A comparison between end-to-end approaches and feature extraction based approaches for Sign Language recognition

Marlon Oliveira*, Houssem Chatbri†, Suzanne Little†, Noel E. O'Connor†, and Alistair Sutherland*
*School of Computing, Dublin City University, Ireland
†Insight Centre for Data Analytics, Dublin City University, Ireland
Emails: marlon.oliveira2@mail.dcu.ie, {houssem.chatbri, suzanne.little, noel.oconnor, alistair.sutherland}@dcu.ie

***Abstract*—In this work we use a new image dataset for Irish Sign Language (ISL) and we compare different approaches for recognition. We perform experiments and report comparative accuracy and timing. We perform tests over blurred images and compare results with non-blurred images. For classification, we use end-to-end approach, such as Convolutional Neural Networks (CNN) and feature based extraction approaches, such as Principal Component Analysis (PCA) followed by different classifiers, i.e. multilayer perceptron (MLP). We obtain a recognition accuracy over 99% for both approaches. In addition, we report different ways to split the training and testing dataset, being one iterative and the other one random selected.**

***Index Terms*—handshape recognition, machine learning, pattern recognition, sign language**

I. Introduction

Sign Language is used by around 5,000 deaf people in Ireland. The Irish sign language is indigenous and it is estimated that is known by some another 50,000 non-deaf people [5]. In this area, computer vision can provide valuable assistance tools for human-machine interaction.

Handshape recognition has been studied for years and we still do not have a good enough implementation using only a regular camera as input, without sensors or multiple cameras. Human Computer Interaction depends strongly on the new developments in CV and handshape recognition is a very active research area in this field.

Earlier works in this area have used rather smaller datasets. For instance, Farouk *et al.* proposed two ISL datasets of relatively limited size [2]. The first dataset is composed of 920 computer generated images that are produced by the Poser software by SmithMicro. The second dataset is composed of 1620 real hand images. Both datasets represent 20 ISL handshapes. We increase the number of real hand images from 1620 to 50,000 and we use 6 different human subjects. We add an additional 3 shapes to the 20 used by Farouk. The images show the hand and arm of a signer against a uniform black background. As for recognition, it was achieved using handcrafted features as well as data-driven models (Sec. II).

In this paper, we are reporting a comparison of end-to-end approaches and feature extraction based approaches on a dataset of Irish handshape images that we collected. The end-to-end approaches include convolutional neural networks (CNN), K-nearest-neighbours (k-NN), decision trees, multilayer perceptrons (MLP), support vector machine (SVM), and linear discriminant analysis (LDA). As for the feature extraction based approaches, we use PCA to extract features, and different classifiers to recognition, including decision trees, LDA, SVM, MLP, and k-NN.

Our main contributions in this work are twofold:

- We propose a public image dataset for ISL containing more than 50,000 images. This is done by recording subjects performing ISL handshapes and designing an iterative process to select the images that keep the dataset diverse by removing redundant frames (Sec. III).
- We show that our dataset can be used to successfully train two different approaches, namely end-to-end approaches and feature extraction based approaches.

II. Related work

Hand sign recognition has been achieved with engineered features including Hidden Markov Models, Fourier Analysis, Points of Interest, Principal Component Analysis (PCA), Orientation Histograms and Kalman Filters [12]. In [10], a PCA-based method has been presented. It was proved that it is possible to interpolate data created by the projection of the images into a PCA space. This interpolation was made in order to improve the accuracy of recognising a hand sign at an unknown rotation angle. In addition, it is possible to interpolate PCA eigenspaces [9] and missing translations [11].

In [16] a novel method for handshape recognition is proposed. The idea is to recognize shapes automatically from simple sketches, such as a "stick-figure". It was introduced in [16] the Hand Boltzmann Machine (HBM) to represent the handshape space of a binary image, and formulate the user provided sketches for sampling to generate realistic handshape samples. These samples were used to train the handshape recogniser. The best recognition rate showed was 85%.

A hierarchical multistage approach was presented in [3]. The authors described an algorithm using a multistage hierarchy to build a pyramid which consists of different eigenspaces at the different levels, to analyse a new incoming pattern and classify it to the nearest neighbour pattern from a set of example images. Image blurring was used to reduce the

small changes between objects and to linearise the manifolds. However, it was only applied for computer generated images.

