Deep Predictive Neural Network: Unsupervised Learning for Hand Pose Estimation

Jamal Banzi, Isack Bulugu, and Zhongfu Ye

Abstract—**The discriminative approaches for hand pose estimation from depth images usually require dense annotated data to train a supervised network. Additionally, generative methods depend on temporal information in generating candidate poses which can be trapped due to local minima during the optimization process. Different from these methods, we propose a hybrid two-stage deep predictive neural network approach that performs predictive coding of image sequences of hand poses in order to capture latent features underlying a given image. Firstly, we train a deep convolutional neural network (CNN) for direct regression of hand joints position. Secondly, we add an unsupervised error term as a part of the recurrent architecture connected with predictive coding portion. An error regression term (ERT) ensures minimal residual errors of the estimated values while the predictive coding portion allows training of the network without the supervision of image sequences, so no dense annotation of data is required. We conduct a complete experiment using two challenging public datasets, ICVL and NYU. Using the ICVL datasets, our approach improved accuracy over the current state of the art methods with an average error joint of 7.5mm. We also achieve 12.2mm average error joint on NYU dataset which is the smallest error to be achieved on all state-of-art approaches.**

Index Terms—**Deep learning, hand pose estimation, joint regression, predictive neural networks.**

I. Introduction

Hand pose estimation from depth is the first step for virtual reality and human-computer interaction applications. It has been an active research area which attracted tremendous attention thanks to the advent of commercial short-range depth sensors such as Intel Real Sense, Prime Sense Carmine and Leap Motion. An accurate hand pose estimation provides a natural way for interaction between a user and a virtual space that achieves greater user experience.

Indeed, there is a growing interest in developing hand pose estimation systems [1], [2] Early approaches utilized augmentation of the hand using specialized hardware such as optical sensors [2] and data gloves [3] to achieve this mission. The approaches were able to demonstrate the viability of gesture control for applications tracking with a high degree of accuracy. However, these approaches were cumbersome e.g. requires many calibrations, expensive and are not the most natural way for users have to contact material outside their bodies to use the systems. This led to the development of passive augmentation, vision-based approaches providing low-cost computing using RGB cameras. To achieve natural interaction, no augmentation was required. The approaches directly extract hand features from RGB sequences to define a customized hand model or a representation and then uses minimization techniques to refine the final hand pose.

Compared with former approaches, vision-based methods were affordable and user-friendly. However, these approaches were not robust and stable enough for real-time application since RGB cameras are challenged by the varying light intensity and the complex background. Additionally, any RGB image provides only 2D information while hand shape topology in 3D space is the most important feature for hand pose estimation. For this reason, it is necessary to look for a better solution to alleviate these limitations.

Fortunately, the advent of depth sensors which were developed for body pose estimation brings about the possibility of accomplishing such a challenging task. Depth sensors provide adequate 3D information of a hand geometry, which simplifies hand detection and in turn enhances the stability of hand pose estimation systems.

On top of that, the fast development of hardware devices expands the computation power, which has led to the rise of deep learning. Deep learning approaches achieve the great success on image classification and analysis. Unquestionably, the utility of deep learning gives potential to hand pose estimation. In fact, recent approaches have confirmed that such deep learning methods outperformed the conventional optimization-based approaches.

On the other hand, the increased demand for a natural way and user-friendly interface to applications such as Human-computer interaction, remote surgery, virtual reality, sign language recognition, and video games also boosts the development of stable hand pose estimation systems.

Despite the fast progress of this field, hand pose estimation is still a difficult task due to some challenges one may face during the process of estimation. Before estimation of the hand pose, one needs to detect the location of a human hand in the input image.

During the hand detection stage, the position of the user's hand from a complex environment in the image is determined which brings about the following problems. Firstly, the

Manuscript received December 21, 2018; revised June 1, 2019. This work is supported by the Fundamental Research Funds for the Central Universities (Grant no. WK2350000002). The author acknowledges a fellowship by the Chinese Academy of Science and The World Academy of Science (CAS-TWAS).

J. Banzi is with the Department of Electronic Engineering and Information Science, School of Information Science and Technology, University of Science and Technology of China, 230026, Hefei city, Anhui Province, P.R China (e-mail: jbanzi@mail.ustc.edu.cn)

I. Bulugu is with the Department of Electronics and Tel. Engineering, University of Dar-es-salaam, Tanzania (e-mail: bulugu@mail.ustc.edu.cn).

Z. Ye is with the Department of Electronic Engineering, University of science and technology of China No. 96, 230026, Hefei city, Anhui Province, P.R China (e-mail: yezf@mail.ustc.edu.cn).

doi: 10.18178/ijmlc.2019.9.4.822

human hand may have various shape and size on account to different subjects, viewpoints and the distance from the sensor, which results in the high intraclass variations. Secondly, the unreliability of the hand detection results, which may cause the failure of the subsequent pose estimation stage. To eradicate this problem, the previous body of works have only employed some strong assumptions to detect the human hands. The work of [4], [5] assume that the human hands will just be the nearest object from the camera appearing in the input images. These assumptions have some faults, for example, failure in training multiple user's hands or a hand being in a complex environment. Recently, Tzu-Yang *et al.* [6] tried to make such system more reliable by combining hand detector and pose estimation system.

