sensors

MDPI

Article

MFA-Net: Motion Feature Augmented Network for Dynamic Hand Gesture Recognition from Skeletal Data †

Xinghao Chen [1], Guijin Wang [1,*], Hengkai Guo [2], Cairong Zhang [1], Hang Wang [3] and Li Zhang [1]

1 Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; chen-xh13@mails.tsinghua.edu.cn (X.C.); zcr17@mails.tsinghua.edu.cn (C.Z.); chinazhangli@tsinghua.edu.cn (L.Z.)
2 AI Lab, Bytedance Inc., Beijing 100086, China; guohengkai@bytedance.com
3 Beijing Huajie IMI Technology Co., Ltd, Beijing 100193, China; wanghang@hjimi.com
* Correspondence: wangguijin@tsinghua.edu.cn; Tel.: +86-010-62781430
† This paper is an extended version of our paper published in: Chen, X.; Guo, H.; Wang, G.; Zhang, L. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017.

Received: 6 December 2018; Accepted: 7 January 2019; Published: 10 January 2019

Abstract: Dynamic hand gesture recognition has attracted increasing attention because of its importance for human–computer interaction. In this paper, we propose a novel motion feature augmented network (MFA-Net) for dynamic hand gesture recognition from skeletal data. MFA-Net exploits motion features of finger and global movements to augment features of deep network for gesture recognition. To describe finger articulated movements, finger motion features are extracted from the hand skeleton sequence via a variational autoencoder. Global motion features are utilized to represent the global movements of hand skeleton. These motion features along with the skeleton sequence are then fed into three branches of a recurrent neural network (RNN), which augment the motion features for RNN and improve the classification performance. The proposed MFA-Net is evaluated on two challenging skeleton-based dynamic hand gesture datasets, including DHG-14/28 dataset and SHREC'17 dataset. Experimental results demonstrate that our proposed method achieves comparable performance on DHG-14/28 dataset and better performance on SHREC'17 dataset when compared with start-of-the-art methods.

Keywords: skeleton; gesture recognition; recurrent neural networks; feature augmentation

1. Introduction

Hand gesture provides an efficient and natural way for human–computer interaction (HCI) due to its flexibility and expressiveness. Hand gesture recognition has great potentials for applications in sign language recognition, remote control and virtual reality and has attracted great research interest in past decades [1–14]. Generally, hand gesture recognitions are categorized into static hand gesture recognition [3,4,15] and dynamic hand gesture recognition [5–7]. Static gesture recognition predicts the configuration or posture of the hand from a single image, while dynamic hand gesture recognition aims to understand what a hand sequence conveys. In this paper, we focus on dynamic hand gesture recognition. It remains a challenging task due to high intra-class variance because the way of performing a gesture differs from person to person.

Existing methods on dynamic hand gesture recognition usually take RGB images and depth images [16,17] as input. Some of them use multi-modal input including IR images [6] or audio stream [7]. Recent progresses on hand pose estimation [18–25] enable the acquisition of more accurate hand skeletons in real-time. Commercial sensors such as LeapMotion [26] and Intel Realsense Camera [27] can capture the hand poses with reasonably good quality. Therefore, the research on dynamic hand gesture recognition from 3D hand skeleton sequences has been greatly promoted.

De Smedt et al. [28] proposed a skeleton-based approach for dynamic hand gesture recognition and demonstrated its superiority over depth-based approaches. Temporal pyramid representation is utilized to model temporal information. Similarly, most recent methods for skeleton-based dynamic hand gesture recognition are based on hand-crafted features [29–31]. The temporal information is not fully exploited. Another family of solutions utilizes recurrent neural networks (RNN) to process the input hand skeleton sequences and predict the gesture class [32,33]. However, these methods only treat the raw skeleton sequences as input and do not fully leverage the properties of dynamic hand gestures, whose most important clues are articulated movements of fingers and the global movements of the hand. To this end, we take advantage of hand-crafted motion features and deep features from RNN to promote the performance of dynamic hand gesture recognition from skeletal data.

In this paper, we propose a Motion Feature Augmented Network (MFA-Net) for skeleton-based dynamic hand gesture recognition. We extract finger articulated features from the hand skeleton by a variational autoencoder (VAE), which is an efficient and concise representation of the finger articulated movements. To describe the global movements of the hand, we extract the global rotation and global translation of the hand. A distance adaptive discretization scheme is exploited to better model the amplitude of the gestures. The finger motion features and global features along with the skeleton sequence are fed into a RNN to predict the class of input gesture. Experiments on the publicly available skeleton-based DHG-14/28 dataset [28] and SHREC'17 dataset [34] demonstrate the effectiveness of our proposed method.

A preliminary version of this paper is presented in [35]. This paper extends the preliminary work [35] in several aspects: (1) A more extensive survey on related work is provided, including RGB-D/skeleton based static/dynamic hand gesture and feature augmentation methods. (2) An improved feature representation is proposed to describe the configurations of hand skeleton articulation via variational autoencoder (VAE) and further promotes the performance on DHG-14/28 dataset. (3) The proposed method was evaluated on a new dataset (SHREC'17 dataset [34]) and outperformed state-of-the-art methods.

The remainder of this paper is organized as follows. In Section 2, we review prior approaches that are related to our proposed method. In Section 3, we present an overview about our proposed motion feature augmented network. In Section 4, we provide details of extracting motion features from hand skeleton sequences. Evaluations on public datasets and ablation studies are provided in Section 5. Section 6 gives a brief conclusion of this paper and discussion of future work.

2. Related Work

In this section, we briefly review recent methods on hand gesture recognition, which are broadly categorized into static hand gesture and dynamic hand gesture recognition. For dynamic hand gesture recognition, we briefly review related work on RGB-D based and skeleton-based methods. We also review some recent methods of feature augmentation. More comprehensive reviews on hand gesture recognition are found in [36–39].

2.1. Static Hand Gesture Recognition

Static hand gesture recognition aims to predict the gesture label for a single image. Ren et al. [15] proposed the finger earth mover's distance metric (FEMD) for classifying hand gestures using Kinect camera. Wang et al. [40] proposed a new distance metric called superpixel earth mover's distance metric (SP-EMD) to measure the dissimilarity between gestures. Chen et al. [3] first located

the hand keypoints from the depth images and exploited angle features of finger roots for hand gesture recognition via finger length weighted Mahalanobis distance. Similarly, hand skeleton is estimated from depth image and joint angle features are used for classification in [4]. Koller et al. [41] exploited convolutional neural network (CNN) in weakly supervised training manner for hand gesture recognition. Despite its advancements in recent years, static hand gesture recognition fundamentally lacks of capability to handle temporal information and exhibits limitations for practical applications.

2.2. RGB-D Based Dynamic Hand Gesture Recognition

Dynamic hand gesture recognition from RGB-D frames has been actively researched for decades [42–49]. Zhu et al. [44] proposed a framework using 3D CNN and Convolutional LSTM to recognize gestures from both RGB and depth sequences. Molchanov et al. [6] proposed a recurrent 3D CNN to perform simultaneous dynamic hand gesture detection and classification from multimodal data, including depth, color, optical flow, and stereo IR streams. Zhang et al. [45] proposed a deep architecture to first learn spatiotemporal features using 3D CNN and bidirectional convolutional LSTM and learn higher-level features via 2D CNN. Köpüklü et al. [50] proposed a data level fusion strategy named Motion Fused Frames (MFFs) to fuse motion information into gesture sequences. However, RGB-D based dynamic hand gesture recognition may suffer from clustered background, heavy input data burden, etc.

