An Enhanced Multi-view Human Action Recognition System for Virtual Training Simulator

Beom Kwon, Junghwan Kim and Sanghoon Lee
Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: {hsm260, junghwan.kim, slee}@yonsei.ac.kr, Tel: +82-2-2123-2767

Abstract—**Virtual military training systems have received considerable attention as a possible substitute for conventional real military training. In our previous work, human action recognition system using multiple Kinects (HARS-MK) has been implemented as a prototype of virtual military training simulator. However, the classification accuracy of HARS-MK is not enough to be utilized for virtual military training simulator. In addition, the experiments are carried out under just two simple action types; walking and crouching walking. In order to overcome these limitations, in this paper, we propose an enhanced multi-view human action recognition system (EM-HARS). Compared to HARS-MK, in EM-HARS, feature extractor is enhanced by employing covariance descriptor. In addition, the feasibility test of EM-HARS is conducted under various human actions including military training actions which are newly captured. The experiment results show that EM-HARS achieves higher classification accuracy than that of HARS-MK.**

I. Introduction

Recently, virtual reality technology has gained more attention in various fields including game, military training, education, and rehabilitation [1]–[4], since 3D visual discomfort problems have been solved [5]–[12]. In the case of military training, military industry is focusing on replacing real military training with virtual military training to improve its effect, and to reduce training cost. To achieve this goal, it is required to implement a virtual military training system in which high stability and vivid virtual military training environment have to be provided to users. A famous example of virtual military training system is the dismounted soldier training system (DSTS) in United States army [13]. In DSTS, user has to wear various wearable devices including motion sensors for posture tracking. Although wearable motion sensors provide accurate position values, the main disadvantage of wearable motion sensors is error accumulation over time. In general, since soldiers participate in a military training in a long time, wearable motion sensors-based posture tracking methods which have error accumulation problem are not appropriate for virtual military training system.

For avoiding the error accumulation problem, the authors of [14]–[16] proposed human action recognition methods based on single camera. However, since these methods in [14]–[16] are vulnerable to occlusion, it is difficult to apply these methods to virtual military training system. To overcome the occlusion problem, multi-view human action recognition methods using color and depth data are proposed in [17] and [18]. In addition, the authors of [19] proposed a multi-view human action recognition method using color, depth, and skeleton data. Many researches on human action recognition including [16]–[19] have been studied by using Microsoft Kinect due to its cheap cost and convenience in use. Furthermore, Kinect provides real-time skeleton data, and no markers are necessary to be attached to user. Kinect captures user under the assumption that the user looks the Kinect in the face. Therefore, if user does not look Kinect in the face, the Kinect may provide imprecise skeleton data of the user. In multi-view human action recognition system, it is impossible for user to look all Kinects in the face at the same time. In addition, since the imprecise skeleton data may lead to degradation of classification accuracy, integration of skeletons is one of the most challenging issues in multi-view human action recognition systems.

To address this problem, we have developed a weighted integration method [20]. Skeleton data obtained from each Kinect can be integrated more accurately by using the weighted integration method in which a higher weight value is assigned to skeleton data obtained from the Kinect which faces user. Based on previous study in [20], we have implemented a human action recognition system using multiple Kinects (HARS-MK) [21]. In order to recognize human actions, snapshot and temporal features are extracted from skeleton data sequences. Based on the snapshot and temporal features, the classifier model is trained by using support vector machine (SVM). The average accuracy rate of HARS-MK are about 88.375% and 88.875% for walking and crouch walking, respectively. However, the average accuracy rate is not sufficient to utilize HARS-MK as a virtual military training system. In addition, the number of types of action tested in the experiments is only two.

Therefore, in this paper, we propose an enhanced multi-view human action recognition system (EM-HARS) which is the improved version of HARS-MK. In order to improve the classification accuracy, we enhance temporal feature extractor by employing covariance descriptors at three-dimensional (3D) joint locations [22]. In addition, we newly capture multi-view skeleton data of various human actions related with military training. Using the new dataset of skeleton data sequences, we demonstrate the feasibility of the proposed EM-HARS.