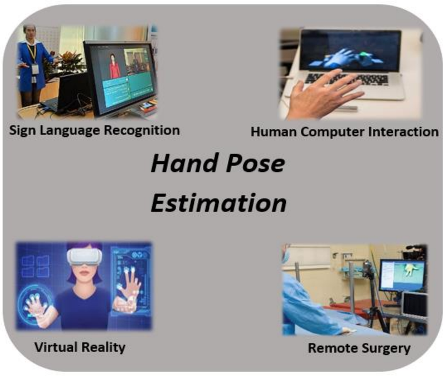

Fig. 1. Various application of human hand pose.

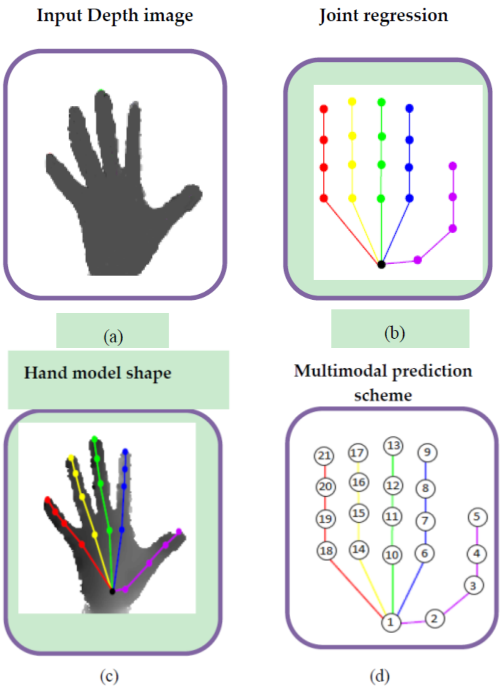

Fig. 2. Hand Model structure showing an overview of the input and output hand image. (a) Input depth frame (b) Hand model showing joint distribution. (c) Hand model fitted to hand shape (d) Our hand model with 21 joints and with 29 degrees of freedom.

The subsequent hand pose estimation stage is equivalently challenging. The hand is a small part of the body that produces self-occluded poses during estimation. The hand is also very dexterous, has many degrees of freedom and has fast movement. Due to this high degree of freedom, the domain of pose configurations becomes higher, and so pose estimation becomes difficult. Moreover, every pose definition requires a number of labels which make it difficult as a result. Labeling such a dataset is both hard and time-consuming which consequently makes training samples quite insufficient.

Inspired by the challenge of labeling large numbers of a training date we discuss above; this paper presents a deep predictive neural network algorithm (PredNet) that captures latent parameters of hand poses and corresponding depth images for estimating 3D hand pose. We use standard autoencoder (SAE) and ladder network (LN) for modeling the generative process of hand poses and the depth map respectively. By combining with an unsupervised error term as a part of the recurrent architecture, the predictive coding portion of the network was trained without the supervision of the image sequences, so no dense annotation of the data is required. We consider a one to one mapping between a depth map and a hand pose. In this way, a latent hand poses parameter and a latent depth map parameter can be shared. Having a shared parameter is highly beneficial since a point sampled in either latent space can be defined both as 3D pose through SAE's decoder or a depth map via LN generator.

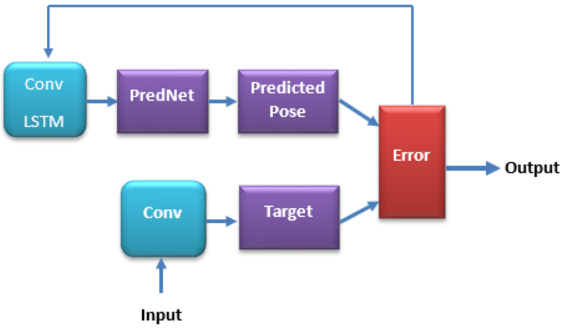

Fig. 3. General overview of the proposed framework.

The general overview of the proposed framework is presented in Fig. 3. The core idea is to learn the basic implicit hand structure and transformations a hand can possibly undergo. To achieve this, our model is trained to estimate multiple frames from the image sequence while considering each initial estimation as an actual input and recursively iterating. Additionally, we include an ERT that provides a correction vector to improve the estimation accuracy. The ERT allows the network to learn from its own mistakes and correct them automatically giving an output with much greater accuracy.

The gist of our approach is that our estimation does not require the labeled training examples, and therefore to the best of our knowledge we are the first to present a fully unsupervised approach for hand pose estimation. Unlike previous generative methods for hand pose estimation which depend on temporal information which can be trapped due to local minima and subsequently cause optimization failures, herein, our method works independently from temporal information. Instead, we train the network to learn the internal representation of the hand joint positions that are well suited to the subsequent estimation and decoding of the latent hand parameters.

II. Related Work

There is a large body of works focusing on hand pose estimation in the past few years and we refer the reader to [1], [2] for an overview. This section highlights recent approaches and examine their relevance to our approach. Generally, hand pose estimation aims to improve Human-Computer Interaction through augmentation free interaction. Due to the importance of this interaction, the focus should be offering a reduced cost of production, improve accuracy and extends research tremendously by adopting new methods (e.g. vision-based methods) to address the existing challenges.