2.3. Skeleton-Based Dynamic Hand Gesture Recognition

As the recent progress of fast and accurate hand pose estimation algorithms [18–24] and related sensors or cameras, 3D hand poses are more easily obtained. More research interests have been focused on skeleton-based dynamic hand gesture recognition.

De Smedt et al. [28] proposed a skeleton-based dynamic hand gesture recognition algorithm and suggested that skeleton-based method achieved superior performance over depth-based methods. In their approach, a new descriptor named Shape of Connected Joints (SoCJ) is encoded by Fisher vector representations to describe the hand skeleton. Histogram of hand directions and wrist orientations are adopted to represent the hand movements in global space. Temporal pyramid is exploited to model the temporal information. Similar representations are further exploited in [29] for hand gesture recognition via learning on Riemannian manifold. Boulahia et al. [30] adopted a feature set named Handwriting-Inspired Features (HIF3D) [51] which was originally proposed for skeleton-based action recognition to address the problem of skeleton-based dynamic hand gesture recognition.Caputo et al. [31] applied several processing methods (such as rotation, smoothing, scaling, etc.) on the hand gesture trajectory and matched it with templates using gesture distance metrics. These methods are based on carefully designed hand-crafted features, which may be not optimal for hand gesture recognition.

There are arising trends to use deep learning methods for skeleton-based dynamic hand gesture recognition. Núñez et al. [32] adopted the combination of CNN and LSTM for dynamic hand gesture recognition and action recognition from skeletal data. A two-stage training strategy is used to first train the CNN and then fine tune the whole CNN + LSTM network. Ma et al. [33] focused on addressing noisy skeleton sequences and proposed a LSTM network together with a nested interval unscented Kalman filter (UKF) to improve performance for noisy datasets.

Different from above existing methods, our proposed method takes advantages of both hand-crafted features and deep learning methods to obtain optimal features for hand gesture recognition.

2.4. Feature Augmented Method

There have been some attempts to enhance the capability of deep neural network by fusing hand-crafted features into the network. Sadanandan et al. [52] proposed the feature augmented deep neural networks that augmented the raw input images with eigen images to improve the performance

of cell segmentation. Egede et al. [53] fused HOG features, geometric features and deep learned features into a Relevance Vector Regressor (RVR) to estimate pain intensity. Similarly, Manivannan et al. [54] concatenated hand-crafted features with CNN features for gland segmentation. Wang et al. [55] adopted the idea of combining hand-crafted features and CNN features to address the problem of action recognition.

Inspired by these methods, we exploited motion features to augment the neural network for better performance of skeleton-based dynamic hand gesture recognition.

3. Overview of the Proposed Framework

The framework of our proposed motion feature augmented network (MFA-Net) is shown in Figure 1. MFA-Net takes a hand skeleton sequence as input and predicts the class label of dynamic hand gesture. It consists of three branches, which process finger motion features, global motion features and skeletons, respectively. The most important clues for a dynamic hand gesture are articulated movements of fingers and the global movements of the hand. Therefore, augmenting the original skeletons with finger and global motion features is beneficial to dynamic hand gesture recognition.

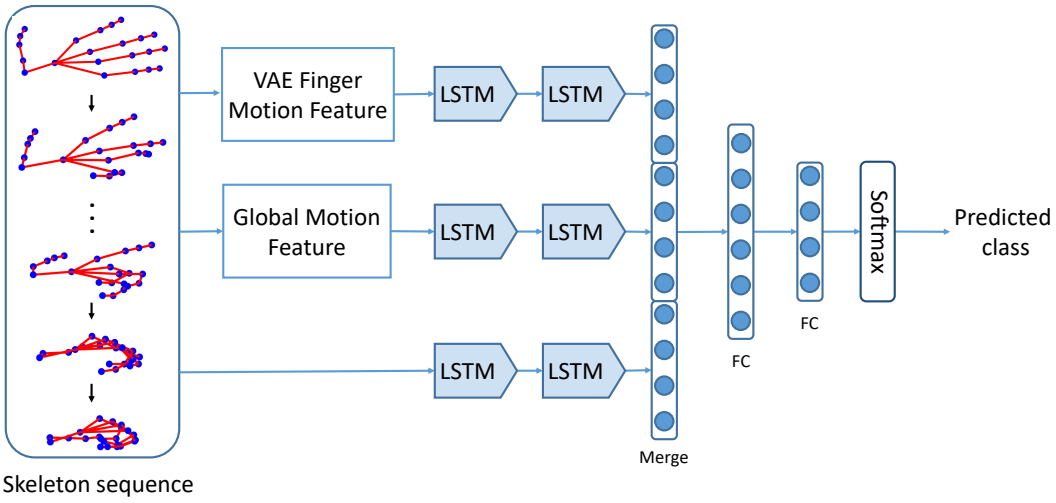

Figure 1. The framework of our proposed motion feature augmented network (MFA-Net). Finger motion features and global motion features are extracted from the input dynamic hand gesture skeleton sequence. These motion features, along with the skeleton sequence, are fed into different branches of a Long-Short Term Memory (LSTM) network to get the predicted class of input gesture.

Firstly the global motion features and finger motion features are extracted from the input skeleton sequence. The global movement of a dynamic hand gesture can be represented by the global translation and rotation of the hand. For finger motion features, we explore two kinds of representations in this paper: kinematic features and variational autoencoder (VAE) features. The hand skeleton can be directly and effectively represented by a kinematic hand model whose parameters are the angles of bones, the global translation and global rotation [20,21]. Therefore, these kinematic hand parameters can serve as efficient and discriminating features for dynamic hand gesture recognition. We also explore latent features extracted from a hand skeleton by a variational autoencoder (VAE), which captures the latent representations of a hand pose. In our approach, theses features with offset and dynamic pose modeling are utilized as the motion features to represent dynamic hand gestures. The details of motion feature extraction are presented in Section 4.

We exploit the recurrent neural network (RNN) to model temporal information for its great successes in temporal sequences recognition tasks [6,56]. More specifically, we adopt the Long-Short Term Memory (LSTM) network, which is a successful variant of RNN that can model long temporal

information of sequences. Although LSTM can somehow learn features from the input skeleton sequences, some information may be absent or weakened, which will hinder the classification performance. To this end, we augment features for LSTM by combining the global and finger motion features and the original skeleton. The finger motion features and global motion features are extracted from the input skeleton sequence. These motion features and the input skeleton sequence are fed into the LSTM. Each branch contains two LSTM layers and one fully connected (FC) layer. Outputs from three branches are concatenated together, followed by three FC layers and a softmax layer for class prediction. All layers are followed by a dropout layer and FC layers are followed by a ReLU function.

4. Motion Feature Extraction

In this section, we describe how to extract finger motion features $\mathcal{H}(\mathcal{S})$ and global motion features $\mathcal{G}(\mathcal{S})$ from the input hand skeleton sequence $\mathcal{S} = \{s^t\}_{t=1}^{T}$, where $s^t = \{x_i^t, y_i^t, z_i^t\}_{i=1}^{J}$ denotes a hand skeleton for frame t, T is the number of frames of this sequence and J is the number of joints for hand skeleton.