The remainder of this paper is organized as follows. Section II presents the system architecture of EM-HARS. Section III shows the experimental results and discusses the feasibility of EM-HARS as a virtual military training system. Conclusion is given in Section IV.

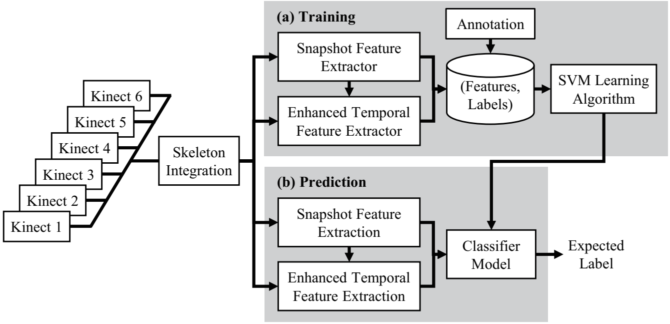

Fig. 1. Block diagram of the proposed EM-HARS.

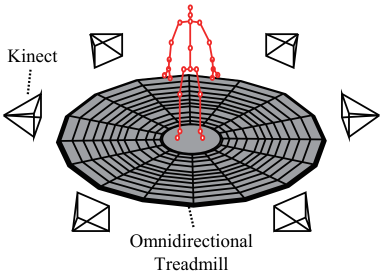

Fig. 2. Kinects configuration in EM-HARS.

II. System Architecture

This section describes the system architecture of EM-HARS. Fig. 1 shows a block diagram of EM-HARS. In EM-HARS, six Kinects, which are arranged in a ring with a radius of 3 m and 60° separated, are used to capture the whole user's body as shown in Fig. 2. Each Kinect provides skeleton data of user in real-time. The skeleton data consists of a set of 25 joints and provides 3D coordinates for each joints. Table I shows the set of joints provided from Kinect.

A. Skeleton Integration

TABLE I
A set of joints provided from Kinect.

Name	Join Index (j)	Name	Joint Index (j)
Spine Base	0	Knee Left	13
Spine Mid	1	Ankle Left	14
Neck	2	Foot Left	15
Head	3	Hip Right	16
Shoulder Left	4	Knee Right	17
Elbow Left	5	Ankle Right	18
Wrist Left	6	Foot Right	19
Hand Left	7	Spine Shoulder	20
Shoulder Right	8	Hand Tip Left	21
Elbow Right	9	Thumb Left	22
Wrist Right	10	Hand Tip Right	23
Hand Right	11	Thumb Right	24
Hip Left	12		

TABLE II
Values of $w_1(\cdot)$ according to tracking states.

Tracking State	Tracked	Not-tracked	Inferred
$w_1(\cdot)$	1	0	0.5

As shown in Fig. 2, since the location of each Kinect differs from each other, the coordinate systems of Kinects differ from each other. Therefore, it is necessary to calibrate them to a world coordinate system. After the calibration is complete, the skeleton data have to be transformed into the world coordinate system. By referring to the calibration method of [23], each Kinect coordinate system is transformed into the world coordinate system by

$$\left[X^{(w)}\ Y^{(w)}\ Z^{(w)}\right]^T = \mathbf{R}^{(i)}\left[X^{(i)}\ Y^{(i)}\ Z^{(i)}\right]^T + \mathbf{t}^{(i)} \tag{1}$$

where $X^{(w)}$, $Y^{(w)}$ and $Z^{(w)}$ are coordinates in the world coordinate system, $\mathbf{R}^{(i)}$ is the rotation matrix of the ith Kinect, $X^{(i)}$, $Y^{(i)}$ and $Z^{(i)}$ are coordinates in the 3D Cartesian coordinate system of the ith Kinect, and $\mathbf{t}^{(i)}$ is the translation matrix of the ith Kinect.

In our previous work [20], we have developed a front vector tracing algorithm for tracking the front of a skeleton, and a weighted integration method for achieving accurate skeleton integration. In this paper, in order to improve performance of multi-view skeleton integration, the front vector tracking algorithm and the weighted integration method are employed in EM-HARS.