The early approaches were based on generative methods, whereby a multiple hand model is synthesized and matched with the best fit of the observed data. Then it defines the energy function to quantify the discrepancy between the synthesized and observed images and optimize the function to obtain the final pose. However, there was an overhead in rendering candidate poses with these generative approaches [7]. Martin *et al.* [8] addressed this challenge by recovering 3D hand pose from a monocular image using model-based Bayesian inference method. The method synthesized the corresponding hand silhouette projection in the image plane measuring its likelihood in a given generative model for background and the hand skin pixels. Hand pose is then iteratively refined through minimization of negative log likelihood. Qian *et al.* [9] responded to the model-based challenges by sub-sampling the observed data and uses spheres to model the hand. Then Melax *et al.* [10] used a rigid body simulation for optimization. Point to surface constraints works similarly to ICP in the optimization process. As for many generative based optimization methods, ICP falls into local minima and this account for a reason why learning methods seem to be preferable.

With the upsurge in the availability of short-range depth sensors, a number of discriminative learning approaches have been also proposed for hand pose estimation. Discriminative approach infers the hand pose directly from the observed visual data without depending on a model. Hand pose is then recovered from the single frame through forest regression or classification techniques [11].

Forests were applied to perform offset regression of the bodies joint using traditional Hough regression style voting [12]. Kinematic limitations would be implicitly modeled but would break for unseen examples hence the need for large datasets. Afterward [13] demonstrated the efficiency of forest regression using adaptive hierarchical classification approach, regressing all joints in one channel of forest per frame. However, the approach was susceptible to error propagation leading to wrongly estimated poses. Poier *et al.* [14] use a random regression forest to estimate hand joint distribution, and then builds a more reliable quality measurement scheme based on consistency between generated joint locations and predicted distribution. This approach distinguishes the joint estimation and model fitting in two stages. First, employed a learning regressor to deliver multiple initial hypotheses for the 3D position of each joint. Then, the kinematic parameters of a 3D hand model are found by deliberately exploiting the inherent uncertainty of the inferred joint proposals.

On the other hand, classification techniques are recently employed in machine learning approach for discriminative modeling. Classification techniques make use of classifiers or artificial neural networks and deep learning to classify, regret, and estimate a hand joint positions for hand estimation. Tompson *et al.* [15] predict joint locations with the convolutional neural network (CNN). CNN also used for feature extraction and generates small heat maps for joint location. Joints were converted to hand skeleton using Inverse Kinematics (IK) process. However, Tompson's approach predicted only 2D locations of joints and was unable to predict hidden joints. Moreover, accuracy is dependent to heat map resolution which means for every pixel, heat map has to be created as CNN has to be evaluated at each pixel location. This is computationally deprived. Guo *et al.* [16] recovered 3D hand pose, using tree-structured Region Ensemble Network (REN) which partitions the convolution outputs into regions and integrate results from multiple regressors on each region. Sridhar *et al.* [17] used a pixel classification to predict the joint position and then applies a similarity function to a model fitting and compare directly the generated joint locations to the predicted joint locations. Recently [18], [19] presented three neural network architecture that trains a feedback loop to predict 3D joint locations of a hand given a depth. The architecture combines generative network, a discriminative pose estimation network and a pose update network. The first two architectures estimate the joint locations and the third architecture refines the joint location estimates. To improve the accuracy of the location estimates, refinement step was iterated several times by centering the network on the location predicted at the previous iteration. Although the approach successfully improves localization accuracy and speed, training this architecture is, however, a complex difficult task.

III. Method Overview

The proposed hand pose estimation system aims to reduce the mutual failure cases of both generative and discriminative approaches to improve the accuracy of the estimation. We consider hand pose estimation problem as a statistical learning problem of a set of depth images. We combine generative and discriminative methods for generation of hand poses and the regression of the hand joint positions respectively. As a consequence, we use two networks to accomplish this mission, one for pose estimation and the other for joint regression. We pre-train each network separately to capture the latent features of an individual domain. We then learn a mapping between the two latent spaces. The complete hand pose estimation network is then trained end to end for hand pose estimation task.

The functional description of our approach is shown in Fig. 4. The main part of the architecture is covered by predictive neural networks which is the extension of the work of [20], these networks learn to predict positions of hand joints in an image sequence, with each layer in the network making local prediction and only forwarding deviations from those predictions to the subsequent network layers. These networks are able to robustly predict hand movements and that in so doing, the networks learn internal representations that are useful for decoding latent hand parameters internal representation that supports hand pose estimation with fewer training views. We use the latent variable as a way to combine supervised and unsupervised learning in a principled way.

The purpose of combining an auxiliary task is to help train the neural network as suggested by [21]. By sharing the hidden representation among more than one task, the network generalizes better. The idea of using unsupervised learning to complement supervision is not new. Different methods utilized different choices for unsupervised tasks, for example, reconstruction of the inputs at every level of the model [22], or classification of each input sample into its own class. Other methods have been able to simultaneously apply both unsupervised and supervised learning [23], often these unsupervised auxiliary tasks are only applied as pre-training, proceeded by normal supervised learning. However, in a complex task such as estimation of the human hand pose, there is much more structure in the input than it can be represented, and unsupervised learning cannot by definition know what will be useful for the task at hand. To solve that problem, this paper proposed the ladder networks where the auxiliary task is to denoise representations at every level of the model. The model structure is an autoencoder with skip connections from the encoder to the decoder and the learning task is similar to that in denoising autoencoders but applied to every layer, not just the inputs. Using skip connections, the decoder can recover any details discarded by the encoder.

