Hand Gesture Recognition Using Local Histogram Feature Descriptor

Dandu Amarnatha Reddy
Department of EC
NIT Rourkela, India
Email: damarnath402@gmail.com

Jaya Prakash Sahoo
Department of EC
NIT Rourkela, India
Email: sahoo.jprakash@gmail.com

Samit Ari
Department of EC
NIT Rourkela, India
Email: samit.ari@gmail.com

***Abstract*—Hand gesture recognition system is widely used in the development of human-computer interaction. The vision based hand gesture recognition is achieved by the following steps: preprocessing, feature extraction and classification. The aim of preprocessing stage is to localize the hand region from the image frame. The Laplacian of Gaussian filtering technique along with zero crossing detector is applied on hand gesture images to detect the edges of hand region. This paper proposes a novel feature extraction technique, which is based on local histogram feature descriptor (LHFD). The proposed feature is extracted by finding the local histogram of the gray scale gesture image. This technique uses the whole region of the hand to extract the features. The proposed method is invariant to the scaling and illumination. Two standard datasets viz. Massey University gesture dataset (MUGD) and Jochen Triesch static hand posture database are used to evaluate the recognition performance of the proposed technique. The gesture recognition performance of the proposed technique is 99.5% and 95% on Massey University gesture dataset and Triesch dataset respectively, using multi-class support vector machine (SVM) classifier.**

***Index Terms*—Feature extraction, Hand gesture recognition, Human computer interaction, Local histogram, Support Vector Machine.**

I. Introduction

The usage of computers became essential to the real life activities like education, medical field, entertainment, business, visual surveillance etc. The modern society is emerging with human-computer interaction (HCI) to increase the smartness in the daily activities [1]. Hand gesture is a way of nonverbal communication between the human being and computer. Nowadays, research on Hand gesture recognition is increasing due to its potential applications in real life [2].

The major steps in vision based hand gesture recognition system are, data acquisition, hand localization, feature extraction and gesture classification. A web-cam is used for data acquisition in vision based system. The objective of hand localization stage is to segment the hand region from an image frame. After hand localization, the next step is to extract an efficient feature from the gesture image. To represent the characteristics of gesture images, various features techniques have been reported, they are shape context [3], elastic graph matching [4], eigenspace Size Function and Hu moments features [5] and Kernel based feature [6]. Belongie *et al.* [3] have introduced the shape context descriptor. The descriptor comprises histogram of relative polar co-ordinates of any single boundary point. Shape contexts are translation invariant but variant to the scaling. The elastic graph matching algorithm which is proposed by Triesch *et al.* [4] to classify the 10 ASL alphabets gestures. This technique does not require segmentation of hand from the background. However, it is facing more computational complexity. Just *et al.* [7] used the Modified Census Transform (MCT): The approach is based on a local non-parametric pixel operator. This method doesn't requires segmentation of hand region in an image. Kelly *et al.* [5] proposed a technique based on eigenspace Size Function and Hu moments features to classify different hand postures. Kernel based feature extraction methods which are kernel principle component analysis (KPCA) and kernel discriminant analysis (KDA) along with SVM and neural networks are used in [6]. After the extraction of features, some of the existed possible techniques for hand gesture classification are neural network [6], Gaussian mixture model [8] and support vector machine (SVM) [9].

From the above literature survey [4], [6], [7], it is found that, vision based static hand gesture recognition system is facing problems at preprocessing and feature extraction stages. The feature of the gesture image is greatly affected by the lighting condition of the gesture images. Therefore, the gesture images should be illumination invariant at preprocessing stage to localize the hand in gesture frame. The recognition of hand gesture is achieved by an efficient discriminative feature of the gesture image. In this paper, a novel feature descriptor is proposed for efficient recognition of hand gesture image. At pre-processing stage, the Laplacian of Gaussian filtering technique with zero crossing detector is used to detect the edges of the hand gesture images. The advantage of this technique is to obtain the edges of the hand in an image frame effectively by removing the shadows present in the background. The novel local histogram feature descriptor (LHFD) is obtained from the preprocessed hand gesture images.