4.1. Global Motion Feature

The global motion features (global rotation and global translation) are important for dynamic hand gesture. Typically, the global status of the hand can be determined by the wrist joint, palm joint and metacarpophalangeal (MCP) joints, which are denoted by p^t. We use Kabsch algorithm [57] to infer the global rotation $\mathcal{G}_r$ and global translation $\mathcal{G}_l$ of a hand skeleton:

$$[\mathcal{G}_l, \mathcal{G}_r] = Kabsch(p^t, p_0), \tag{1}$$

where $\mathcal{G}_r = (r_x, r_y, r_z)$ represents the rotations along three axes, $\mathcal{G}_l = (\rho, \theta, \phi)$ is the spherical coordinate of global translation, p_0 is a reference palm that centers at $(0, 0, 0)$ and faces the camera with the palm upwards.

The amplitudes of hand gestures differ from person to person for the same gesture. Therefore, previous work [28] ignored the amplitude part ρ of global translation. However, sometimes the amplitude is critical for gestures. For example, gesture *Grab* and gesture *Pinch* are quite similar except for the amplitude of the gesture. To this end, we propose a distance adaptive discretization (DAD) method to extract global translation amplitude feature, inspired by Distance Adaptive Scheme [4,58], which is used for feature selection for hand pose estimation. The DAD method discretizes ρ into M bins using the threshold $\{\eta_i\}_{i=1}^{M}$. A Gaussian distribution kernel $g(x)$ is used to generate the thresholds:

$$\int_0^{\eta_i} g(x)dx = \frac{i}{M} \int_0^{\sigma} g(x)dx, \tag{2}$$

where σ is the standard deviation of the Gaussian function. In our experiments, we set $\sigma = 1.5 r_{palm}$ where r_{palm} is the radius of the palm. The global feature for a hand skeleton can be written as:

$$\Phi^t = [\rho_{bin}, \theta, \phi, r_x, r_y, r_z], \tag{3}$$

where ρ_{bin} is the discrete representation of ρ using the thresholds determined by Equation (2).

Similar to previous work [59], we use offset pose Φ_{op}^t and dynamic pose Φ_{dp}^t for global features Φ^t to model the global motion features. The offset pose represents the offset from current global features to those of the first frame of gesture sequence:

$$\Phi_{op}^t = \Phi^t - \Phi^1. \tag{4}$$

The dynamic pose represents the difference of global features between current frame and several previous frames:

$$\Phi_{dp}^t = \{\Phi^t - \Phi^{t-s} | s = 1, 5, 10\}. \tag{5}$$

There features can enhance the representability of the global motion of the hand and thus can model the temporal information of dynamic hand gesture. All above features are concatenated to form the global motion features $\mathcal{G}^t(\mathcal{S}) = [\Phi^t, \Phi^t_{op}, \Phi^t_{dp}]$ for frame t.

4.2. Finger Motion Feature

For finger motion features, we explore two kinds of representations, namely kinematic features and variational autoencoder (VAE) features, which are presented in Sections 4.2.1 and 4.2.2, respectively. Kinematic finger motion features are exploited in our preliminary work [35] and in this paper we propose a more effective representation for finger motion feature using VAE. The impact of kinematic and VAE finger motion features is discussed in Section 5.3.

4.2.1. Kinematic Finger Motion Feature

For many dynamic hand gestures, the articulated movements of fingers are critical because the global movements may be insignificant, especially for fine-grained gestures. We use 20 DoFs (degree of freedoms) to model the finger articulation movements. For the MCP joints, there are 2 DoFs for each joint. For proximal interphalangeal (PIP) and distal interphalangeal (DIP) joints, 1 DoF is used to describe the angle of bone. These kinematic parameters retain rich information for the shape of the hand skeleton. We use $\mathcal{IK}(\cdot)$ to denote the inverse kinematics function that derives hand kinematic parameters from the original hand skeleton s^t:

$$\Theta^t_{km} = \mathcal{IK}(s^t). \tag{6}$$

Similarly, we use dynamic pose Θ^t_{dp} and offset pose Θ^t_{op} to model the finger motion feature:

$$\Theta^t_{op} = \Theta^t_{km} - \Theta^1_{km}, \tag{7}$$

$$\Theta^t_{dp} = \{\Theta^t_{km} - \Theta^{t-s}_{km} | s = 1, 5, 10\}. \tag{8}$$

These features are concatenated to obtain the kinematic finger motion features $\mathcal{F}^t_{km}(\mathcal{S}) = [\Theta^t_{km}, \Theta^t_{op}, \Theta^t_{dp}]$ for frame t.

4.2.2. VAE Finger Motion Feature

The pose variational autoencoder (PoseVAE) consists of an encoder ($Enc(\cdot)$) and a decoder ($Dec(\cdot)$), as shown in Figure 2. Both the encoder and the decoder have two fully connected (FC) layers, with the dimensions of 32 and 20, respectively. The encoder projects the original hand skeleton into latent representations:

$$\Theta^t_{vae} = Enc(s^t). \tag{9}$$

The decoder produces a decoded hand skeleton given the latent features:

$$\tilde{s^t} = Dec(\Theta^t_{vae}). \tag{10}$$

The PoseVAE tries to minimize the distance between the original skeleton s^t and the decoded skeleton $\tilde{s^t}$. We use the encoder to obtain latent features of the hand skeleton to describe the articulated movements of fingers. Similar to Equations (7) and (8), dynamic pose and offset pose are concatenated with the VAE features to obtain the VAE finger motion features $\mathcal{F}^t_{vae}(\mathcal{S})$.

There are several benefits of using VAE features to represent the finger articulated motion. Firstly, learning latent representations for hand skeletons by PoseVAE has the potential to obtain more representative features than hand-crafted features such as kinematic features. Secondly, PoseVAE reduces the noises in hand skeletons that are introduced by inaccurate annotations. As shown in Figure 2, the input hand pose contains noise in the middle and ring fingers whose joints are inaccurately

annotated, resulting in a physically implausible hand skeleton. The output pose of PoseVAE is much smoother and thus removes the unnecessary noise. Therefore, the latent representations learned by the PoseVAE are more robust and insensitive to the noise, which is beneficial to hand gesture recognition.

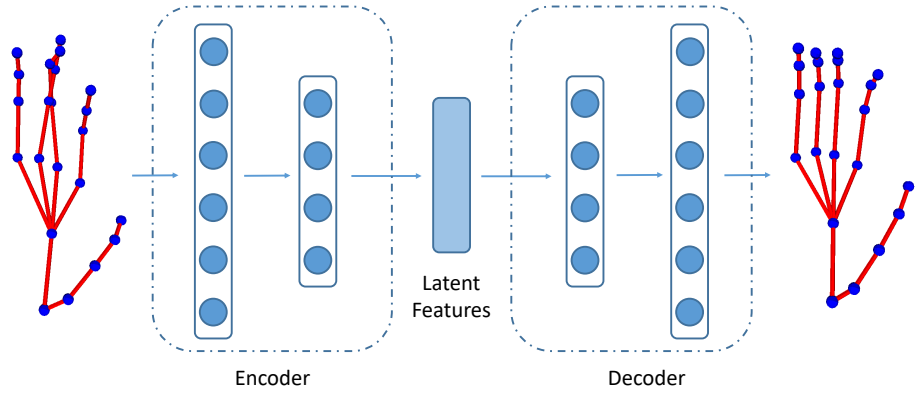

Figure 2. PoseVAE: Variational autoencoder for hand pose. We use the encoder to obtain latent features of the hand skeleton to describe the articulated movements of fingers.

5. Experiments

In this section, we show the experimental results of our proposed method. Firstly, the datasets used for the experiments and some implementation details are briefly introduced. Secondly, comparisons with state-of-the-art methods and ablation studies are shown and discussed.