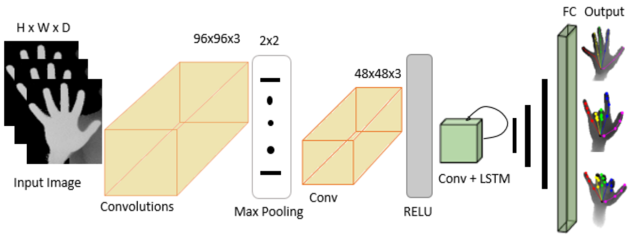

Fig. 4. A general overview of the deep prednet used for joint regression.

A. Unsupervised Learning

Herein, the ladder network Fig. 5 will decode the fully connected encoder in order to learn meaningful abstractions which help in denoising. We use the denoising function, Gaussian latent variable to support unsupervised learning.

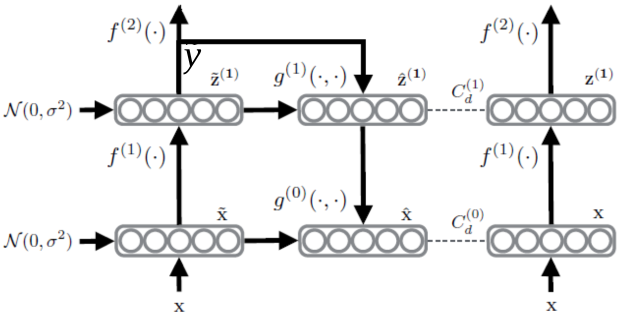

Fig. 5. An illustration of the operation of the ladder network showing feedforward path with sharing mappings.

Consider the supervised cost C_c, as the average negative log probability of the noisy output y matching the target $t(n)$ given the inputs $x(n)$.

$$C_c = -\frac{1}{N}\sum_{n=1}^{N}\log P(y = t(n) \mid x(n)) \tag{1}$$

This structure regularizes supervised learning.

Now then, when designing a suitable decoder to support unsupervised learning, a parameterization that supports the optimal denoising of gaussian latent variables is chosen. To derive the chosen parameterization and justify why it supports Gaussian latent variables, let assume that the noisy value of one latent variable $\tilde{z}$ that we want to denoise has the form of $\tilde{z} = z + n$, where z is the clean latent variable value that has Gaussian distribution with variance σ_z^2, and n is the Gaussian noise with variance σ_n^2. We will estimate $\hat{z}$, a denoised version of $\tilde{z}$, so that the estimate minimizes the squared error of difference to the clean latent variable values z. It can be shown that the functional form of $\hat{z} = g(\tilde{z})$ has to be linear in order to minimize the denoising cost. Therefore, the result will be a weighted sum of the corrupted $\tilde{z}$ and a prior μ. The weight w of the corrupted $\tilde{z}$ will be a function of the variance of z and n according to:

$$w = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_n^2} \tag{2}$$

The denoising function will therefore be in the form of

$$\hat{z} = g(\hat{z}) = w * \tilde{z} + (1-w) * \mu = (\tilde{z} - \mu) * w + \mu$$

w and μ can be trainable parameters of the model, where the model would learn some estimate of the optimal weighing w and μ. The final unsupervised denoising cost function C_d is thus

$$C_d = \sum_{t=0}^{L}\lambda_l C_d^l = \sum_{l=0}^{L}\frac{\lambda_l}{Nm_l}\sum_{n=1}^{N}\| z^l(n) - \hat{z}_N^l(n)\|^2 \tag{3}$$

where m_l is the layer's width, n the number of training samples, and the hyperparameter λ_l a layer-wise multiplier determining the importance of the denoising cost.

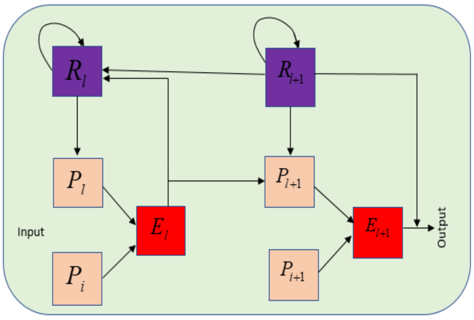

Fig. 6. Operation of an error regression term.

B. Error Regression Term

Initially, the input image sequence enters the model and the local estimation of this input is made Fig. 6. This estimated input is subtracted from the actual input and passed along to the next layer. The network takes the difference from the input P_i and the estimated hand (P_l), and output an error regression (E_l) which splits into rectified positive and negative error populations. The error E_l, is then passed forward through convolutional layer to become the input of the next layer P_{i+1}. The representation layer (R_l) receives the copy of error signal E_l, together with the top-down input from the representation layer of the next level layer of the network R_{i+1}.To improve the accuracy of the location estimates of the hand joints, this step is iterated several times while forwarding an error to the input and allow the network to learn from its previous mistakes.

IV. Experiment

This section describes the experiments which were conducted to validate the proposed hand pose estimation system. Prior to the discussion of the experiments, a brief description of the used hand pose datasets is given. Two of them are publicly available and one was collected by the authors.