This paper presents the results for two types of standard datasets. Dataset I is Massey University gesture dataset (MUGD) [10]. Gesture dataset contains postures with varying illuminations. Dataset II is Jochen Triesch static hand posture dataset [4], which comprises 10 different hand gestures. The

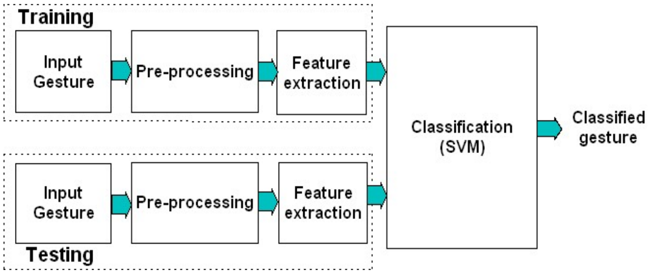


Fig. 1: Block diagram of hand gesture recognition system.

recognition performance of the proposed technique on dataset I and dataset II is 99.5% and 95% respectively.

The remaining topics of the paper are as follows. Section II explains the proposed methodology for vision based static hand gesture recognition system. Section III describes the standard dataset and evaluation methods. Finally, the paper is concluded in section IV.

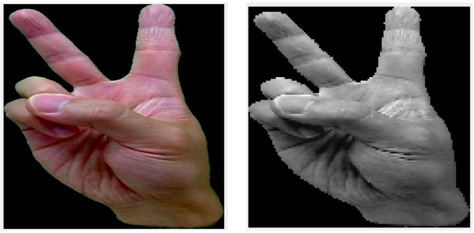
II. Methodology

The framework of proposed hand gesture recognition technique is shown in Fig. 1. Block diagram comprises two phases, training phase and testing phase. Both phases follow preprocessing, feature extraction and classification stages for recognition of hand gestures.

A. *Preprocessing*

Preprocessing stage plays a vital role to extract the hand region in static image, which includes image enhancement, image binarization, and cropping the hand region (segmentation).

Dataset I [10] contains color images. In order to convert the color images as illumination invariant, Gray world algorithm is used effectively. In gray world algorithm [11], compensation of intensity variation is achieved by calculating the mean of red, green and blue channels in an image. All the normalized images are converted into gray-scale images and each image is resized to the equal size of 128×128 pixels as shown in Fig. 2.



(a) (b)

Fig. 2: Simulation results of preprocessing stage on gesture of ASL number class '2' of dataset I. (a) Original color image, (b) Image obtained after intensity normalization by using Gray world algorithm and resized to the size of 128×128.

Preprocessing stage for Dataset II is carried by using edge detection followed by morphological filtering. The simulation results are demonstrated in Fig. 3.

1) Edge detection: The Laplacian of Gaussian filtering technique [12] along with zero crossing detector is applied on hand gesture images to eliminate the shadows present in the background and to detect the edges of hand region. Fig. 3(a) shows original image having shadow. A fine edge detected image is shown in Fig. 3(b).

2) Morphological filtering: Morphological operations includes dilation and erosion. Dilation adds a layer of pixels to the both inner and outer boundaries of regions in a binary image, while erosion removes layer of pixels from the both inner and outer boundaries of the region in a binary image. To extract neat segmented image, dilation followed by erosion is performed. The morphological filtered image is shown in Fig. 3(c).

3) Replacing the hand region of binary image with gray-scale pixels of original image: After morphological filtering, the resultant image is the binary image with segmented hand region as shown in Fig. 3(c). After obtaining the segmented image, pixels of hand region are replaced by the pixels of the hand region in the original gray-scale image. Resultant segmented image is shown in Fig. 3(d). Finally, the hand region is cropped from the obtained gray-scale image and it is resized to the size of 128×128 pixels. The image in Fig. 3(e) is the final preprocessed image, used for feature extraction.