In [6], Nagi *et al.* reported a method for hand gesture classification using a CNN with 3 convolutional layers, two Max Pooling layers, and one fully connected layer. They collected hand gesture images of six classes corresponding to finger counts from zero to five, and they used a glove with specific color to do segmentation and hand area extraction before classification. They compared their approach with engineered features (e.g. PHOG, FFT) classified with SVMs, and obtained improved results.

III. Image dataset

In this work we use the Irish Sign Language (ISL) alphabet which is composed of 23 static gestures (corresponding to characters from the English alphabet apart from J, X and Z). Figure 1 shows cropped images of the dataset. Dataset contains hands of 6 different people (3 males and 3 females) performing the finger spelling ISL alphabet. Each shape appears 3 times. Each of the 23 static gestures is performed by moving the arm from the vertical to the horizontal position. This is to include rotation as a variation in order to train classifiers to be robust to the sign rotation angle [7].

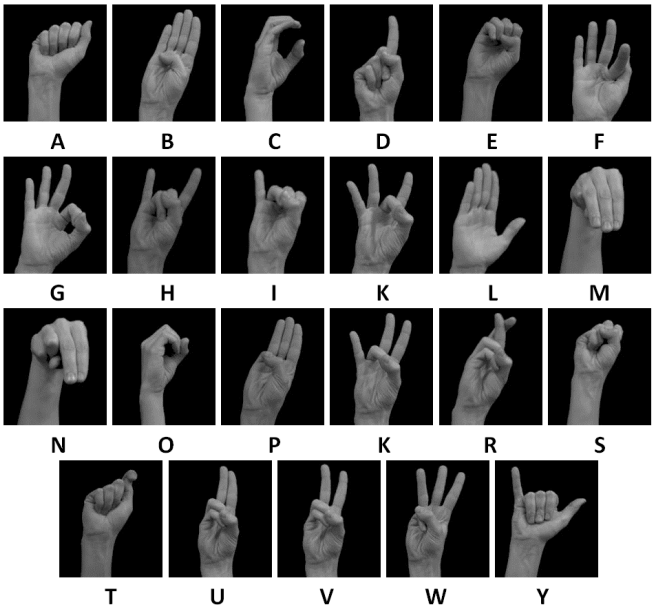

Fig. 1: Irish Sign Language static alphabet

Images resolution is 640x480 pixels and the dataset contains a total of 52,688 frames. Note that some of these frames are naturally blurred because of the arm movement. In set of images numerous frames were similar due to the speed variation of the subjects. For this reason, we designed a method to filter redundant frames (i.e. frames with insignificant difference). Our method works as follows: Each image I_u of the original dataset is represented with a compact feature vector $\overrightarrow{V_u}$.Then, a *diversity score* is introduced to express image heterogeneity, and images are iteratively selected so to optimise the *diversity score*.

We then selected 50,000 with the highest *diversity score* (roughly before the sharp decrease). We take this subset as the final dataset for subsequent experiments. In order to run experiment in a CPU we resized each image to 160×120 pixels [8].

We create two sets of testing and training dataset. The first dataset is created by iterating through the images and assigning every image to either the training or the testing set in an alternating manner, and we call it DB_i. The second set is created by a cross validation algorithm (random selection) called DB_r. Therefore, both our training and testing dataset contains 25,000 images each.

IV. Handshape classification

We compare different approaches for handshape recognition. The first approach is end-to-end (Sec. IV-A), and the second approach is based on feature extraction followed by different classifiers (Sec. IV-B).Table I shows the different methods we used.

We applied two-dimensional Gaussian blurring over the images to test how the classifiers behave on blurred images, motivated by earlier results by Farouk [2], which showed that such image filtering is beneficial for PCA accuracy. Therefore, we are interested in seeing how other approaches respond over blurred images. In this stage, a kernel of different sizes were used and the standard deviation is computed according to $\sigma = 0.3 * ((ksize - 1) * 0.5 - 1) + 0.8$. We tested $ksize$ = 5, 15 and 25. These datasets are called DB_b.

