DBN-Based Fingerspelling Recognition Approach using Feature fusion

Yong Hu
1. The Laboratory for Internet of Things and Mobile Internet Technology of Jiangsu Province, Huaiyin Institute of Technology, 1 Meicheng East Road, Huai’an, Jiangsu Province, China
2. School of Software Engineering, Jinling Institute of Technology, 99 Hongjing Avenue, Jiangning University Town, Nanjing, Jiangsu Province, China
Email: huyong@jit.edu.cn

Abstract. Sign language recognition offers effective and precise approach of recognizing gestures or postures. In this work, a vision-based framework is presented for recognizing fingerspelling alphabets and a comparison is conducted to show the efficiency of feature fusion. Fused features and Deep Belief Network are used in the proposed framework. In the experiments stage, a comparison between the fused features and the individual ones is performed by using two public fingerspelling datasets. Experiment results show the improvement of the feature fusion.

1. Introduction
Sign language recognition (SLR) offers effective and precise approach of recognizing gestures or postures. A lot of published research summarized this field [1]. Generally, gestures are composed of 2 parts; one is static gestures and the other is dynamic gestures. Static sign is presented by stationary contour of hand(s), while a dynamic one is presented by continuous hand(s) motions. There are 2 classes of approaches in SLR; electromechanical devices based approaches (such as data-gloves) and vision-based approaches. Electromechanical devices based approach is costly and user-unfriendly. Vision-based method conducted solely machine vision techniques for recognition schemes. It handles the vision of bare hand(s), which enable the interaction naturally. It is universally utilized due to the performance of user-friendly and lower cost.

D. Kelly et al. [2] proposed a user-independent algorithm by using Support Vector Machines (SVMs). An eigenspace size function and Hu moments features were used for classification. F. H. CHOU et al. [3] presented a GMM-based (Gaussian Mixture Model) processing algorithm for the gesture images detection and recognition. Based on their presented algorithm, the correct recognition rate is about 94% in average. K. C. Otiniano-Rodriquez et al. [4] proposed two methods for SLR using the SVM classifier and features extracted from Hu and Zernike Moments. In the experiments, the proposed methods are performed using a database composed of 2040 images for recognition of 24 symbol classes. P. Rajathi and S. Jothilakshmi [5] proposed a three stage system to recognize static gestures representing Tamil words.

In this study, a vision-based framework is presented for recognizing fingerspelling alphabets. The recognition system is based on the fused features and Deep Belief Network (DBN). Local Binary Patterns (LBP), Zernike moments and Histogram of Oriented Gradient (HOG) are extracted and then fused as discriminating features. DBN with three hidden layers is applied for training and testing. The comparison between the fused features and the individual one shows the promising effectiveness of the fused features.
This manuscript is arranged as follows. Section 2 presents the materials and methods concisely. Section 3 introduces the proposed framework which contains system architecture, feature normalization/ fusion and classification/ testing in detail. Section 4 shows the experimental results. Finally, the summary and conclusions are provided in Section 5.

2. Related work

2.1 Local Binary Patterns
LBP offers an efficient and straight way for feature extraction. It has the significant characteristics of less computation cost, rotation- and translation-invariant. Lots of LBP-based approaches have been utilized in texture segmentation, face recognition, gestures recognition, et al. In this approach, the texture $T$ in neighboring pixels is specified by the joint distribution of the gray value of $P$ ($P>1$) image pixels:

$$
T = t(g_c, g_0, \ldots, g_{p-1})
$$

(1)

Where $g_c, g_p$ represents the gray value of the center point and its neighboring pixels respectively. The following equation defines a LBP operator:

$$
LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p
$$

(2)

On the basis of this formula, LBP operator represents its local neighborhood by compare with the center point. If the gray value of center point is less than the pixel in local neighborhood, the value will define as 1, otherwise defined as 0. The result of the LBP operator is a P-bit binary number with $2^P$ unique values.