Let $\mathbf{P}_j^* = [x_j^*\ y_j^*\ z_j^*]$ be a vector representation of an optimal jth joint position after skeleton integration. Then, $\mathbf{P}_j^*$ is can be obtained by solving an optimization problem as follows:

$$\min_{\mathbf{P}_j} \sum_{i\in K} w_1\left(s_j^{(i)}\right)\cdot w_2\left(d_j^{(i)}\right)\cdot \parallel \mathbf{P}_j - \mathbf{P}_j^{(i)} \parallel \tag{2}$$

where $\mathbf{P}_j = [x_j\ y_j\ z_j]$ is a vector representation of the possible joint positions, K is the set of indices of Kinects, $s_j^{(i)}$ is the tracking state of the jth joint provided from the ith Kinect, $d_j^{(i)}$ is the distance between the jth joint provided from the ith Kinect and the ith Kinect, $w_1(\cdot)$ and $w_2(\cdot)$ are weight functions whose values are determined according to $s_j^{(i)}$ and $d_j^{(i)}$, respectively, $\parallel\cdot\parallel$ indicates the Euclidean distance, and $\mathbf{P}_j^{(i)} = \left[x_j^{(i)}\ y_j^{(i)}\ z_j^{(i)}\right]$ is a vector representation of the jth joint position provided by ith Kinect.

The tracking state $s_j^{(i)}$ provided via Kinect software development kit (SDK) is categorized as *Tracked*, *Not-tracked*, and *Inferred*. If the jth joint position data is obtained accurately, $s_j^{(i)}$ becomes *Tracked*. $s_j^{(i)}$ becomes *Not-tracked* when the jth joint position data cannot be obtained. In addition, if the jth joint is occluded, $s_j^{(i)}$ becomes *Inferred*. Table II shows the values of $w_1(\cdot)$ according to $s_j^{(i)}$.

In general, the accuracy of a joint position data is inversely proportional to the distance between the joint and a Kinect [24]. The authors of [25] measured the noise in the joint position data according to distance from Kinect. The measurement was performed in the range of $1.2-3.5\ m$ from Kinect. In addition, it is shown that the noise can be fitted to the function $0.4946e^{0.7\cdot d_j^{(i)}}$. A minimum (maximum) value of the

TABLE III
DESCRIPTION OF *Angles*.

Kinematic Chain	Angle	Notation
Spine Base - Spine Mid - Spine Shoulder	Yaw	A_1
Spine Shoulder - Shoulder Left - Elbow Left	Roll	A_2
	Yaw	A_3
Spine Shoulder - Shoulder Right - Elbow Right	Roll	A_4
	Yaw	A_5
Shoulder Left - Elbow Left - Wrist Left	Roll	A_6
	Yaw	A_7
Shoulder Right - Elbow Right - Wrist Right	Roll	A_8
	Yaw	A_9
Spine Base - Hip Left - Knee Left	Roll	A_{10}
	Yaw	A_{11}
Spine Base - Hip Right - Knee Right	Roll	A_{12}
	Yaw	A_{13}
Hip Left - Knee Left - Ankle Left	Roll	A_{14}
	Yaw	A_{15}
Hip Right - Knee Right - Ankle Right	Roll	A_{16}
	Yaw	A_{17}

average noise is 1.1456 (5.7315) at 1.2 (3.5) meters. By using these values, the noise is normalized. Then, $w_2(\cdot)$ is defined as follows:

$$w_2(d_j^{(i)}) = 1 - \underbrace{\left(\frac{0.4946e^{0.7\cdot d_j^{(i)}} - 1.1456}{5.7315 - 1.1456}\right)}_{=\text{ normalized noise}}. \quad (3)$$

B. Feature Extraction

In EM-HARS, the feature extraction is done in two stages; snapshot and enhanced temporal feature extractions. The output of the snapshot feature extractor consists of *Joint Velocities*, *Angles*, and *Angular Velocities*.