B. *Feature extraction technique*

In the present work, a novel local histogram feature descriptor (LHFD) is proposed. Histogram of an image is the representation of probability distribution of intensity levels. Extraction of local histogram feature descriptor of an image is obtained as follows:

Step.1:The preprocessed gray-scale image is divided into 16 small blocks as shown in Fig. 4. Here, size of the each block is 32×32 pixels.

Step.2:Then the histogram for each individual block is calculated. As the histogram is computed for each block individually in the image, it is called as Local histogram. Each 32×32 block has 1024 pixels, it can be shown as

$$Block = \{P_1, P_2, P_3, P_4, ..., P_{1024}.\} \tag{1}$$

where, P represents pixel of the respective block.

To find histogram:

$$\begin{aligned} C_0 &= count(Intensity \quad values \quad at \quad P_i = 0) \\ C_1 &= count(Intensity \quad values \quad at \quad P_i = 1) \\ &\vdots \\ &\vdots \\ C_{255} &= count(Intensity \quad values \quad at \quad P_i = 255) \end{aligned}$$

where,

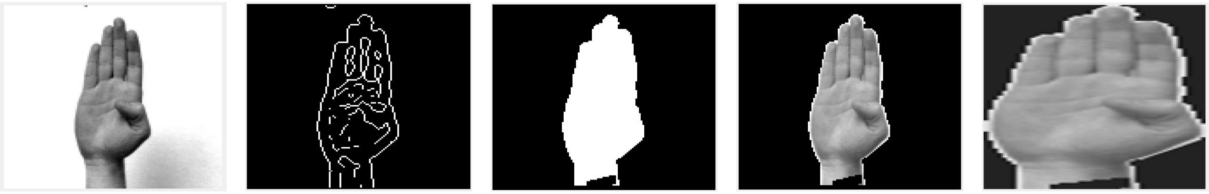

(a) (b) (c) (d) (e)

Fig. 3: Simulation results of preprocessing stage on gesture class ‘B’ of dataset II. (a) Original image, (b) Edge detected image, obtained by applying Laplacian of Gaussian filter followed by zero-crossings detector, (c) Morphological filtered image, (d) An image obtained after replacing the pixels of the hand region in the binary image (Fig. 3(c)) with gray-scale pixels of the hand region in the original image, (e) Cropped and resized image (128 × 128 pixels).

i = 1 to 1024,
C_0, C_1, C_2,........, C_{255} are the counts.

Step.3: In this step, histogram of each block is divided into eight bins. These obtained bins of each block are concatenated into a single array. Hence, the length of the final array is 128. This array represents the feature vector of that particular image. This feature vector is normalized with maximum value of that feature vector. Fig. 5 shows the plot of histogram bins of 6th block of Fig. 4(b). The concatenated vector of histogram bins of Fig. 4(b) is shown in Fig. 7. Let $b_0, b_1, ...b_8$. are the bins of a particular block. Then,

$$\begin{aligned} b_0 &= c_0 + c_1 + c_2....c_{31} \\ b_1 &= c_{32} + c_{33} + c_{34}....c_{63} \\ b_2 &= c_{64} + c_{65} + c_{66}....c_{95} \\ &\vdots \\ b_7 &= c_{224} + c_{225} + c_{226}....c_{255} \end{aligned}$$

$$B_0 = \{b_0, b_1, b_2....b_7\}_{(1\times8)}$$

similarly, for other blocks,

$$\begin{aligned} B_1 &= \{b_0, b_1, b_2.....b_7\}_{(1\times8)} \\ B_2 &= \{b_0, b_1, b_2.....b_7\}_{(1\times8)} \\ &\vdots \\ B_{15} &= \{b_0, b_1, b_2....b_7\}_{(1\times8)} \end{aligned}$$

$$Featurevector = \{B_0, B_1, B_2....B_{15}\}_{(1\times128)} \quad (2)$$

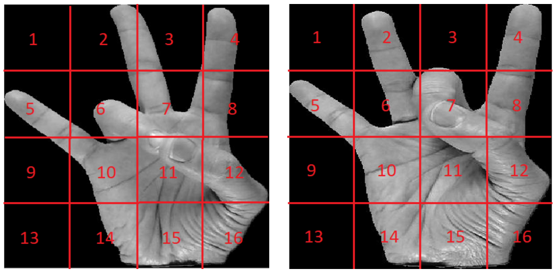
$$B_{max} = max(featurevector) \quad (3)$$

Finally, the normalized feature vector of particular image is =

$$\{B_0, B_1, B_2....B_{15}\}/(B_{max}) \quad (4)$$

Step.4: Above three steps are repeated for all images.