A. End-to-end approaches

For this category of methods, we used different models including a convolutional neural network (CNN), linear discriminant analysis (LDA), support vector machines (SVM), a multilayer perceptron (MLP), decision trees, and a k-nearest-neighbour (k-NN) [1]. Inputs of each model is raw image pixels (blurred or non-blurred).

The CNN is configured as follows (Fig. 2): has 4 convolutional layers with ReLU non-linearity, and ends with 2 fully connected layer with 128 and 23 neurons in each layer, and Relu and Softmax non-linearity respectively. The 23 output neurons are activated correspondingly to the image class. Dropout layers are used to prevent overfitting [13], [14]. An Adadelta optimizer [15] is used with a learning rate of 1.0, and a categorical cross entropy is used as a loss function. The filter size of the convolutional layers decreases throughout the model since that has been proven to be beneficial [4].

The other classifiers are configured as follows: The MLP has a first layer area that is connected to all pixels of the input image, a single hidden layer has 256 neurons, an output layer with 23 neurons to indicate the class of the hand shape, and it was fitted 100 times. The k in k-NN is set to $k = 1$ and for LDA we used singular value decomposition solver. Finally, for SVM we used polynomial kernel (degree 3).

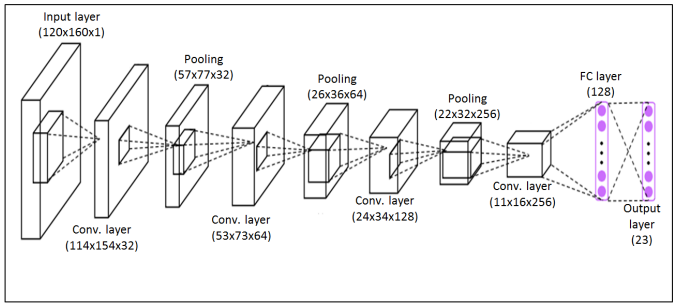

Fig. 2: Architecture of the convolutional neural network

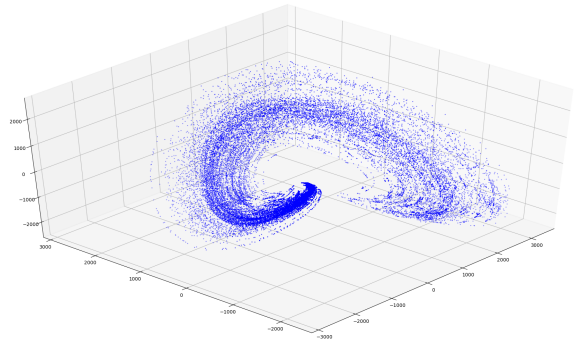

Fig. 3: Projection of training dataset images onto the PCA space in 3D

Fig. 4: Training and testing curves of the CNN model

B. Feature extraction based approaches

For this category of methods, we have extracted features using Principal Component Analysis (PCA). These features were used as input to the classifiers k-NN, LDA, MLP, SVM, and Decision Trees [1].

In order to apply PCA over the training dataset, we combined all images into the same array. After vectorization, every image was represented by a row array with 9,600 (pixels) entries. As a result, we have an eigenspace with the same dimension.

By projecting the images from the training set onto the most significant D_i eigenvectors, we obtained a D_i-dimensional space containing (N_{im}) points for each pose angle. Each point represents an image. In this study, we have tested a different number of eigenvectors and to evaluate how it affects the accuracy.

Figure 3 shows the 3 dimensions (axes) D_1, D_2 and D_3, where each point represents one image in our training dataset. In the same way, we have projected the images from the testing dataset onto the eigenspace, in order to have both in the same space.

At this point, PCA is applied to the dataset and then used with the aforementioned classifiers to measure accuracy.

The classifier are configured as follows: for Decision trees the minimum number of samples required to split an internal node is set to 2 and the the minimum number of samples required to be at a leaf node set to 1. For k-NN, LDA, MLP and SVM we keep the same configuration as end-to-end approach, a part from taking PCA's feature vectors are the inputs.