Joint Velocities are calculated as the difference between current and previous joint position vectors which are captured from two consecutive frames, respectively. Let $\mathbf{V}_j[n]$ be a vector representation of a joint velocity of the jth joint at the nth frame. Then, $\mathbf{V}_j[n]$ can be computed as follows:

$$\mathbf{V}_j[n] = \frac{\mathbf{P}_j[n] - \mathbf{P}_j[n-1]}{\Delta n}, \quad j \in \{1, \cdots, 25\}, \quad (4)$$

where Δn is a time interval between the nth and $(n-1)$th frames. Here, Δn is set to $1/30$ ms because the frame rate of Kinect is 30 frames per second.

Angles mean Tait-Bryan angles (also known as Cardan angles or nautical angles) and are calculated from three consecutive joints (kinematic chain). Let $l \in \{1, \cdots, 17\}$ be a index of angle, and A_l be a value of the lth angle. Table III shows the details of description of *Angles*.

Angular Velocities are computed as the difference between current and previous angles at two consecutive frames. Let $U_l[n]$ be a angular velocity of the lth angle at the nth frame. Then $U_l[n]$ can be computed as follows:

$$U_l[n] = \frac{A_l[n] - A_l[n-1]}{\Delta n}, \quad l \in \{1, \cdots, 17\}. \quad (5)$$

The dimensions of *Joint Velocities*, *Angles*, and *Angular Velocities* are 75, 17, and 17, respectively. In other words, the dimension of snapshot feature vectors is 109.

In the enhanced temporal feature extractor, the snapshot features over M frames are stored in a buffer to capture temporal characteristics of human action. The output of the enhanced temporal feature extractor are *Average of Joint Velocities*, *Average of Angles*, *Average of Angular Velocities*, and *Covariance of Joints*.

Average of Joint Velocities are computed as follows:

$$\overline{\mathbf{V}}_j[n] = \frac{1}{M} \sum_{m=n-(M-1)}^{n} \mathbf{V}_j[m], \quad j \in \{1, \cdots, 25\}. \quad (6)$$

Average of Angles are calculated as follows:

$$\overline{A}_l[n] = \frac{1}{M} \sum_{m=n-(M-1)}^{n} A_l[m], \quad l \in \{1, \cdots, 17\}. \quad (7)$$

Average of Angular Velocities are calculated as follows:

$$\overline{U}_l[n] = \frac{1}{M} \sum_{m=n-(M-1)}^{n} U_l[m], \quad l \in \{1, \cdots, 17\}. \quad (8)$$

Covariance of Joints are calculated as follows:

$$Cov\,(\mathbf{S}[n]) = \frac{1}{M} \sum_{m=n-(M-1)}^{n} [\mathbf{S}[m] - \overline{\mathbf{S}}[n]]^T \cdot [\mathbf{S}[m] - \overline{\mathbf{S}}[n]], \quad (9)$$

where $\mathbf{S}[m] = [\mathbf{P}_1[m]\ \mathbf{P}_2[m] \cdots \mathbf{P}_{25}[m]]$ is a vector expressing the positions of 25 joints at the mth frame, $\overline{\mathbf{S}}[n] = \frac{\mathbf{S}(n-(M-1))+\cdots+\mathbf{S}[n-1]+\mathbf{S}[n]}{M}$ is the sample mean from the $n-(M-1)$th frame to the nth frame, and the superscript T indicates transpose. Since $Cov\,(\mathbf{S}[n])$ is a symmetric matrix with dimension 75×75, it is efficient to use the upper triangle of $Cov(\mathbf{S}[n])$ as temporal features. However, in this case, the number of elements of the upper triangle of $Cov(\mathbf{S}[n])$ is $75(75+1)/2 = 2850$. In other words, the dimension of *Covariance of Joints* is 2850, which makes SVM classifier model unsuitable for real-time virtual military training system. In order to reduce the feature dimension of *Covariance of Joints*, we select 6 joints (Elbow Left, Elbow Right, Wrist Left, Wrist Right, Hand Left, and Hand Right) among 25 joints. Therefore, in (9), 1×75-dimensional $\mathbf{S}[m] = [\mathbf{P}_1[m] \cdots \mathbf{P}_{25}[m]]$ becomes 1×18-dimensional $\mathbf{S}[m] = [\mathbf{P}_5[m]\ \mathbf{P}_6[m]\ \mathbf{P}_7[m]\ \mathbf{P}_9[m]\ \mathbf{P}_{10}[m]\ \mathbf{P}_{11}[m]]$. In addition, the number of elements of the upper triangle of $Cov(\mathbf{S}[n])$ is reduced from 2850 to $18(18+1)/2 = 171$.