A. Dataset Creation

Herein we consider three datasets which are used in the experiments. Two of these datasets are publicly available and have been used by many recent research works. Therefore, a thorough comparison with other bodies of work will be conducted on these two datasets to evaluate the performance of our system. The third dataset is generated by the authors, it contains 10 participants with different hand shapes (6 male, 4 females, aged between 20 to 30. In the experiment, we collected a sequence from 10 different subjects with the varying hand sizes by allowing each subject to make various hand poses with an illustration of 29 different postures. The hand shape has 21 joints position rotating with 29 degrees of freedom. Each sequence is sampled at 3fps generating a total of 20K images with 96x96 pixels in size. We use 1.6K images for both validation and testing. Estimations generated by PredNet are presented in Fig. 10.

B. Training

We first trained the network using synthetic images since for these we have access to the underlying generative model and all the latent parameters. The input for the network is the cropped hand shape using ground truth joint location. As stated in Section IV.A, images were 96×96 in size and in grayscale with values normalized between 0 and 1. Focusing on image sequence data, for layer zero, the target is set to the actual sequence itself i.e. $P_0^t = x_t \forall t$. The target for higher layers P_l^t for $l > 0$, are computed by the convolution through the error unit from the layer below, P_{l-1}^t, succeeding by rectified linear units ReLU activation and the max-pooling. Therefore, the targets will generate an abstraction of features as the error propagates up the network. Specifically, a convolutional LSTM unit is used as it can replace a dense multiplication matrix in a standard LSTM with a sparse convolutional matrix. As in the case of [24], the hidden state R_l^t is updated according to R_l^{t-1}, E_l^{t-1}, and R_{l+1}^t so its spatial dimension is up-sampled by the pooling effect present in the feedforward path. The estimations P_l^t at a later time *t* are made through convolution of the R_l^t stack followed by a ReLU non-linearity. For pixel layer zero, P_l^t is also passed through a saturating non-linearity set at a maximum pixel level. The error term E_l^t is then calculated from the difference between P_l^t and P_i^t, then splits by ReLU-activation into positive and negative estimation errors, which are concatenated along the feature dimension. The model is trained to minimize the weighted sum error. These error units consist of subtraction of an input and estimated output followed by ReLU activation which corresponds to L_1 error estimation. The model was trained to estimate hand joint position post learning its previous position. The loss was taken as the sum of the firing rates of the error neurons in the zeroth pixel layer at time step 2-10. A random hyperparameter search was conducted over fourth- and fifth-layer models. The model is chosen with respect to the performance on the validation test.

Our network model consists of 5 layers with 2×2 filter sizes for all convolutions and stack size per layer of (1,24,48,96,192). Model weight was optimized using Adam algorithm [25] with all parameters set to default values. The PredNet Model was implemented with python library using Theano [26] and Keras and is trained on a desktop computer with python 3.4 on window platform.

V. Evaluation

We evaluate our approach on the two public benchmark datasets for hand pose estimation. The NYU datasets and ICVL datasets. For comparison with other methods, we focus on the works that are published recently to compare the state-of-the-art with our method. Different evaluation metrics have been used in the literature for hand pose estimation. We report the values stated in papers or measured from graphs if provided, and or plot-relevant graphs for comparison.

A. Evaluation Metrics

We employ two different commonly used metrics to evaluate accuracy.

- The fraction of sample error distance within a threshold. Here we measure the fraction of success frames whose error distance for each joint is less than a certain threshold. This is the most challenging evaluation criterion since a single mistaken joint may cause the entire hand pose judgment to be considered a failure.
- Mean error distance of different joints and their average. This is the most commonly used criteria in the literature of hand pose estimation because of its simplicity of evaluation; using it allows comparison with many contending baselines.

B. Quantitative Evaluation with the NYU Dataset

We compare our method to four state-of-the-art methods: Oberweger *et al.* [18], Deng *et al.* [27], Zhou *et al.* [28] and Tompson *et al.* [15]. The results of testing examples using max joint error below the threshold are shown in Fig. 7 and the values are presented in Table I. Depicted from the figure, our proposed method outperformed all other approaches. The performance is nearly comparable to [18] when the threshold is very high i.e. the requirements for estimation quality are low. However, when the requirement for estimation accuracy is very tough i.e. a low threshold, our method performed better than all the contending methods. The error distances of different joints are shown in Fig. 8. And the quantitative results of the average error distance are shown in Table II. It shows that our method has less error in the average of all joints than all other approaches.

TABLE I: PERCENTAGE OF FRAMES IN TEST EXAMPLES OVER DIFFERENT THRESHOLDS ON NYU DATASETS

Threshold(mm)	≲ 20	≲ 50
Ours	35%	90%
Deng *et al.* [27]	20%	75%
Oberweger *et al.* [18]	12%	65%
Zhou *et al.* [28]	10%	59%
Tompson *et al.* [15]	5%	51%

TABLE II: QUANTITATIVE ERROR JOINT RESULTS OF DIFFERENT METHODS ON NYU DATASETS

Method	Avg. error distance
Oberweger *et al.* [18]	19.8
Tang *et al.* [13]	18.5
Deng *et al.* [27]	17.6
Zhou *et al.* [28]	16.9
Ours	12.2

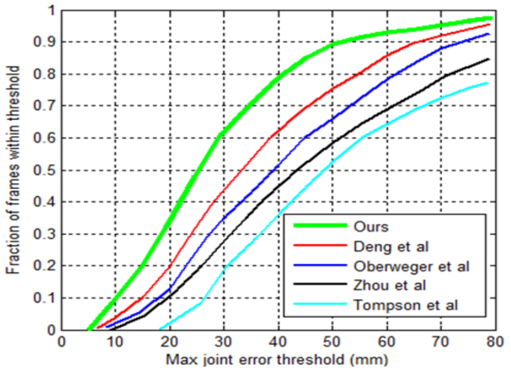

Fig. 7. Comparison with the state-of-the-art on NYU dataset: Percentage of success frames in the test examples error threshold.