TABLE I: Recognition approaches

Approach	Models
End-to-end	Decision Trees; CNN; LDA; MLP; SVM; k-NN
Feature extraction	PCA+k-NN; PCA+SVM; PCA+MLP; PCA+LDA; PCA+Decision Trees

V. Experimental results

In this section, we report results of comparing our end-to-end and featured extraction based approaches on the Irish Sign Language dataset. Evaluation is reported in terms of $Recognition\ Accuracy$, which is defined as follows:

$$Recognition\ Accuracy = \frac{1}{N_h}\sum_{N_h=1}^{N_h}(1-\frac{|\overrightarrow{y_{true}}-\overrightarrow{y_{predicted}}|}{N_h}) \quad (1)$$

where $\overrightarrow{y_{true}}$ and $\overrightarrow{y_{predicted}}$ refer to the ground truth and predicted outputs respectively.

Experiments were made in Python 3.5, scikit-learn 0.19.0, and Keras 2.0.6, running on Windows 7, on a CPU with 16GB RAM. CNN was trained on an Nvidia Titan X GPU with 12GB RAM.

A. End-to-end classification

For the CNN model, the images were used without any pre-processing or resizing, apart from dividing the pixel intensities on 255 for normalization.

Figure 4 shows the performance progress of the model during the training. The model starts to improve quite fast in the initial iterations.

For CNN and the remaining classifiers, such as decision trees, SVM, MLP, LDA and k-NN accuracy is shown in Table II. As we can note the highest accuracy is with CNN. However, k-NN showed a very high accuracy. MLP showed the worse accuracy among the classifiers, because this technique mostly depend on a small number of features instead of raw pixels.

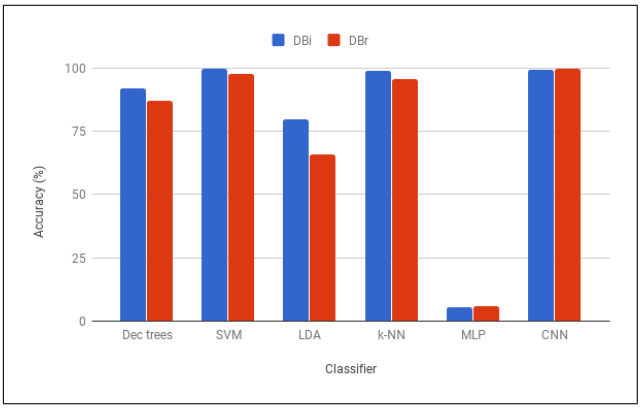

Fig. 5: Accuracy according to dataset and classifiers for end-to-end approach

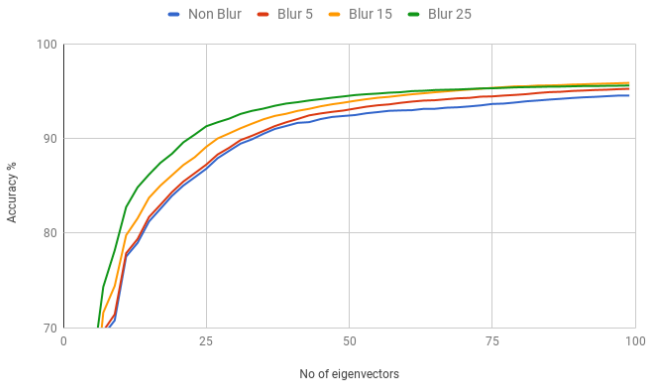

Fig. 6: Accuracy according to the blurring level for DB_r using PCA+k-NN

Some classifiers showed improved results for blurred images, i.e. k-NN, LDA and SVM.

Figure 5 shows the accuracy for the iterative dataset DB_i and for the random dataset DB_r for testing and training dataset for non blurred images. Note that iterative method showed improved accuracy for decision trees, LDA, and k-NN whereas MLP, SVM and CNN showed slightly improved results for random selection.

B. Feature extraction based classification

As stated in Sec. IV-B we used PCA in order to reduce the dimension and extract features. The highest number of eigenvectors used were 100.

The first classifier we tested was the k-NN algorithm, with $k = 1$ and Euclidean distance. We projected each testing image into the training dataset eigenspace and classified according to the nearest point (shortest Euclidean distance). We run this on two different datasets, DB_i and DB_r. In addition, we tested how different level of blurring affected the accuracy in DB_b.