5.1. Implementation

The proposed framework was implemented in Keras [60]. We used Adam [61] algorithm with mini-batch of 32 to train the network. The parameters of Adam were set to the default settings suggested in [61], with learning rate $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. During training, the network minimized the cross entropy loss between the predicted labels and the ground truth labels. The network was trained for 100 epochs. In our experiments, M was set to $M = 5$ in Equation (2). Every skeleton sequence was subtracted by the palm position of the first frame and scaled the amplitude to 1 before fed into third branch in Figure 1.

5.2. Comparison with State-Of-The-Art Methods

Although there are many dynamic hand gesture datasets, e.g. Chalearn IsoGD [49], Nvidia Gesture Dataset [6], etc., these datasets only provide RGB-D images and do not contains skeleton information. To evaluate our proposed method of skeleton-based dynamic hand gesture recognition, we conducted experiments on DHG-14/28 dataset [28] and SHREC'17 dataset [34], which provide hand skeleton annotations for each gesture sequence. We report the classification accuracy to evaluate our proposed method, which is the most commonly used evaluation metric for hand gesture recognition.

5.2.1. DHG-14/28 Dataset

DHG-14/28 [28] is a public dynamic hand gesture dataset that provides hand gesture sequences with depth images and corresponding skeletons. The depth images were captured by Intel Realsense Camera and hand skeletons were obtained by Intel Realsense SDK. DHG-14/28 is a challenging dataset since it contains hand gesture from 20 subjects and has 14 gestures with two different finger configurations. Totally, DHG-14/28 consists of 2800 sequences. Each hand skeleton is represented by

22 joints. Since our proposed method focuses on dynamic hand gesture recognition from skeletal data, we only used the skeleton information of the datasets to conduct our experiments.

On DHG-14/28 dataset, we followed the same experimental setup as previous work [28,29,32,33,62–64], using the leave-one subject-out cross-validation (LOOCV) strategy for all following experiments. The proposed network was trained on data from 19 subjects and tested on the remaining one. Therefore, these experiments were repeated 20 times, with a different subject being used for testing. There are five fine-grained gestures and nine coarse gestures in DHG-14. In the experiments presented below, the MFA-Net was trained to classify 14 gestures and the classification accuracies of all gestures, five fine-grained gestures and nine coarse gestures are reported.

We compared our work with several state-of-the art methods [28,29,32,33,62–64] on DHG-14/28 dataset. The recognition rates of different methods on DHG-14 and DHG-28 dataset are shown in Table 1. It shows that our proposed method outperforms most state-of-the-art methods on DHG-14 dataset (14 gestures setting) and achieves comparable performance with SL-fusion-Average [64] and CNN + LSTM [32]. It should be noted that SL-fusion-Average [64] exploited both depth images and skeletons as input, while our method only relies on skeletons. CNN + LSTM [32], NIUKF-LSTM [33] and our proposed MFA-Net are all based on LSTM. NIUKF-LSTM [33] focuses on handling noisy skeleton data using nested interval unscented Kalman filter (NIUKF) and CNN + LSTM [32] utilizes CNN to learn spatiotemporal features from skeleton sequence. They both exploit LSTMs afterwards for gesture classification. Different from these work, our proposed MFA-Net focuses on augmenting the motion features for LSTM, which well preserves the properties of dynamic hand gestures. Moreover, the proposed MFA-Net can be compatible with existing LSTM-based methods by replacing the third branch of our method with prior methods (e.g., NIUKF-LSTM [33]). However, it is out of the focus of this paper to fully explore the combinations with prior work and we leave it for future work. To better illustrate the performance of our proposed algorithm, the confusion matrix of 14 classes is shown in Figure A1. It can be observed that the confusion between gesture *Grab* and *Pinch* is severe, due to the high similarity of these two gestures. However, our algorithm does improve the performance of these two gestures compared with those of SoCJ + HoHD + HoWR [28], with 60.25% average recognition rate for these two gestures of our method and 59.0% for SoCJ + HoHD + HoWR [28]. It can be observed that our method promotes the classification accuracy of fine-grained gestures and coarse gestures when compared with SoCJ + HoHD + HoWR [28].

Table 1. Comparison of recognition rates (%) with state-of-the-art methods on DHG-14/28 dataset.

Method	DHG-14			DHG-28
	Fine	**Coarse**	**Both**	**Both**
HON4D [62]	-	-	75.53	74.03
HOG^2 [63]	-	-	80.85	76.53
Smedt et al. [29]	-	-	82.50	68.11
SoCJ + HoHD + HoWR [28]	73.60	88.33	83.07	80.0
NIUKF-LSTM [33]	-	-	84.92	80.44
SL-fusion-Average [64]	76.00	90.72	85.46	74.19
CNN + LSTM [32]	**78.0**	89.8	85.6	**81.1**
MFA-Net (Ours)	75.60	**91.39**	**85.75**	81.04

As shown in Table 1, our method is also better than most prior methods and achieves comparable performance with CNN + LSTM [32] when considering the more complicated 28-gesture classification task, which demonstrates the effectiveness of our proposed algorithm. Specifically, MFA-Nets boosts the accuracy by 6.85% when compared with a most recent method [64]. The confusion matrix of 28 classes is shown in Figure A2.

5.2.2. SHREC'17 Dataset

SHREC'17 dataset [34] was first introduced in SHREC 2017 track to evaluate the performance of dynamic hand gesture recognition. Similar to DHG-14/28 dataset, SHREC'17 dataset also consists of 14 gestures performed by 28 participants executing the same gesture with two different configurations of fingers. There are 1960 sequences in the training set and another 840 sequences in the testing set.

Following the evaluation protocol of SHREC'17 track [34], we trained our MFA-Net on 1960 samples and evaluated on the other 840 samples, which is the same with previous work [28,30,31,34,65]. We followed a similar data augmentation strategy as in [32] to add random scaling, shifting, time interpolation and noise to the original sequences. After data augmentation, the whole training set contained 9800 samples.

As shown in Table 2, our proposed MFA-Net achieves the accuracy of 91.31% for 14 gestures and 86.55% for 28 gestures and outperforms all prior methods for both experimental settings. Specifically, MFA-Net improves the accuracy for 28 gestures by about 4.7% when compared with existing best performance by SoCJ + HoHD + HoWR [28] and 6.07% when compared with more recent work [30].

Table 2. Comparison of recognition rates (%) with state-of-the-art methods on SHREC'17 dataset.

Method	14 Gestures	28 Gestures
HOD4D [62]	78.53	74.03
Riemannian Manifold [65]	79.61	62.00
Key Frames [34]	82.90	71.90
HOG^2 [63]	83.85	76.53
SoCJ + HoHD + HoWR [28]	88.24	81.90
3 cent + OED + FAD [31]	89.52	-
Boulahia et al. [30]	90.48	80.48
MFA-Net (Ours)	**91.31**	**86.55**

The confusion matrices for 14 gestures and 28 gestures recognition on SHREC'17 dataset are shown in Figures A3 and A4, respectively. Our MFA-Nets achieves accuracy higher than 85.0% for 12 out of 14 gestures. For the more challenging 28 gestures task, MFA-Net obtains accuracy higher than 85.0% for 15 out of 28 gestures and accuracy higher than 80.0% for 22 out of 28 gestures.

5.3. Ablation Studies

To verify the contributions of different modules of our proposed method, we conducted several self-comparison experiments on DHG-14 dataset, which has 14 gestures.