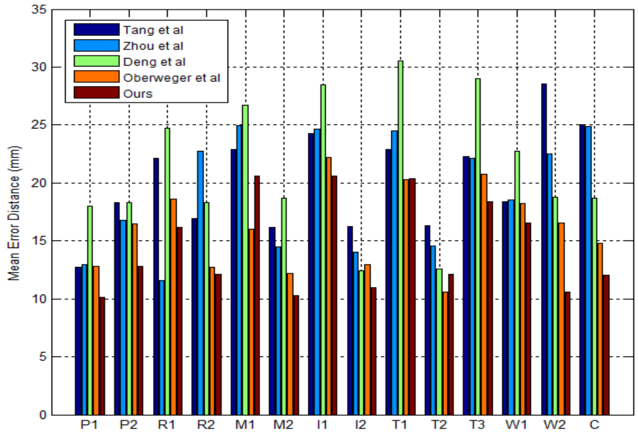

Fig. 8. Comparison with the state-of-the-art on NYU dataset: Mean error distance for each joint across all the test examples.

C. Quantitative Evaluation with ICVL Datasets

The ICVL dataset is a public dataset released by Imperial College Vision Lab (ICVL) and captured by an Intel Creative Interactive Gesture Camera. The ICVL dataset contains a training set of about 180k depth frames having various hand poses recorded from 10 different subjects. The test set contains two sequences with approximately 700 frames each. Each hand pose has 16 annotations of joints. In contrast to the traditional structured light camera, this camera can capture depth image with a low noise level via the advanced time-of-flight component. The depth images have high quality with no or very few missing depth values and sharp outlines with an only slight noise. Hence this dataset is very appropriate for hand pose estimation systems. However, the pose variability of this dataset is limited as reported in [18] and other publications have reported inaccurate annotations as discussed in [31].

TABLE III: PERCENTAGE OF FRAMES IN THE TEST EXAMPLES OVER DIFFERENT THRESHOLDS ON ICVL DATASETS

Threshold (mm)	≲ 20	≲ 50
Ours	70%	98%
Deep prior++ [19]	68%	95%
Zhou *et al* [28]	62%	96%
Deng *et al* [27]	55%	97%
Tang *et al* [14]	50%	91%
Guo *et al* [16]	20%	85%

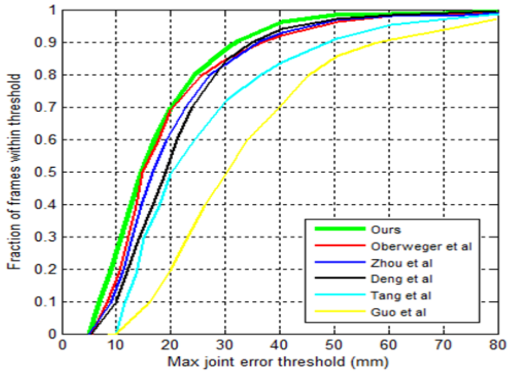

Fig. 9. Comparison with the state-of-the-art on ICVL dataset.: Percentage of success test frames in the test examples error threshold.

We compare our experimental results on the ICVL hand posture dataset with five state-of-the-art techniques: Deep prior ++ [19], Tang *et al.* 2014[13], Zhou *et al.* [28], Deng *et al.* [27], and Guo *et al.* [16]. Fig. 7 illustrates the results, showing our method achieves consistent improvement overall error thresholds compared to other methods using the first evaluation criteria. For example, if we require all the error distance of joints to be below an error threshold of 50mm, the yield of our approach is almost 98% while other contending approaches achieve only 80% or even less with the exception of deep prior. The results affirm that our system is robust and practical for real-time human-computer interaction applications. We also present results using the error distance of different joints and their average on the ICVL hand posture dataset as shown in Fig. 9 with quantitative results listed in Table IV. It should be noted that we only compare the mean error distance for 11 joints in our second evaluation criteria since many published works [19], Tang *et al.* [14] only provide the results for 11 joints. The results show the supremacy of our system by producing the lowest error for 15 joints out of 16 joints, achieving 7.5mm average error distance. Some of the results in the list are not shown in the graph due to limited space. The results of deep prior++ [19] are obtained from their paper; their results and our results are quite close using both the first and second evaluation criteria. However, when considering each joint individually, our method is better for most joints. Although our data for a few joints is the same as the deep prior+ [19] results, our results have a better average mean error distance than all other techniques, thereby indicating that the combination of deep learning and predictive coding can greatly improve the performance of hand pose estimation systems.

