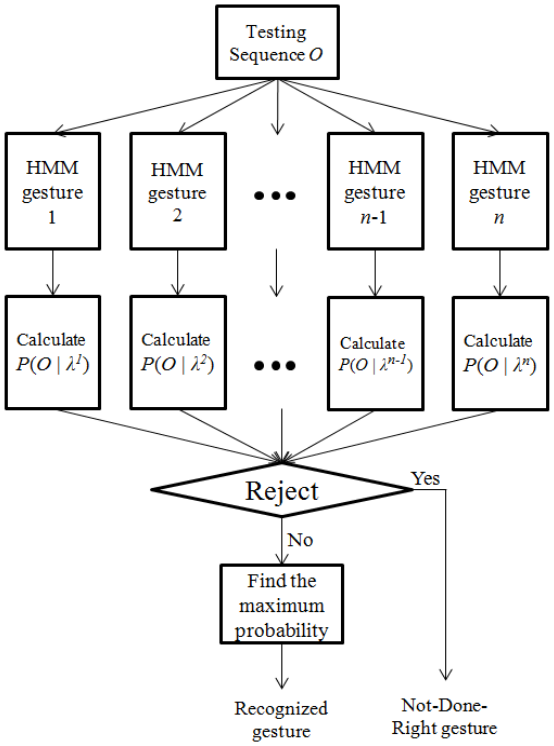
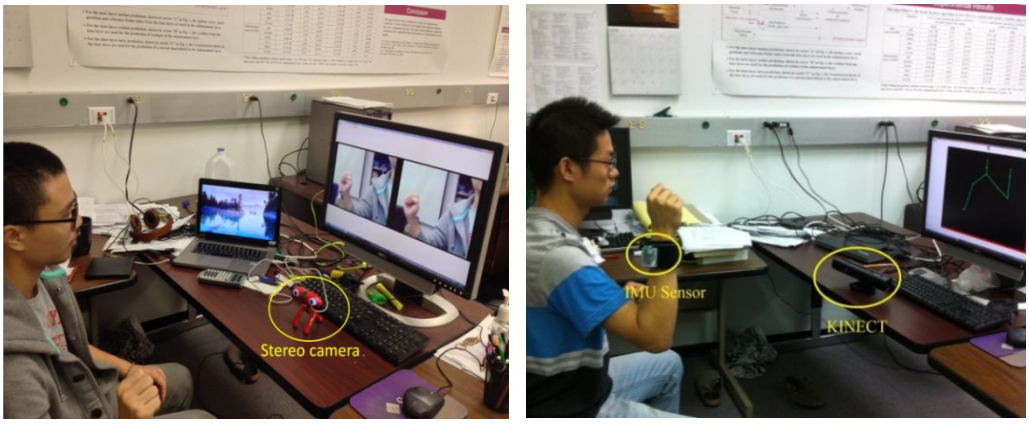


Figure 4. Flowchart of HMM testing or recognition of dual-modality sensor fusion

3. COMPARISON OF RECOGNITION RATES AND REAL-TIME PROCESSING

The codes for the above fusion systems are written in C. Both types of fusion systems run in real-time on a PC platform having a quad core 1.7GHz processor and 4G of memory. The input images in the stereo camera system are captured by the stereo webcam Novo Minoru, which is an inexpensive stereo webcam generating low resolution images of size 640*480, see Figure 5(a). The input data in the dual-modality fusion system are captured by a Microsoft Kinect camera and a wireless inertial sensor developed in the ESSP Laboratory at the University of Texas at Dallas [11]. The wireless inertial sensor is tied to a subject's wrist, see Figure 5(b).



(a) (b)

Figure 5. (a) Stereo camera system (b) dual-modality system (Kinect + inertial sensor)

For comparison of the systems, two gesture sets were considered. One gesture set consisted of the single hand gestures in the Microsoft Action Dataset [12] and the other gesture set consisted of the $1 Gesture Recognizer Dataset [13]. There are 5 single hand gestures in the Microsoft Action Dataset and 15 single hand gestures in the $1 Gesture Recognizer Dataset. The gestures in the Microsoft Action Dataset are illustrated in Figure 6. The hand gestures in the $1 Gesture Recognizer Dataset are used to manage and navigate an Opera Web Browser [14]. These gestures are illustrated in Figure 7 with the beginning of a gesture indicated by a solid dot.