After snapshot and enhanced temporal feature extractions, the dimension of feature vector which is utilized for training SVM classifier model is $109 + 109 + 171 = 389$.

III. EXPERIMENTAL EVALUATION AND RESULTS

In this section, we evaluate the proposed EM-HARS using newly captured human action datasets. This dataset was captured by using six Kinects which are configured as shown in Fig. 2. In addition, this dataset consists of 19 human actions performed by 4 different subjects. Each subject performed every action nine or ten times. SVM with radial basis kernel is utilized for classification. In order to train the SVM classifier

TABLE IV
CLASSIFICATION ACCURACY RESULTS FOR EXPERIMENTS.

Action Type	EM-HARS	HARS-MK
Change Weapon Pistol	93.52	91.33
Change Weapon Rifle	91.34	89.75
Change Weapon Grenade	92.19	90.47
Change Weapon Sword	92.43	90.20
Throw High Left	91.62	90.75
Throw High Right	91.73	90.33
Throw Low Left	93.47	91.27
Throw Low Right	93.68	92.01
Reload Pistol	90.79	86.75
Reload Rifle	89.82	85.33
Shoot Sword	92.03	90.87
Lean Left	88.43	87.25
Lean Right	89.02	87.14
Pick Up	91.70	89.25
Pick Down	93.37	91.50
Open	92.08	89.73
Close	91.58	89.35
Telescope	92.72	90.26
Gasmask	93.70	91.50

model, data sequences of three subjects were used, and those of the remaining subject are used for testing.

Table IV shows the classification accuracy results for our experiments. The figures shown in Table IV are expressed as a percentage. The average classification accuracy of the proposed EM-HARS is roughly 91.85%. On the other hand, the average classification accuracy of conventional HARS-MK is approximately 89.74%. As shown in Table IV, the proposed EM-HARS achieves a higher average classification accuracy for all human action types than that of HARS-MK. These results suggest that the proposed EM-HARS is more suitable as a virtual military training simulator than HARS-MK.

IV. CONCLUSION

In this paper, we proposed EM-HARS for virtual training simulator. Compared to our previous HARS-MK, in EM-HARS, temporal feature extractor which captures the temporal characteristics human action is enhanced by employing *Covariance of Joints*. Also we newly captured skeleton data of 19 human actions including military training actions. In order to classify human actions, SVM was applied to train the classifier model. The experiment results demonstrated the feasibility of EM-HARS as a virtual military training system.

ACKNOWLEDGMENT

This work was supported by the ICT R&D program of MSIP/IITP. [R0101-15-0168, Development of ODM-interactive Software Technology supporting Live-Virtual Soldier Exercises]

REFERENCES

[1] D. Lee *et al.*, “A development of virtual reality game utilizing Kinect, Oculus Rift and smartphone,” *International Journal of Applied Engineering Research*, vol. 11, no. 2, pp. 829-833, 2016.

[2] A. Lele, “Virtual reality and its military utility,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 1, pp. 17-26. Feb. 2013.

[3] Z. Merchant *et al.*, “Effectiveness of virtual reality-based instruction on students’ learning outcomes in K-12 and higher education: A meta-analysis,” *Computers & Education*, vol. 70, pp. 29-40, Jan. 2014.