Since DHG-14 exploits LOOCV strategy for evaluation, there are totally 20 splitting protocols for 20 subjects. Existing studies only report the average classification accuracy of these 20 experiments, which is not sufficient to evaluate the performance and robustness of hand gesture recognition algorithms for different participants. In the ablation studies, we report the worst, best and average results of 20 different splitting protocols as well as the standard deviation, which is a more comprehensive metric for taking the inter-subject effects into account.

5.3.1. The Contributions of Motion Features Augmentation

We conducted several baseline experiments to explore how the motion feature augmentation strategy affects the accuracy of dynamic hand gesture recognition. The first baseline (*Skeleton*) only took the skeleton sequences as input and adopted LSTM for gesture recognition. The second baseline (*MF(Kinematic)*) only took motion features as input and removed the third branch of the framework shown in Figure 1. In this baseline, we exploited kinematic features for finger motion features.

As shown in the first three rows of Table 3, in most cases, combining skeleton and motion features outperforms two baselines in terms of worst, best, and average accuracy and the stand derivation.

Overall, the *Skeleton + MF (Kinematic)* has better average accuracies for fine, coarse and all gestures, which verify the effectiveness of the proposed strategy of augmenting LSTM with motion features.

Table 3. Recognition rates (%) of self-comparison experiments on DHG-14 dataset.

Method	Fine			Coarse			Both		
	Best	**Worst**	**Avg ± Std**	**Best**	**Worst**	**Avg ± Std**	**Best**	**Worst**	**Avg ± Std**
Skeleton	86.0	42.0	61.2 ± 12.37	97.78	74.44	86.44 ± 7.94	93.57	67.86	77.43 ± 6.82
MF(Kinematic)	84.0	46.0	71.5 ± 11.44	96.67	64.44	81.94 ± 8.17	90.0	58.57	78.21 ± 7.49
S + MF(Kinematic)	90.0	**56.0**	**76.9 ± 9.19**	97.78	72.22	89.0 ± 7.55	94.29	67.86	84.68 ± 6.67
S + MF(VAE)	**96.0**	48.0	75.6 ± 10.29	**100.0**	**76.67**	**91.39 ± 7.30**	**96.43**	**71.43**	**85.75 ± 6.71**

5.3.2. The Contributions of VAE Features

We then explored the effects of using VAE features or kinematic features for finger motion feature extraction. Kinematic features were exploited in our preliminary version of MFA-Net [35] and it achieves accuracy of 84.69% for DHG-14 dataset, as also shown in the third row of Table 3. When using PoseVAE to extract latent representations to describe hand skeleton, the accuracy increases to 85.75%, as shown in the last row of Table 3. In addition, the accuracies for best and worse subject for all 14 gestures are also improved, which indicates the effectiveness of the VAE features.

5.3.3. The Contributions of DAD Strategy

In Section 4.1, we introduce the distance adaptive discretization (DAD) strategy to handle with the amplitude of dynamic hand gesture. We conducted experiments to remove the term ρ_{bin} produced by DAD strategy in Equation (3). As shown in Table 4, adding DAD term increases the accuracies in terms of worst, best, average accuracy and the stand derivation, which demonstrates the contributions of DAD method. Specifically, the overall accuracy increases from 84.60% to 85.75% for 14 gestures on DHG dataset.

Table 4. Recognition rates (%) of MFA-Net with/without DAD strategy on DHG-14 dataset.

Method	Fine			Coarse			Both		
	Best	**Worst**	**Avg ± Std**	**Best**	**Worst**	**Avg ± Std**	**Best**	**Worst**	**Avg ± Std**
MFA-Net w/o DAD	92.0	42.0	74.2 ± 11.81	100.0	75.56	90.39 ± 6.89	**97.14**	67.86	84.60 ± 7.22
MFA-Net	**96.0**	**48.0**	**75.6 ± 10.29**	**100.0**	**76.67**	**91.39 ± 7.30**	96.43	**71.43**	**85.75 ± 6.71**

5.3.4. The Impacts of Different Classifiers

In the proposed MFA-Net, fully connected (FC) layers are utilized to classify gestures from deep features from LSTM blocks, as shown in Figure 1. To explore the impacts of different choices of classifiers and demonstrate the discriminability of the learned features, we extracted the deep features before the last two FC layers and fed them into different classifiers, including *k*-NN, an enhanced *k*-NN algorithm (Centroid Displacement-Based *k*-NN) [66] and random forest. The hyper-parameters for these classifiers were chosen using cross-validation on the training set. The recognition rates for different classifiers on SHREC'17 dataset [34] are shown in Table 5. One of the observations is that using FC layers as classifier performs better than others. It is intuitive since the deep features are learned together with the weights of FC layers. Another observation is that, even using simple classifier such as *k*-NN, the performances are good when compared with state-of-the-art methods. For example, the best prior performance on 14 gestures classification is 90.48% by Boulahia et al. [30], while *k*-NN obtains recognition rate of 90.60% and CD *k*-NN achieves 90.85%. On 28 gestures classification task, the highest recognition accuracy of existing methods is 81.90% [28], while both *k*-NN and CD *k*-NN achieve the accuracy of 86.07%. The considerably good performances of these classifiers demonstrate that the features produced by our proposed MFA-Net are quite discriminative for hand gesture recognition. To

better understand this, we used t-SNE [67] to visualize the 2D embedding of the features. Figure 3 shows that the features of MFA-Net exhibit separable feature distributions in manifold and can be easily distinguished. Moreover, the feature embeddings on testing set are highly similar to those on training set, which is beneficial to obtain a good classifier.

Table 5. Comparison of recognition rates (%) for different classifiers on SHREC'17 dataset.

Method	14 Gestures	28 Gestures
k-NN	90.60	86.07
CD *k*-NN [66]	90.83	86.07
Random Forest	90.36	85.24
FC Layers (Ours)	**91.31**	**86.55**

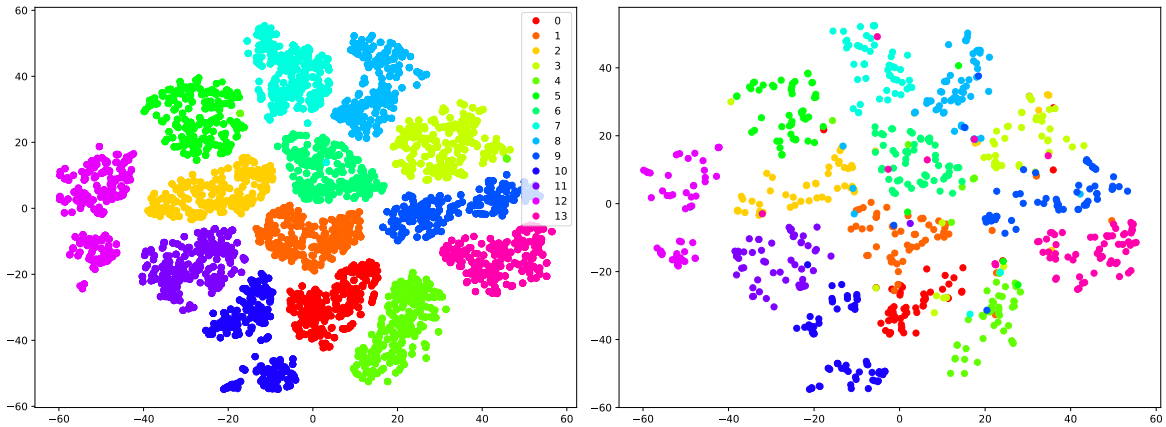

Figure 3. 2D t-SNE visualization of features before FC layers: (**Left**) feature embeddings of training set on SHREC'17 dataset; and (**Right**) feature embeddings of testing set.