Figure 6. Single hand gestures in the Microsoft Action Dataset:

"wave", "hammer", "punch", "drawX", "circle"

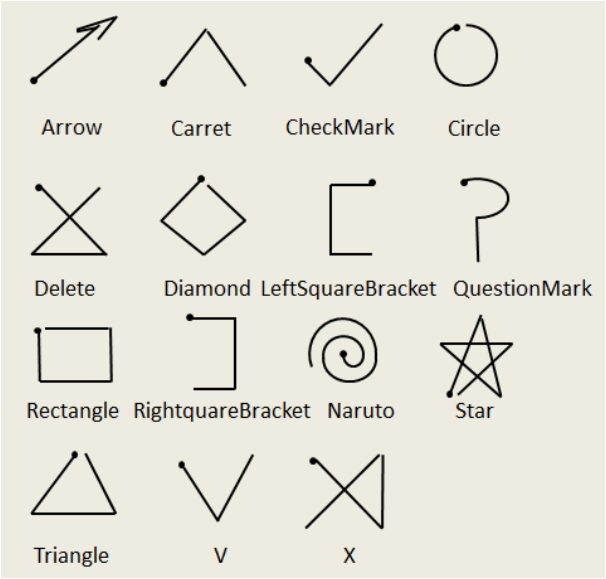


Figure 7. Single hand gestures in the $1 Gesture Recognizer Dataset

Ten subjects were asked to perform each gesture in the two gesture sets 30 times with different speeds in front of different backgrounds. For the stereo camera fusion system, the DTW distance indicated how well an unknown signal matched a number of template or reference signals. A template or reference signal of a hand gesture was set up by taking the average of the training sample signals. The signals consisted of the 3-axis {*X*, *Y*, *Z*} gradient signals after merging the information from the stereo images. For the dual-modality sensor fusion, 8-12 HMM states were used. The 3-axis accelerometer and the 3-axis gyroscope signals from the wireless inertial sensor and the 3-axis {*X*, *Y*, *Z*} gradient or position difference signals from the Kinect camera were captured simultaneously in a synchronized manner to form the observation sequence $O = \{O_1, O_2, \ldots, O_T\}$ of the HMM classifiers. The recognition process was repeated 10 times. Each time we chose a different set of 9 training subjects. The recognition rates obtained were averaged to remove any bias to a particular subject. In addition to performing the hand gestures correctly, incorrect gestures were also performed such that half of them were the same gestures but done in an incomplete way and the other half were random gestures. The incorrectly performed gestures is named "Negative" here while the correctly performed gestures named "Positive".

Table 1 shows the performance outcome of the stereo fusion system and Table 2 shows the performance outcome of the dual-modality fusion system based on the Microsoft Action gesture set. The experiments included a positive database containing a total of 5*30=150 correctly done gestures and a negative database containing 50 incorrectly done gestures, named "Not-done-right (N)", with 25 gestures done in incomplete ways and 25 done by random hand movements. In these tables, PPV (positive predictive value) and NPV (negative predict value) indicate the recognition rates of the correctly and incorrectly done hand gestures, respectively; TP, FP, TN and FN denote true positive, false positive, true negative and false negative, respectively. Table 3 and Table 4 summarize the performance outcomes of the two fusion systems for the $1 Recognizer gesture set. For this gesture set, the experiments included a positive database containing 15*30=450 correctly done gestures and a negative database containing 100 incorrectly done gestures.

As an alternative way to show the recognition outcomes of the two fusion systems, Tables 4 through 8 provide the confusion matrices obtained. In these tables, the following abbreviations are used for the hand gestures: "Arrow (A)", "Carret (Ca)", "Check mark (Ch)", "Circle (Ci)", "Delete (De)", "Diamond (Di)", "Left square bracket (L)", "Question mark (Q)", "Rectangle (Re)", "Right square bracket (Ri)", "Spiral (Sp)", "Star (St)", "Triangle (T)", "X", "V" and "Not-done-right (N)".