[4] D. Meldrum *et al.*, “Virtual reality rehabilitation of balance: assessment of the usability of the Nintendo Wii® Fit Plus,” *Disability and rehabilitation: assistive technology*, vol. 7, no. 3, pp. 205-210, 2012.

[5] J. Park *et al.*, “3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 415-427, Jun. 2014.

[6] T. Kim, J. Kang, S. Lee and A. C. Bovik, “Multimodal interactive continuous scoring of subjective 3D video quality of experience,” *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 387-402, Feb. 2014.

[7] K. Lee, A. K. Moorthy, S. Lee and A. C. Bovik, “3D visual activity assessment based on natural scene statistics,” *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 450-465, Jan. 2014.

[8] H. Kim, S. Lee and A. C. Bovik, “Saliency prediction on stereoscopic videos,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1476-1490, Apr. 2014.

[9] J. Park, H. Oh, S. Lee and A. C. Bovik, “3D visual discomfort predictor: Analysis of disparity and neural activity statistics,” *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1101-1114, Mar. 2015.

[10] T. Kim, S. Lee and A. C. Bovik, “Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3D images,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4335-4347, Nov. 2015.

[11] K. Lee and S. Lee, “3D perception based quality pooling: Stereopsis, binocular rivalry, and binocular suppression,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 533-545, Apr. 2015.

[12] H. Oh, S. Lee and A. C. Bovik, “Stereoscopic 3D visual discomfort prediction: a dynamic accommodation and vergence interaction model”, *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 615-629, Feb. 2016.

[13] B. W. Knerr, “Immersive simulation training for the dismounted Soldier (No. ARI-SR-2007-01),” *United States Army Research Institute for the Behavior and Social Sciences*, 2007.

[14] F. Lv and R. Nevatia, “Single view human action recognition using key pose matching and viterbi path searching,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Jun. 2007.

[15] H. Liu and L. Li, “Human action recognition using maximum temporal inter-class dissimilarity,” *In The Proceedings of the Second International Conference on Communications, Signal Processing, and Systems*, Springer International Publishing, pp. 961-969, 2014.

[16] G. T. Papadopoulos *et al.*, “Real-time skeleton-tracking-based human action recognition using kinect data,” *In MultiMedia Modeling*, Springer International Publishing, pp. 473-483, Jan. 2014.

[17] Z. Cheng *et al.*, “Human daily action analysis with multi-view and color-depth data,” *In Computer Vision ECCV 2012. Workshops and Demonstrations*, Springer Berlin Heidelberg, pp. 52-61, Oct. 2012.

[18] B. Ni, G. Wang and P. Moulin, “Rgbd-hudaact: A color-depth video database for human daily activity recognition,” *In Consumer Depth Cameras for Computer Vision*, Springer London, pp. 193-208, 2013.

[19] A. A. Liu *et al.*, “Single/multi-view human action recognition via regularized multi-task learning,” *Neurocomputing*, vol. 151, pp. 544-553, Mar. 2015.

[20] J. Kim, I. Lee, J. Kim and S. Lee “Implementation of an omnidirectional human motion capture system using multiple kinect sensors,” *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 98, no. 9, pp. 2004-2008, 2015.

[21] B. Kwon *et al.*, “Implementation of human action recognition system using multiple Kinect sensors,” *In Advances in Multimedia Information Processing-PCM 2015*, Springer International Publishing, pp. 334-343, 2015.

[22] M. E. Hussein, M. Torki, M. A. Gowayyed and M. El-Saban, “Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations,” *In International Joint Conference on Artificial Intelligence*, vol. 13, pp. 2466-2472, Aug. 2013.

[23] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[24] J. Kim, D. Kim, I. Lee, J. Kim, H. Oh and S. Lee, “Human gait prediction method using Microsoft Kinect,” *International Workshop on Advanced Image Technology (IWAIT 2016)*, Jan. 2016.

[25] M. A. Livingston, J. Sebastian, Z. Ai and J. W. Decker, “Performance measurements for the Microsoft Kinect skeleton,” *In IEEE Virtual Reality Short Papers and Posters (VRW)*, pp. 119-120, Mar. 2012.