5.4. Run Time Analysis

We evaluated the inference speed of MFA-Net on a computer equipped with 3.4 GHz i7-4770 CPU and Nvidia Tesla K40c GPU. MFA-Net takes about 1.47 ms to extract motion features for one hand skeleton and about 83.2 ms to predict the gesture for an input dynamic hand gesture sequence with 64 skeletons. Therefore, on average MFA-Net takes 2.77 ms to process one hand skeleton. In other words, MFA-Net can process 361 skeletons per second, which is sufficient for real-time performance.

6. Conclusions

This paper proposes the motion feature augmented network (MFA-Net) to recognize skeleton-based dynamic hand gestures. Finger motion features are extracted via a variational autoencoder from the hand skeleton sequence to describe finger articulated movements. Global motion features are utilized to represent the global movements of hand skeleton. The motion features, along with the skeleton sequence, are fed into three branches of RNN to predict the label of input gesture. Experiments demonstrate that our proposed MFA-Net achieves comparable performance with state-of-the-art methods on the public DHG-14/28 dataset and best performance on SHREC'17 dataset. Future work may focus on a hierarchical coarse to fine framework to achieve better classification performance.

Author Contributions: X.C. proposed the ideas, and conceived and designed the experiments; X.C. and H.G. performed the experiments; X.C., G.W. and C.Z. analyzed the data; G.W. supervised this work; H.W. and L.Z. provided useful discussions. and X.C. drafted the paper, which was revised by all authors.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Confusion Matrices

For DHG-14/28 [28] datasets, the confusion matrix of 14 classes is shown in Figure A1 and the confusion matrix of 28 classes is shown in Figure A2.

The confusion matrices for 14 gestures and 28 gestures recognition on SHREC'17 dataset [34] are shown in Figures A3 and A4, respectively.

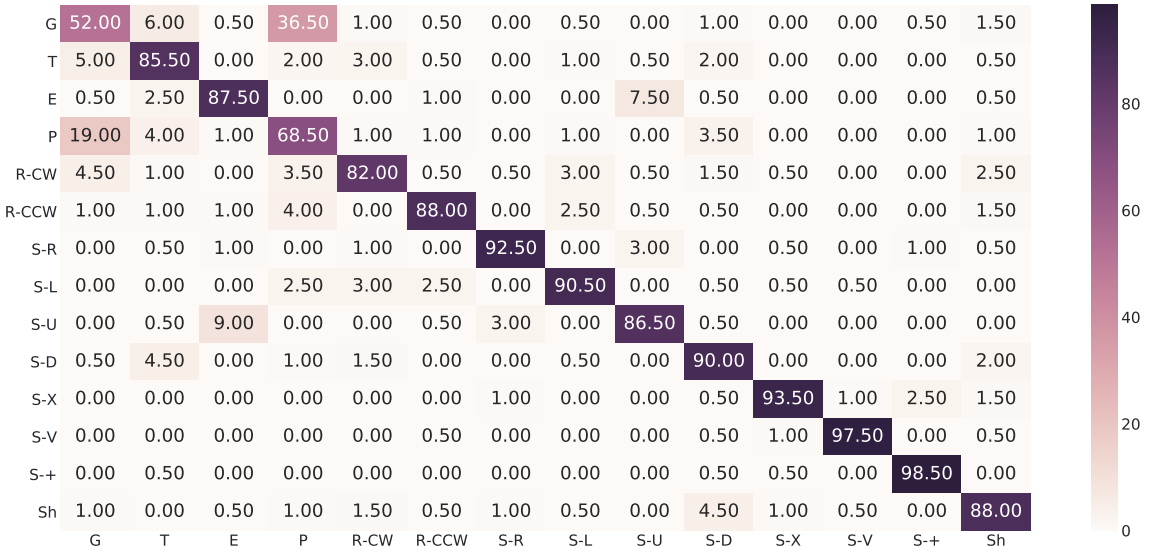

Figure A1. The confusion matrix of the proposed approach for DHG-14.

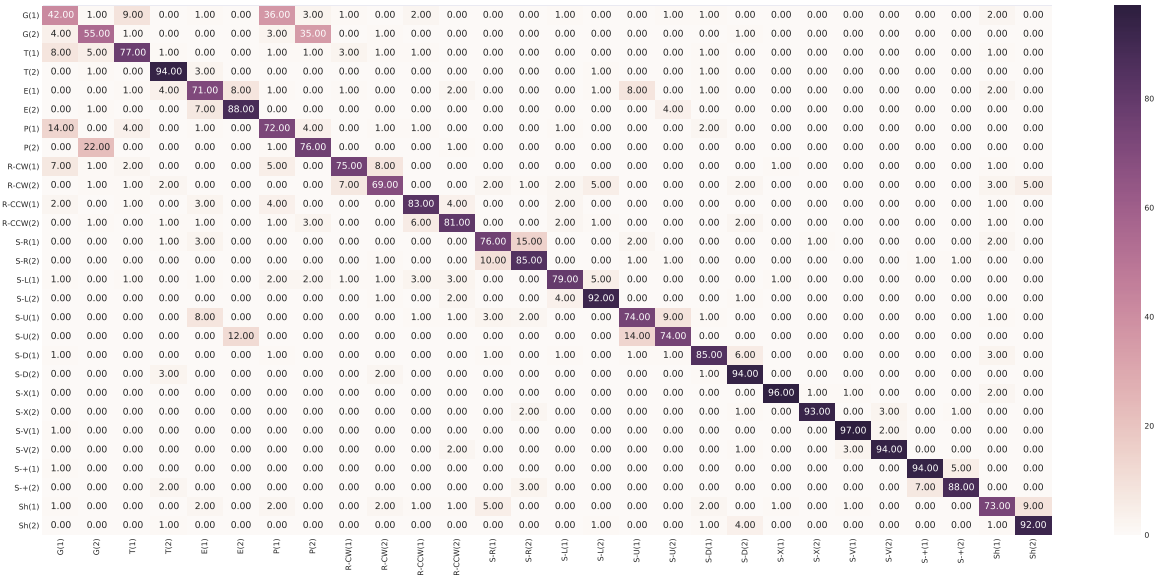

Figure A2. The confusion matrix of the proposed approach for DHG-28.

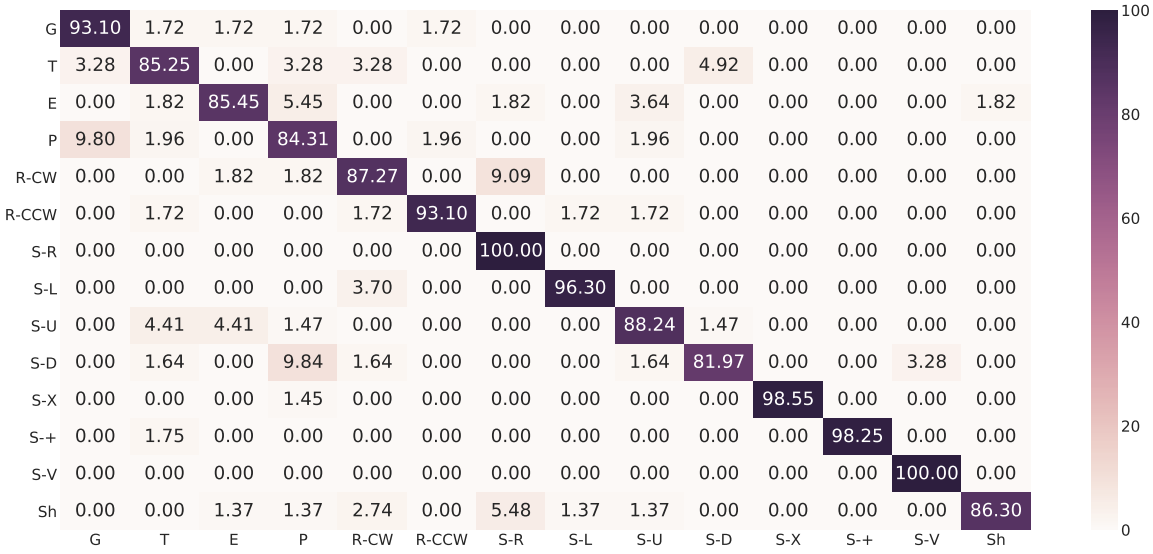

Figure A3. The confusion matrix of the proposed approach for SHREC'17 (14 gestures) dataset.