Figure 6 shows the accuracy according to the number of eigenvectors and according to the Gaussian blurring kernel size for DB_r testing dataset. It is clear that blurring helps to reduce the number of eigenvectors needed to obtain a good accuracy. This finding is consistent with earlier studies showing the positive effect of image filtering with blurring on PCA by reducing the non-linearity in the manifolds within the eigenspaces [3]. In addition, we can note an optimal result in the blurring with kernel size $(15, 15)$. For that reason, all the other experiments from now on will use this kernel size.

Finally, Figure 7 shows the accuracy for the iterative dataset DB_i and for the random dataset DB_r for testing and training dataset for non blurred images. Note that iterative method showed improved accuracy for almost all classifiers except for PCA+LDA.

In addition to k-NN we tested different classifiers over our PCA data, such as decision tree, LDA, SVM and MLP. Table II reports the accuracy for each classifier, over the DB_r dataset, for either blurred and non blurred images. Note that blurring helps all tested classifiers and the best accuracy is obtained with PCA+SVM for feature extraction approach. PCA+LDA showed the lowest accuracy comparing to the other techniques.

TABLE II: Classifier performances. Best ones are the CNN model, the second best the PCA+MLP model and the third k-NN

Model	End-to-end accur		Featured accur.	
	Non-blur	Blur	Non-blur	Blur
CNN	99.80%	99.58%	-	-
MLP	6.01%	5.61%	98.56%	99.50%
Decision Trees	87.21%	88.81%	76.45%	77.36%
k-NN (k=1)	95.50%	96.91%	94.56%	95.90%
LDA	65.94%	29.97%	32.72%	38.82%
SVM	97.86%	99.80%	99.61%	99.87%

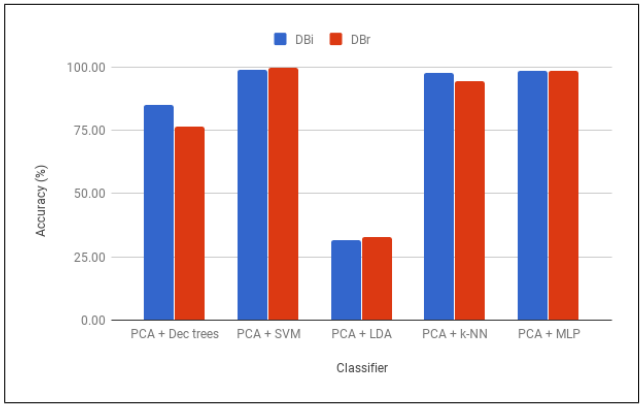

Fig. 7: Accuracy according to dataset and classifiers for feature extraction approach

Table II shows how classifiers behave with blurred and non blurred images. It is possible that CNNs carry out their own blurring on the images. Blurring is a convolution operation and it showed an insignificant change in testing accuracy for CNN, proving that pre-processing is not important for this technique. Decision tree, k-NN and MLP showed a slightly improvement with blurred images, being decision trees the most benefited.

Figures 8 and 9 show the average time and standard devi-

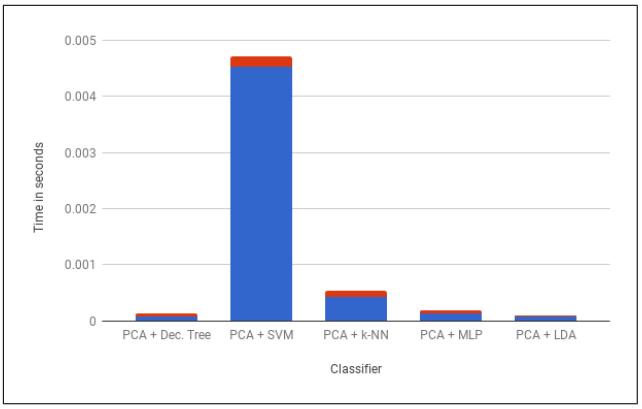

Fig. 8: Average time to classify one image according to the classifier with PCA

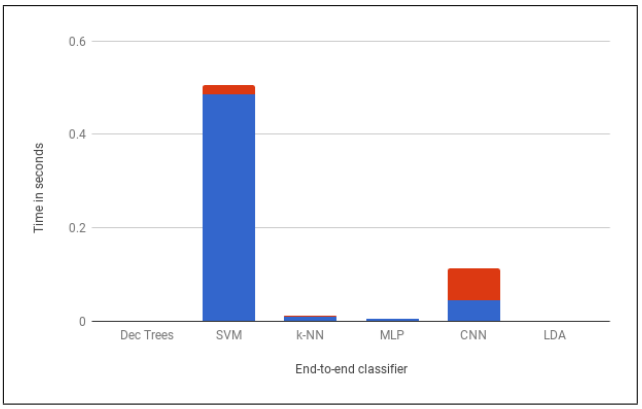

Fig. 9: Average time to classify one image according to the classifier end-to-end approaches