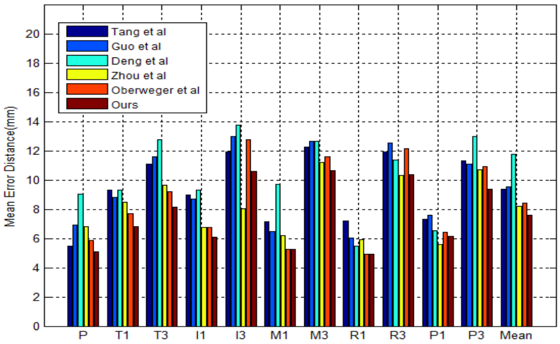

Fig. 10. Comparison with the state-of-the-art on ICVL dataset: Mean error distance for each joint across all the test examples.

TABLE IV: QUANTITATIVE ERROR JOINT RESULTS OF DIFFERENT METHODS ON ICVL DATASETS

Method	AVG error distance
Tang *et al.* [14]	8.9
Deng *et al.* [27]	8.7
Deep prior ++ [19]	8.4
Zhou *et al.* [28]	8.2
Guo *et al.* [16]	9.0
Ours	7.5

TABLE V: EFFICIENCY COMPARISON OF CONTENDING APPROACHES SHOWING THE DEVICE USED FOR COMPUTATION AND THE FRAME RATE OF EACH METHOD

Method	Time(fps)	Device used
PredNet	63.5	CPU
Tang *et al* [14]	63	CPU
Melax *et al* [10]	60	CPU
Sharp *et al* [29]	30	GPU
Qian *et al* [9]	25	CPU
Xu *et al* [30]	12	CPU

II. QUALITATIVE ANALYSIS

We present qualitative results for hand pose estimation produced by our network trained on real data in Fig. 11. We also present the visualization comparison with the state-of-art methods on NYU datasets on Fig. 12. We report some failed cases caused by the wrong hand joint location in the second row which represents our method. However, it can be observed in the fail cases that despite the wrong location of some joint positions, the kinematic structure of the hand is preserved. Furthermore, our method outperformed all other methods in terms of efficiency and estimation accuracy. We present the efficiency performance of our approach in comparison of the state of art approaches in Table V.

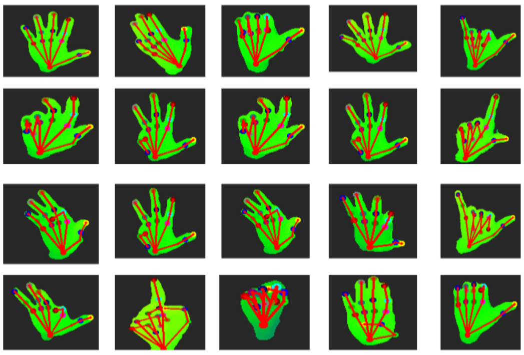

Fig. 11. Visualization of estimated hand skeletons produced by our network trained on real data.

III. CONCLUSIONS

In this paper, we propose a novel approach for a 3D-hand pose estimation system which can accurately estimate hand poses. The experimental results show that our system can estimate human hand pose with over 90% accuracy and achieve about 7.5mm of an average error distance. This achievement is attributed to the following features.

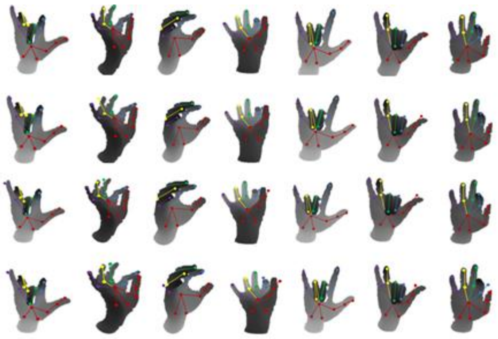

Fig 12. Visual comparison of the estimated hand skeletons of our system and the comparison with state-of-art-works on the NYU hand pose dataset. The first row is the Ground truth, the second row is Our method, the third row is Deng *et al.*, and the last row is for Oberweger *et al.*

Firstly, a powerful model architecture, the PredNet which is capable of learning the latent hand features underlying a given image and estimate hand joints location with much more perfection.

Secondly, we present an unsupervised learning paradigm to expand the utility of deep learning on human hand pose estimation by incorporating an unsupervised error term as a part of the recurrent architecture, the predictive coding portion of the network was trained without the supervision of image sequences.

Thirdly, in order to improve the estimation accuracy, our networks perform feed-forward generation of estimation errors after every local estimation which is used as input to the subsequent network layer. This allows the network to learn from its own previous mistakes and correct them automatically and therefore increases accuracy.

In the future, we plan to further improve estimation accuracy, develop a deeper network to accommodate a larger amount of training data with many more learned features and extend our architecture to allow for more complex hand configurations. Our system can be integrated with some advanced applications for human-computer interaction to provide the most natural way of interaction between users and cyberspace to achieve a better user experience.

References

[1] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, pp. 52-73, 2007.

[2] E. Barsoum, "Articulated hand pose estimation review," *arXiv preprint arXiv:1604.06195*, 2016.

[3] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Real-time hand pose estimation using depth sensors," *Consumer Depth Cameras for Computer Vision*, pp. 119-137, 2013, Springer, London.

[4] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

[5] J. F. Banzi, Z. Ye, and I. Bulugu, "A novel hand pose estimation using discriminative deep model and transductive learning approach for occlusion handling and reduced discrepancy," in *Proc. 2016 2nd IEEE International Conference on Computer and Communications*, pp. 347-352, October 2016.