Finally, the normalized feature vectors of all images are applied to the SVM (Support Vector Machine) classifier by dividing some of the images as training images and remaining images as testing images.



(a) (b)

Fig. 4: Division of image into blocks of size 32 × 32: (a) Gesture image of number class 7 from dataset I, (b) Gesture image of number class 8 from dataset I.

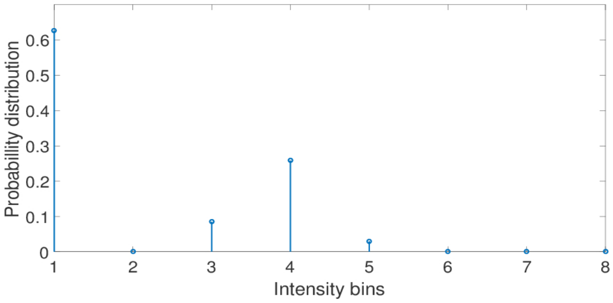


Fig. 5: Histogram plot for block 6 in Fig. 4(b). X-axis represents intensity bins and Y-axis represents probability distribution of intensity bins.

C. Support Vector Machine (SVM)

SVM is one of the popular supervised machine learning algorithms, which can be used for data classification [9]. In this paper, a kernel based SVM is used to classify different classes of data by using optimal boundary between the classes. The advantages of SVM are the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin.

Learning an SVM has been formulated as a constrained optimization problem over $\mathbf{w}$ and ξ

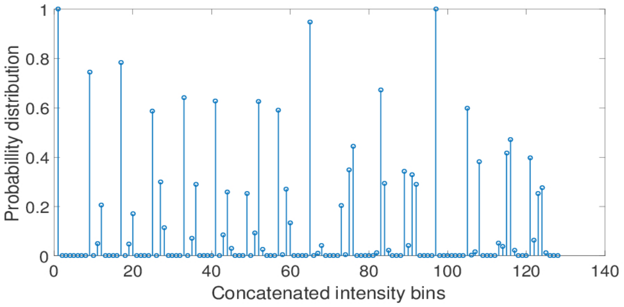


Fig. 6: Histogram plot for Fig. 4(a). X-axis represents concatenated intensity bins of 16 blocks (each block is having 8 bins) and Y-axis represents probability distribution of intensity bins.

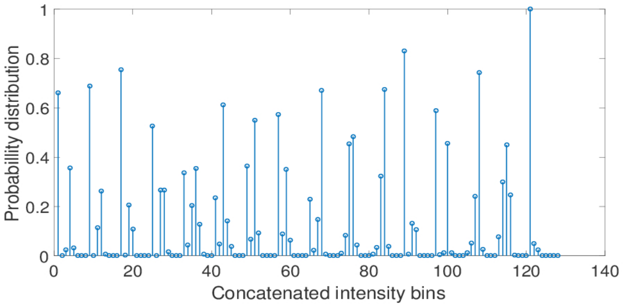


Fig. 7: Histogram plot for Fig. 4(b). X-axis represents concatenated intensity bins of 16 blocks (each block is having 8 bins) and Y-axis represents probability distribution of intensity bins.

$$\underset{\mathbf{w}\in\mathbb{R}^{\mathrm{d}},\xi_{\mathrm{i}}\in\mathbb{R}^{+}}{min} ||\mathbf{w}||^2 + \mathrm{C}\sum_{\mathrm{i}}^{\mathrm{N}}\xi_{\mathrm{i}} \tag{5}$$

subject to $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i$ for $i = 1, 2, ..., N$.