ation to recognize one image in seconds. We used 10 images to measure the time. For PCA approach we considered 100 eigenvectors, for CNN we considered 100 iterations because this number provided a good accuracy as an empirical result. Figure 8 shows the speed for classifying one image using PCA and different classifiers, note that the shortest classification time is with PCA+decision trees and for PCA+LDA, and the classifier that took the longest is PCA+SVM. Taking into account speed and accuracy PCA+MLP provided the best accuracy with a still short time. Figure 9 shows the same information for end-to-end classifiers, being again decision trees the fastest one.

VI. Conclusions and Future Work

In this paper, we reported a comparative study of using end-to-end approaches and feature extraction based approaches in recognising hand shapes. Our results showed that, opposed to common belief, handcrafted features are still strongly competitive against deep features extracted using convolutional neural networks (CNN).

Furthermore, we introduced and tested different techniques over a large dataset that contains 50,000 images and 23 handshapes for the Irish Sign Language. This paves the way for more in-depth research in this area that were hindered by the lack of large public datasets.

Building on our results, we plan to extend our algorithms to video and build an automatic transcript system. For this purpose, we will exploring sequential models that take into account the time dimension, such as recurrent neural networks and hidden Markov models.

Acknowledgements

This research was funded by CAPES/Brazilian Science without Borders, process no.: 9064-13-3 and is co-funded under the European Regional Development Fund. This research also emanated from a grant in part from the IRC under Grant no. GOIPD/2016/61, and in part from SFI under Grant no. SFI/12/RC/2289 (Insight).

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.
[2] M. Farouk. *Principal Component Pyramids using Image Blurring for Nonlinearity Reduction in Hand Shape Recognition*. PhD thesis, Dublin City University, Ireland, 2015.
[3] M. Farouk, A. Sutherland, and A. Shokry. Nonlinearity Reduction of Manifolds using Gaussian Blur for Handshape Recognition based on Multi-Dimensional Grids. *ICPRAM*, 2013.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
[5] L. Leeson and J. I. Saeed. *Irish Sign Language : A Cognitive Linguistic Account*. Edinburgh University Press, 2012.
[6] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347. IEEE, 2011.
[7] M. Oliveira, H. Chatbri, Y. Ferstl, M. Farouk, S. Little, N. E. O'Connor, and A. Sutherland. A dataset for irish sign language recognition. In *Irish Machine Vision & Image Processing Conference*, 2017.
[8] M. Oliveira, H. Chatbri, Y. Ferstl, S. Little, N. E. O'Connor, and A. Sutherland. Irish sign language recognition using principal component analysis and convolutional neural networks. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (in press)*. IEEE, 2017. in press.
[9] M. Oliveira and A. Sutherland. Interpolating eigenvectors from second-stage pca to find the pose angle in handshape recognition. In *Irish Machine Vision & Image Processing Conference*, 2015.
[10] M. Oliveira, A. Sutherland, and M. Farouk. Manifold interpolation for an efficient hand shape recognition in the irish sign language. In *International Symposium on Visual Computing*, pages 320–329. Springer, 2016.
[11] M. Oliveira, A. Sutherland, and M. Farouk. Two-stage pca with interpolated data for hand shape recognition in sign language. In *Applied Imagery Pattern Recognition Workshop*. IEEE, 2016.
[12] A. K. Sahoo, G. S. Mishra, and K. K. Ravulakollu. Sign Language Recognition : State of the Art. *Asian Res. Publ. Netw.*, 9(2):116–134, 2014.
[13] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
[14] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1058–1066, 2013.
[15] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
[16] X. Zhu, R. Sang, X. Jia, and K. Y. K. Wong. A hand shape recognizer from simple sketches. In *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, pages 130–135, Nov 2013.