2.2 Zernike Moments
Zernike moments usually considered as holistic features in recognition systems. Scale-, rotation- and translation-invariance are the properties of Zernike moments. The multinomial of Zernike moments is established on a unit circle $x^2+y^2\leq1$ by completely orthogonal basis. For a gray image $f(x,y)$, the following equation defines Zernike moments:

$$
Z_{nm} = \frac{n+1}{\lambda} \sum_{r=0}^{\infty} \sum_{\gamma} f(x,y)V^*_{nm}(x,y)
$$

(3)

Where: $n, m, \lambda$ are order, repetition and normalization factor respectively.

In order to reach scale- and translation-invariance, the Zernike moments could be normalized by using the following equation

$$
g(x,y,t) = f(\frac{x}{a} + \bar{x}, \frac{y}{a} + \bar{y})
$$

(4)

2.3 Histogram of Oriented Gradient
Many approaches based on HOG operator have been certified its efficiency in recognition systems. The HOG operator catches the important features of local figure - gradient structure. Rotation- and translation-invariance are the properties of HOG. The operator is computed in each cell and combined to obtain a feature vector. The size and distance of the grids are the parameters which affect the generalization performance well.

HOG descriptor can be computed by the steps list below. Firstly, calculate gradients of the image by apply two 1-D filters $[-1 0 1]$ and $[-1 0 1]^T$, one for horizontal and the other for vertical. Secondly, Divide the image into grids of equal size with over-lapping or non-overlapping and then calculate the histogram of gradient in every grid. More number of bins means greater extent of details. Finally,
Normalization scheme is conducted for each cell and the factor could be obtained by L1-norm or L2-norm; Concatenation every histogram into one feature vector.

3. The Proposed Recognition Framework

3.1 System Architecture

The proposed recognition framework consists of 4 stages and shown in Figure 1. Firstly, the fingerspelling image is segmented, denoised and enhanced in the pre-processing step. And then, the segmented hand area is conducted for extracting features. Several features including Local Binary Patterns, Zernike Moments and Histogram of Oriented Gradient are extracted. The features are normalized and fused as discriminating features. In the training/classification stage, a DBN with three hidden layers is applied for recognition. In the last stage, a comparison between the fused features and the individual ones is performed by using two public fingerspelling datasets.

![Figure 1. The proposed recognition framework](image)

3.2 Feature Normalization & Fusion

Feature normalization & fusion are crucial steps before pattern recognition. Each feature component will be the same contribution to the distance and the features with bigger range could not lead the computations in classification. Thus, setting each feature dimension of the entire feature to the range of [0, 1] individually becomes a common method to normalize features.

It is hard to build a clear and strict principle of data fusion. It can be defined as a combination of multiple sources to obtain improved information; which means less expensive, higher quality, or more relevant information [6]. One of the most well-known data fusion classification systems was provided by Dasarathy. The data fusion techniques can be classified into five nonexclusive categories: Data In-Data Out (DAI-DAO), Data In-Feature Out (DAI-FEO), Feature In-Feature Out (FEI-FEO), Feature In-Decision Out (FEI-DEO) and Decision In-Decision Out (DEI-DEO).

In this work, a feature normalization & fusion step is implemented to improve extracted features. FEI-FEO is used as intermediate level fusion method to obtain new features.

3.3 Deep Belief Network

DBN was first proposed by Hinton in 2006. The novel deep structured learning architecture was considered as a breakthrough on deep learning techniques. Due to the inherent capability of overcoming the drawback of traditional algorithms, deep learning techniques (include DBNs) have drawn ever-increasing research interests [7].