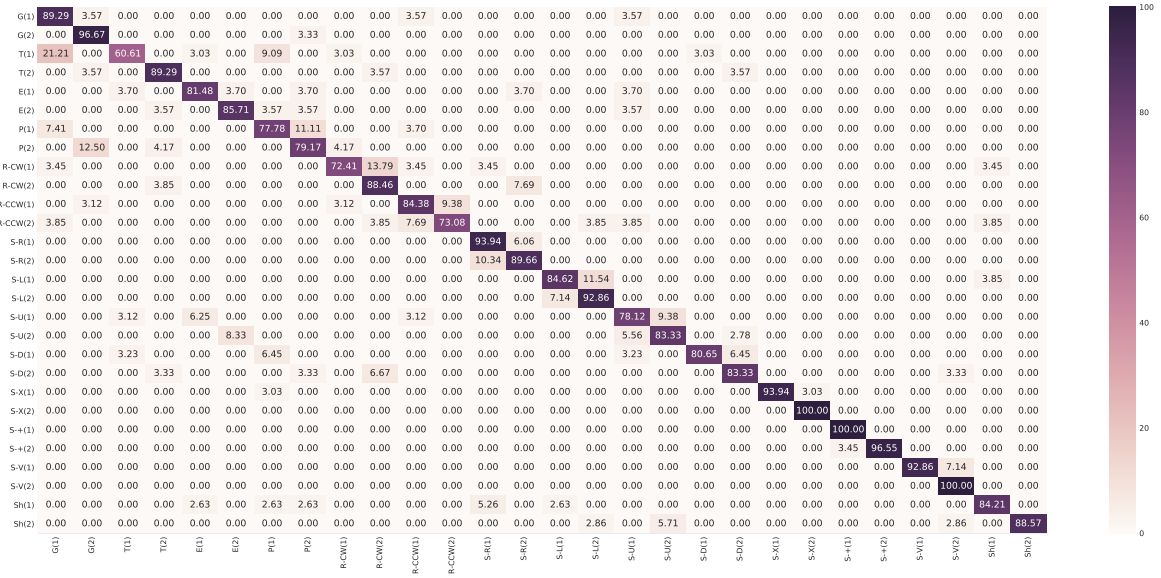

Figure A4. The confusion matrix of the proposed approach for SHREC'17 (28 gestures) dataset.

References

1. Mitra, S.; Acharya, T. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2007**, *37*, 311–324. [CrossRef]
2. Zeng, B.; Wang, G.; Lin, X. A hand gesture based interactive presentation system utilizing heterogeneous cameras. *Tsinghua Sci. Technol.* **2012**, *17*, 329–336. [CrossRef]
3. Chen, X.; Shi, C.; Liu, B. Static hand gesture recognition based on finger root-center-angle and length weighted Mahalanobis distance. In Proceedings of the SPIE 9897, Real-Time Image and Video Processing, Brussels, Belgium, 3–7 April 2016.
4. Dong, C.; Leu, M.C.; Yin, Z. American sign language alphabet recognition using microsoft kinect. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015.
5. Ohn-Bar, E.; Trivedi, M.M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377. [CrossRef]

6. Molchanov, P.; Yang, X.; Gupta, S.; Kim, K.; Tyree, S.; Kautz, J. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
7. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. [CrossRef]
8. Palacios, J.M.; Sagüés, C.; Montijano, E.; Llorente, S. Human-computer interaction based on hand gestures using RGB-D sensors. *Sensors* **2013**, *13*, 11842–11860. [CrossRef] [PubMed]
9. Choi, H.R.; Kim, T. Combined dynamic time warping with multiple sensors for 3D gesture recognition. *Sensors* **2017**, *17*, 1893. [CrossRef]
10. Abraham, L.; Urru, A.; Normani, N.; Wilk, M.; Walsh, M.; O'Flynn, B. Hand tracking and gesture recognition using lensless smart sensors. *Sensors* **2018**, *18*, 2834. [CrossRef]
11. Zhou, Q.; Xing, J.; Chen, W.; Zhang, X.; Yang, Q. From Signal to Image: Enabling Fine-Grained Gesture Recognition with Commercial Wi-Fi Devices. *Sensors* **2018**, *18*, 3142. [CrossRef]
12. Wang, X.; Tanaka, J. GesID: 3D Gesture Authentication Based on Depth Camera and One-Class Classification. *Sensors* **2018**, *18*, 3265. [CrossRef]
13. Wen, R.; Tay, W.L.; Nguyen, B.P.; Chng, C.B.; Chui, C.K. Hand gesture guided robot-assisted surgery based on a direct augmented reality interface. *Comput. Methods Programs Biomed.* **2014**, *116*, 68–80. [CrossRef]
14. Wen, R.; Nguyen, B.P.; Chng, C.B.; Chui, C.K. In situ spatial AR surgical planning using projector-Kinect system. In Proceedings of the Fourth Symposium on Information and Communication Technology, Danang, Vietnam, 5–6 December 2013.
15. Ren, Z.; Yuan, J.; Meng, J.; Zhang, Z. Robust part-based hand gesture recognition using kinect sensor. *IEEE Trans. Multimed.* **2013**, *15*, 1110–1120. [CrossRef]
16. Wang, G.; Yin, X.; Pei, X.; Shi, C. Depth estimation for speckle projection system using progressive reliable points growing matching. *Appl. Opt.* **2013**, *52*, 516–524. [CrossRef]
17. Shi, C.; Wang, G.; Yin, X.; Pei, X.; He, B.; Lin, X. High-accuracy stereo matching based on adaptive ground control points. *IEEE Trans. Image Process.* **2015**, *24*, 1412–1423.
18. Supancic, J.S.; Rogez, G.; Yang, Y.; Shotton, J.; Ramanan, D. Depth-based hand pose estimation: Data, methods, and challenges. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
19. Oberweger, M.; Wohlhart, P.; Lepetit, V. Training a feedback loop for hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
20. Tang, D.; Taylor, J.; Kohli, P.; Keskin, C.; Kim, T.K.; Shotton, J. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
21. Ye, Q.; Yuan, S.; Kim, T.K. Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation. In Proceedings of the The European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
22. Guo, H.; Wang, G.; Chen, X.; Zhang, C.; Qiao, F.; Yang, H. Region Ensemble Network: Improving Convolutional Network for Hand Pose Estimation. In Proceedings of the 24th IEEE International Conference on Image Processing, Beijing, China, 12–17 September 2017.
23. Chen, X.; Wang, G.; Guo, H.; Zhang, C. Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation. *arXiv* **2017**, arXiv:1708.03416.
24. Wang, G.; Chen, X.; Guo, H.; Zhang, C. Region Ensemble Network: Towards Good Practices for Deep 3D Hand Pose Estimation. *J. Vis. Commun. Image Represent.* **2018**, *55*, 404–414. [CrossRef]
25. Chen, X.; Wang, G.; Zhang, C.; Kim, T.K.; Ji, X. SHPR-Net: Deep Semantic Hand Pose Regression from Point Clouds. *IEEE Access* **2018**, *6*, 43425–43439. [CrossRef]
26. Motion, L. Leap Motion Controller. 2015. Available online: https://www.leapmotion.com (accessed on 2 December 2018).
27. Keselman, L.; Iselin Woodfill, J.; Grunnet-Jepsen, A.; Bhowmik, A.; Gupta, M.; Jauhari, A.; Kulkarni, K.; Jayasuriya, S.; Molnar, A.; Turaga, P.; et al. Intel RealSense Stereoscopic Depth Cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Venice, Italy, 21–26 July 2017.