[6] T.-Y. Chen, P.-W. Ting, M.-Y. Wu, and L.-C. Fu, "Learning a deep network with the spherical part model for 3D hand pose estimation," *Pattern Recognition*, vol. 33, p. 3203, 2018.

[7] P. Krejov, A. Gilbert, and R. Bowden, "Combining discriminative and model-based approaches for hand pose estimation," in *Proc. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, May 2015, vol. 1, pp. 1-7.

[8] M. D. L. Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793-1805, 2011.

[9] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Real-time and robust hand tracking from depth," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1106-1113.

[10] S. Melax, L. Keselman, and S. Orsten, "Dynamics based 3D skeletal hand tracking," *Proceedings of Graphics Interface*, Canadian Information Processing Society, pp. 63-70, May 2013.

[11] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 824-832, 2015.

[12] D. Tang, T. H. Yu, and T. K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in *Proc. 2013 IEEE International Conference on Computer Vision*, pp. 3224-3231, December 2013.

[13] D. Tang, H. J. Chang, A. Tejani, and T. K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3786-3793.

[14] G. Poier, K. Roditakis *et al.*, "Hybrid one-shot 3D hand pose estimation by exploiting uncertainties," *arXiv preprint arXiv:1510.08039*, 2015.

[15] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (ToG)*, vol. 33, no. 5, p. 169, 2014.

[16] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," *arXiv preprint arXiv:1702.02447*, 2017.

[17] S. Sridhar, A. Oulasvirta, and C. Theobalt, "Interactive markerless articulated hand motion tracking using RGB and depth data," in *Proc. 2013 IEEE International Conference on Computer Vision*, December 2013, pp. 2456-2463.

[18] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit, "Efficiently creating 3D training data for fine hand pose estimation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4957-4965.

[19] M. Oberweger and V. Lepetit, "Deep prior++: Improving fast and accurate 3d hand pose estimation," in *Proc. ICCV Workshop*, August 2017, vol. 840, p. 2.

[20] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104*, 2016.

[21] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," *Advances in Neural Information Processing Systems*, pp. 3546-3554, 2015.

[22] M. A. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in *Proc. ICML*, 2008, pp.792–799.

[23] I. Goodfellow, Y. Bengio, and A. C. Courville, "Large-scale feature learning with spike and-slab sparse coding," in *Proc. ICML*, pp. 1439–1446, 2012.

[24] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Proc. International Conference on Machine Learning*, June 2015, pp. 843-852.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] R. Al-Rfou, G. Alain, A. Almahairi *et al.*, "Theano: A python framework for fast computation of mathematical expressions," *arXiv preprint arXiv:1605.02688*, pp. 472-473, 2016.

[27] X. Deng, S. Yang, Y. Zhang, P. Tan, L. Chang, and H. Wang, "Hand3d: Hand pose estimation using 3d neural network," *arXiv preprint arXiv:1704.02224*, 2017.

[28] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, "Model-based deep hand pose estimation," *arXiv preprint arXiv:1606.06854*, 2016.

[29] T. Sharp, C. Keskin *et al.*, "Accurate, robust, and flexible real-time hand tracking," in *Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3633-3642.

[30] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. the IEEE International Conference on Computer Vision*, 2013, pp. 3456–3462.

[31] G. Rogez, Y. Yang *et al.*, "Depth-based hand pose estimation: Methods, data, and challenges," *arXiv Prepr. arXiv1504.06378*, pp. 1868–1876, 2015.

Jamal Banzi received his bachelor of science in information systems degree from the University of Dodoma, Tanzania in 2011. He then joined Tianjin University of Technology and Education, China in 2012 where he graduated with a master's degree of science in signal and information processing engineering in 2015. He is currently with the Department of Electronic Engineering and Information Science, University of Science and Technology of China for his Ph.D. studies. His research interest includes computer vision and pattern recognition, deep learning, and human-computer interaction.

Isack Bulugu received his B.Sc. degree in electronics in 2007 from University of Dar-es-salaam, Tanzania. He received master's degree and Ph.D. in signal and information processing engineering from Tianjin University of Technology and Education and University of Science and Technology of China in 2014 and 2018 respectively. He is currently working at the University of Dar-es-Salaam. His research interests are image processing, hand gesture recognition, and artificial intelligence.

Zhongfu Ye is a professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He obtained his Ph.D. in signal and information processing from the University of Science and Technology of China in 1995. He is the author of over 200 academic papers on national and international major journals and important international conferences like IEEE, IET, PASP, PASA, Acta Electronica Sinica. He has been teaching and researching in USTC since 1995 and was appointed as a full professor in 2000. He is currently the director of signal statistics and processing research center of USTC, head of the discipline of signal and information processing of USTC, Ph.D. supervisor of Institute of Electronics of CAS, Member of the editorial board of Journal of Communication, Acta Armament, and Journal of Data Acquisition and Processing. He is also a reviewer for IEEE, IET, international conferences, and national academic journals, member of the committee of Instrument Science and Control Technology, Chinese Association of Higher Education, and a member of the Committee of Photonics, Chinese Society of Astronautics. His current research interest includes array signal processing, speech/audio signal processing, image, and processing.