DBNs are composed of multiple layers of RBMs (Restricted Boltzmann Machines), in which the hidden visible layers are disconnected. The Boltzmann Machine is a concurrent network with stochastic binary units. The network has a set of visible units \( v \in \{0,1\}^{g_v} \) and a set of hidden units \( h \in \{0,1\}^{g_h} \) where \( g_v \) and \( g_h \) are the number of visible units and the number of hidden units respectively. The joints between hidden units and also between visible units are disconnected. The energy of the joint configuration \( \{v, h\} \) in Boltzmann machine is given as follows:
\[ E(v,h) = -\frac{1}{2} v^T L v - \frac{1}{2} h^T J h - v^T W h \] 

(5)

The bias is removed for simplicity of presentation. The term W is the concurrent weights between visible and hidden units, L is the concurrent weights between visible and visible units and finally J is the concurrent weights between hidden and hidden units. Diagonal values of L and J are zero.

4. Experimental Results

4.1 Dataset

For implementation, two image sets were used in this work. The Jochen-Triesch dataset composed of ten signs conducted by twenty-four various signers with diverse backgrounds. Ten instances of Jochen-Triesch static hand posture dataset in black backgrounds are shown in Figure 2 (one instance for every sign). The Thomas Moeslund’s dataset is known as an arduous one. It is composed of more than 2 thousands images that exemplify twenty-four grey scale static signs. Twenty-four instances of the dataset are shown in Figure 3 (one instance for every sign).

4.2 Experimental setup

In order to make the system simpler and faster, sign images are resized after segmentation to binary images. A fast image segmentation method based on histogram information is used for each sign image. Their size is reduced to a resolution of 80×100 pixels in order to decrease computational cost. The calculation of feature extraction is based on small images. The detail experimental setup of each feature is listed as follows.

Scale and translation normalized Zernike moments are computed by using centered image. 15 dimensions feature vector is conducted by up to order 6 and only the real part is applied in this study. Original LBP approach with R=1 is applied to calculate features, thus the dimensions of the feature vector is 256. In extracting HOG features, local window is set to 3×3 and the number of histogram bins is set to 9. A feature vector of 81 dimensions is extracted from the finger spelling images.

4.3 Recognition results

The detail results of this classification with 20 samples for training in the Jochen-Triesch database and 80% of samples for training in the Thomas Moeslund's database are list in Table 1. It shows the recognition rate for all signs from two databases. It is observed that features have its own advantages and disadvantages of finger spelling recognition. That is the reason why feature fusion becomes an alternative way for recognition. Improved information is obtained by a combination of multiple sources, which lead to higher quality and accuracy.
5. Summary and Conclusions
In this work, a comparison is conducted to show the efficiency of feature fusion. Several features are extracted and then fused as discriminating features. DBN with three hidden layers is applied for training and testing. In the experiments, a comparison between the fused features and the individual ones is performed by using public fingerspelling databases. Experiments show the improvement of the feature fusion. The work done in this paper is actually a preliminary work for SLR. Future research is needed to solve the confusing letters and the isolated signs with motions.

6. References
[1] Murthy G R S and Jadon R S, A review of vision based hand gestures recognition 2009 Int. J. Inf. Tech. Knowl. Manage. 2(2) 405-10
[2] Kelly D, McDonald J and Markham C, A person independent system for recognition of hand posture used in sign language 2010 Pattern Recogn. Lett. 31(1) 1359-68
[3] Chou F H and Su Y C, An Encoding and Identification Approach for the Static Sign Language Recognition 2012 Int. Conf. on Advanced Intelligent Mechatronics (Taiwan) p 885-9
[4] Otiniano-Rodrıguez K C, Cá mara-Chá vez G and Menotti D, Hu and Zernike Moments for Sign Language Recognition 2012 Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition(IPCV’12) p 1-5
[5] Rajathi P and Jothilakshmi S, A Static Tamil Sign Language Recognition System 2013 Int. J. of Advan. Res. Comput. Commun. Eng. 2(4) 1885-91
[6] Castanedo F, A Review of Data Fusion Techniques 2013 Sci. World J. (2013) 1-19, doi:10.1155/2013/704504
[7] Liu W, Wang Z and Liu X, A survey of deep neural network architectures and their applications 2016.Neurocomputing 234 11-26