28. De Smedt, Q.; Wannous, H.; Vandeborre, J.P. Skeleton-based Dynamic hand gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016.
29. De Smedt, Q.; Wannous, H.; Vandeborre, J.P. 3D Hand Gesture Recognition by Analysing Set-of-Joints Trajectories. In Proceedings of the International Conference on Pattern Recognition (ICPR)/UHA3DS 2016 Workshop, Cancun, Mexico, 4 December 2016.
30. Boulahia, S.Y.; Anquetil, E.; Multon, F.; Kulpa, R. Dynamic hand gesture recognition based on 3D pattern assembled trajectories. In Proceedings of the 7th IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA 2017), Montreal, QC, Canada, 28 November–1 December 2017.
31. Caputo, F.M.; Prebianca, P.; Carcangiu, A.; Spano, L.D.; Giachetti, A. Comparing 3D trajectories for simple mid-air gesture recognition. *Comput. Gr.* **2018**, *73*, 17–25. [CrossRef]
32. Núñez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Vélez, J.F. Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognit.* **2018**, *76*, 80–94. [CrossRef]
33. Ma, C.; Wang, A.; Chen, G.; Xu, C. Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network. *Vis. Comput.* **2018**, *34*, 1053–1063. [CrossRef]
34. De Smedt, Q.; Wannous, H.; Vandeborre, J.P.; Guerry, J.; Le Saux, B.; Filliat, D. SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset. In Proceedings of the 10th Eurographics Workshop on 3D Object Retrieval, Lyon, France, 23–24 April 2017.
35. Chen, X.; Guo, H.; Wang, G.; Zhang, L. Motion Feature Augmented Recurrent Neural Network for Skeleton-based Dynamic Hand Gesture Recognition. In Proceedings of the 24th IEEE International Conference on Image Processing (ICIP), Beijing, China, 7–20 September 2017.
36. Cheng, H.; Yang, L.; Liu, Z. Survey on 3D Hand Gesture Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 1659–1673. [CrossRef]
37. Asadi-Aghbolaghi, M.; Clapés, A.; Bellantonio, M.; Escalante, H.J.; Ponce-López, V.; Baró, X.; Guyon, I.; Kasaei, S.; Escalera, S. Deep learning for action and gesture recognition in image sequences: A survey. In *Gesture Recognition*; Springer: Belin, Germany, 2017; pp. 539–578.
38. Cheok, M.J.; Omar, Z.; Jaward, M.H. A review of hand gesture and sign language recognition techniques. *Int. J. Mach. Learn. Cybern.* **2017**. [CrossRef]
39. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based human motion recognition with deep learning: A survey. *Comput. Vis. Image Understand.* **2018**, *171*, 118–139. [CrossRef]
40. Wang, C.; Liu, Z.; Chan, S.C. Superpixel-based hand gesture recognition with kinect depth camera. *IEEE Trans. Multimed.* **2015**, *17*, 29–39. [CrossRef]
41. Koller, O.; Ney, H.; Bowden, R. Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
42. Wang, P.; Li, Z.; Hou, Y.; Li, W. Action recognition based on joint trajectory maps using convolutional neural networks. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016.
43. Hou, Y.; Wang, S.; Wang, P.; Gao, Z.; Li, W. Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition. *IEEE Access* **2018**, *6*, 2206–2219. [CrossRef]
44. Zhu, G.; Zhang, L.; Shen, P.; Song, J. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access* **2017**, *5*, 4517–4524. [CrossRef]
45. Zhang, L.; Zhu, G.; Shen, P.; Song, J.; Shah, S.A.; Bennamoun, M. Learning spatiotemporal features using 3D CNN and convolutional lstm for gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017.
46. Wan, J.; Guo, G.; Li, S.Z. Explore efficient local features from RGB-D data for one-shot learning gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1626–1639. [CrossRef]
47. Wan, J.; Ruan, Q.; Li, W.; An, G.; Zhao, R. 3D SMoSIFT: three-dimensional sparse motion scale invariant feature transform for activity recognition from RGB-D videos. *J. Electron. Imaging* **2014**, *23*, 023017. [CrossRef]
48. Wan, J.; Ruan, Q.; Li, W.; Deng, S. One-shot learning gesture recognition from RGB-D data using bag of features. *J. Mach. Learn. Res.* **2013**, *14*, 2549–2582.

49. Wan, J.; Zhao, Y.; Zhou, S.; Guyon, I.; Escalera, S.; Li, S.Z. Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016.
50. Köpüklü, O.; Köse, N.; Rigoll, G. Motion Fused Frames: Data Level Fusion Strategy for Hand Gesture Recognition. *arXiv* **2018**, arXiv:1804.07187
51. Boulahia, S.Y.; Anquetil, E.; Kulpa, R.; Multon, F. HIF3D: Handwriting-Inspired Features for 3D skeleton-based action recognition. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016.
52. Sadanandan, S.K.; Ranefall, P.; Wählby, C. Feature Augmented Deep Neural Networks for Segmentation of Cells. In Proceedings of the European Conference on Computer Vision Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016.
53. Egede, J.; Valstar, M.; Martinez, B. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Amsterdam, The Netherlands, 8–10 October 2017.
54. Manivannan, S.; Li, W.; Zhang, J.; Trucco, E.; McKenna, S.J. Structure Prediction for Gland Segmentation with Hand-Crafted and Deep Convolutional Features. *IEEE Trans. Med. Imaging* **2018**, *37*, 210–221. [CrossRef]
55. Wang, S.; Hou, Y.; Li, Z.; Dong, J.; Tang, C. Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier. *Multimed. Tools Appl.* **2016**, *77*, 18983–18998. [CrossRef]
56. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
57. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923. [CrossRef]
58. Liang, H.; Yuan, J.; Thalmann, D. Parsing the hand in depth images. *IEEE Trans. Multimed.* **2014**, *16*, 1241–1253. [CrossRef]
59. Chen, H.; Wang, G.; Xue, J.H.; He, L. A novel hierarchical framework for human action recognition. *Pattern Recognit.* **2016**, *55*, 148–159. [CrossRef]
60. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 16 July 2018).
61. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
62. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
63. Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
64. Lai, K.; Yanushkevich, S.N. CNN + RNN Depth and Skeleton based Dynamic Hand Gesture Recognition. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
65. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* **2015**, *45*, 1340–1352. [CrossRef] [PubMed]
66. Nguyen, B.P.; Tay, W.L.; Chui, C.K. Robust Biometric Recognition From Palm Depth Images for Gloved Hands. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *45*, 799–804. [CrossRef]
67. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.