Table 1. Recognition performance of the stereo fusion system for the hand gestures in the Microsoft Action Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	123	27	150	PPV=0.82
Negative	42	8	50	NPV=0.83

*PPV(Positive Predictive Value)=TP/(TP+FP) NPV(Negative Predictive Value)=TN/(TN+FN)

Table 2. Recognition performance of the dual-modality fusion system for the hand gestures in the Microsoft Action Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	138	12	150	PPV=0.92
Negative	48	2	50	NPV=0.97

Table 3. Recognition performance of the stereo fusion system for the hand gestures in the $1 Recognizer Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	338	112	450	PPV=0.75
Negative	73	27	100	NPV=0.73

Table 4. Recognition performance of the dual-modality fusion system for the hand gestures in the $1 Recognizer Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	407	43	450	PPV=0.90
Negative	91	9	100	NPV=0.91

Table 5. Confusion matrix (% recognition rates) for the hand gestures in the Microsoft Action Dataset when using the stereo fusion system

	wave	hammer	punch	drawX	circle	N
wave	80	3	3	5	8	1
hammer	2	85	3	5	2	3
punch	4	6	83	2	3	2
drawX	5	4	3	76	7	5
circle	6	2	1	4	86	1
N	1	1	5	9	1	83

Table 6. Confusion matrix (% recognition rates) for the hand gestures in the Microsoft Action Dataset when using the dual-modality fusion system

	wave	hammer	punch	drawX	circle	N
wave	92	1	1	5	1	0
hammer	5	91	2	2	0	0
punch	3	5	91	0	0	1
drawX	0	0	6	88	6	0
circle	1	0	0	0	99	0
N	0	1	1	1	0	97

Table 7. Confusion matrix (% recognition rates) for the hand gestures in the $1 Recognizer Dataset when using the stereo fusion system

	A	Ca	Ch	Ci	De	Di	L	Q	Re	Ri	Sp	St	T	V	X	N
A	68	12	8	0	1	5	0	1	1	1	1	0	1	0	0	1
Ca	0	74	0	0	1	0	0	0	12	10	0	0	0	1	1	1
Ch	6	0	66	2	2	1	1	0	0	0	0	2	1	13	3	3
Ci	0	0	0	79	0	0	7	0	5	0	6	0	3	0	0	0
De	0	0	0	0	82	1	0	0	0	4	0	0	5	0	6	2
Di	0	0	0	11	1	70	4	0	6	0	4	0	3	0	0	1
L	0	0	0	10	1	5	77	0	0	3	0	0	2	0	0	2
Q	6	0	1	0	1	0	0	79	0	6	0	1	0	0	2	4
Re	0	0	0	5	0	9	1	0	72	0	0	0	8	0	5	0
Ri	1	1	0	2	2	0	4	6	0	76	0	0	0	0	5	3
Sp	0	1	1	11	0	3	0	0	0	3	79	0	0	0	0	2
St	2	0	0	3	3	7	0	0	0	0	2	73	5	0	3	2
T	0	0	0	2	2	6	0	2	5	0	0	0	78	0	4	1
V	1	2	12	3	6	0	0	0	0	0	0	0	2	73	0	1
X	0	1	1	3	7	0	0	0	0	3	0	0	0	0	83	2
N	1	2	2	1	0	1	5	3	1	4	2	1	1	3	0	73

As can be seen from these tables, the dual-modality fusion system outperformed the stereo fusion system. On average, the dual-modality system provided 12% higher recognition rate for the Microsoft Action Dataset and 16% higher recognition rate for the $1 Recognizer Dataset compared to the stereo system. This is attributed to the fact that the sensors in the dual-modality fusion system are of two different modalities capturing different attributes or features of a hand gesture while in the stereo fusion system, both the left and right video images used have the same vision-based modality.

Table 9 provides the comparison between the two real-time systems in terms of frame rates and computational complexity. As can be seen from this table, the frame rates of the two systems were comparable with the computational complexity of the stereo system being slightly higher than the dual-modality system. The computational complexity of the stereo system is $O(\alpha m^2+L^2)$ where α denotes the number of mean shift iterations, m^2 image resolution, L the length of the warping path in the DTW algorithm.

Table 8. Confusion matrix (% recognition rates) for the hand gestures in the $1 Recognizer Dataset when using the dual-modality fusion system

	A	Ca	Ch	Ci	De	Di	L	Q	Re	Ri	Sp	St	T	V	X	N
A	89	3	0	0	0	0	0	0	0	0	0	0	0	2	3	3
Ca	3	90	2	0	3	0	0	0	0	0	0	0	0	0	0	2
Ch	1	2	91	0	2	0	0	0	0	0	0	0	0	4	0	0
Ci	0	0	0	87	1	3	0	0	4	0	0	0	4	0	0	1
De	0	3	4	0	86	3	0	0	0	0	0	0	0	0	3	1
Di	0	0	0	3	0	89	0	0	3	0	0	0	4	0	0	1
L	0	0	0	0	0	2	95	0	1	1	0	0	0	0	0	1
Q	0	0	0	0	0	0	0	94	0	3	0	0	1	0	1	1
Re	0	0	0	4	0	5	0	0	86	0	0	0	3	0	0	2
Ri	0	0	0	0	2	0	5	2	0	87	0	0	0	0	3	1
Sp	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	2
St	2	0	0	0	0	0	0	0	1	0	0	92	2	0	0	3
T	0	0	0	3	0	6	0	0	1	0	0	0	88	0	0	2
V	0	1	7	0	2	0	0	0	0	0	0	0	0	89	0	1
X	0	0	0	3	1	4	0	0	0	0	0	0	0	1	90	1
N	2	0	0	0	0	0	1	1	0	2	0	2	0	0	1	91

Table 9. Frame rates and computational complexity for the stereo fusion system versus the dual-modality fusion system

Hand gesture recognition system	Frame rates per sec	Computational complexity
Stereo Fusion	24±1.6	$O(\alpha m^2 + L^2)$
Dual-Modality Fusion	27±3.0	$O(RNS)$

Since the parameters of the HMM model are pre-trained, the computational complexity of the dual-modality system is basically the same as the complexity of the HMM model testing which is given by $O(RNS)$, where R denotes the number of operations to compute an observation likelihood, N the number of states in HMM, and S the number of observations [15]. It is worth pointing out that the skeleton tracking is done by a dedicated processor as part of the Kinect depth sensor [16]. As a result, the skeleton is retrieved in real-time and the skeleton image resolution has little influence on the computational complexity. Example video clips of the two systems operating in real-time can be seen at the websites [17] and [18].

4. CONCLUSION

In this paper, the two previously developed stereo and dual-modality sensor fusion approaches for hand gesture recognition were compared. It was shown that fusing or merging the data from the dual-modality sensors led to an overall recognition rate of 94% for the five single hand gestures in the Microsoft Action Dataset and 90% for the 15 single hand gestures in the $1 Recognizer Dataset under realistic conditions such as different gesture speeds and backgrounds. These recognition rates were on average 12% and 16%, respectively, higher than the rates for the stereo fusion system.

REFERENCES

[1] Keskin, C., Kirac, F., Kara, Y., and Akarun, L., "Real time hand pose estimation using depth sensors," *Proceedings of IEEE International Conference on Computer Vision Workshops*, 1228-1234 (2011).

[2] Gorce, D., Fleet, D., and Paragios, N., "Model-based 3D hand pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9), 1793-1805 (2011).

[3] Wang, L., Gu, T., Chen, H., Tao, X., and Lu, J., "Real-time activity recognition in wireless body sensor networks: From simple gestures to complex activities," *Proceedings of IEEE Int. Conf. on Embedded and Real-Time Computing Systems and Applications*, 43-52 (2010).

[4] Zhu, R., and Zhou, Z., "A real-time articulated human motion tracking using tri-axis inertial/magnetic sensors package," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 12(2), 295-302 (2004).

[5] Liu, K., and Kehtarnavaz, N., "Real-time robust vision-based hand gesture recognition using stereo images," *Journal of Real-Time Image Processing* (2013).

[6] Liu, K., Chen, C., Jafari, R., and Kehtarnavaz, N., "Fusion of Inertial and Depth Sensor Data for Robust Hand Gesture Recognition," to appear in *IEEE Sensors Journal* (2014).

[7] Rahman, M., Kehtarnavaz, N., and Ren, J., "A hybrid face detection approach for real-time deployment on mobile devices," *Proceedings of IEEE International Conference on Image Processing*, 3233-3236 (2009).

[8] Bradski, G., "Computer video face tracking for use in a perceptual user interface," *Intel Technology Journal* (1998).

[9] Heckbert, P., *Graghics Gems IV*, Academic Press, New York (1994).

[10] Rabiner, L., "A tutorial on Hidden Markov Model and selected application in speech recognition," *Proceedings of IEEE*, 77(2), 257-286 (1989).

[11] "http://www.essp.utdallas.edu/"

[12] "http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/"

[13] Li, Y., "Protractor: a fast and accurate gesture recognizer," *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, 2169-2172 (2010).
[14] "http://www.opera.com/"
[15] Johnson, M., "Capacity and complexity of HMM duration modeling techniques," *IEEE Signal Processing Letters*, 12(5), 407-410 (2005).
[16] Wang, R., *Augmented Reality with Kinect*, Packt Publishing, Birmingham, UK (2013).
[17] "http://www.youtube.com/watch?v=jYg7U2UYeZo"
[18] "http://youtu.be/GSQrExl81mo"