The constraint

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i$$

can be written more concisely as

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i$$

which, together with

$$\xi_i \geq 0$$

is equivalent to

$$\xi_i = max(0, 1 - y_i f(\mathbf{x}_i)) \tag{6}$$

substitute equation 6 in equation 5

$$\underset{\mathbf{w}\in\mathbb{R}^{\mathrm{d}},\xi_{\mathrm{i}}\in\mathbb{R}^{+}}{min} ||\mathbf{w}||^2 + \mathrm{C}\sum_{\mathrm{i}}^{\mathrm{N}}\max(0, 1 - \mathrm{y_i f}(\mathbf{x}_{\mathrm{i}}))$$

where,
ξ_i are slack variables
$\mathbf{x}_i$ is training data
y_i represents class label
C is regularization factor
$f(\mathbf{x}_i)$ is the prediction function
$\mathbf{w}$ is weight vector
b represents biasing term.

III. Evaluation method

A. Datasets

The performance of the proposed technique is evaluated on two standard datasets. First one is Massey University Gesture Dataset (MUGD) and Second one is Jochen Triesch Static Hand Posture dataset.

1) Massey University gesture dataset (MUGD): Dataset I is a standard ASL hand gesture dataset [10], having 36 gesture classes (alphabet gestures a to z and number gestures 0 to 9). The dataset is collected from 5 different subjects with illumination variations in five different directions like bottom, top, left and right with uniform black background. As in the earlier reported techniques [8], [13], the analysis is performed on only number dataset. Therefore, the performance of the proposed technique is evaluated by considering the number gestures of MUGD. From each class in the number dataset, 40 images are used as training data and the remaining 20 images are considered to test the performance of proposed technique.

2) Jochen Triesch Static Hand Posture dataset: Dataset II is a benchmark dataset of 10 static ASL alphabet hand postures [4]. The dataset is developed from 24 subjects in three distinct backgrounds: uniform light, uniform dark and complex. In this work, the experiments are carried out for light background postures only. The performance is evaluated using two protocols. The protocol P1 is based on [7], [5], in which gesture images of 3 subjects are used for training and images of 21 subjects are used to test the proposed technique. Similarly, in protocol P2, gesture images of 8 subjects and 16 subjects are used for training and testing the proposed technique respectively.

B. Evaluation Results

Experiments are performed on dataset I, forty images of each class in the number dataset are used as training data and the remaining 20 images are considered as testing data based on [8]. The confusion matrix for gestures number 0 to number 9 is as shown in Table I. Misclassification is occurred in gesture seven with gesture eight, since the gesture seven and gesture eight are less differentiable in shape. The recognition performance of propose method on dataset I is 99.5%, for the block size of 32 × 32.

The proposed method is also applied to the dataset II. The obtained recognition accuracy based on protocol-1 is 91.90% and the recognition accuracy based on protocol-2 is 95%, since protocol-2 is having more number of training images than that of protocol-1.

TABLE I: Confusion matrix for dataset I

Class	0	1	2	3	4	5	6	7	8	9
0	20	0	0	0	0	0	0	0	0	0
1	0	20	0	0	0	0	0	0	0	0
2	0	0	20	0	0	0	0	0	0	0
3	0	0	0	20	0	0	0	0	0	0
4	0	0	0	0	20	0	0	0	0	0
5	0	0	0	0	0	20		0	0	0
6	0	0	0	0	0	0	20	0	0	0
7	0	0	0	0	0	0	0	19	1	0
8	0	0	0	0	0	0	0	0	20	0
9	0	0	0	0	0	0	0	0	0	20

TABLE III: Recognition performance on dataset II

Protocol	Method	Accuracy (%)
P1	Kelly *et al.* [5]	85.1
	Triesh *et al.* [4]	95.2
	Moghaddam *et al.* [6]	89.5
	Proposed method	91.9
P2	Just *et al.* [7]	89.9
	Kelly *et al.* [5]	91.8
	Moghaddam *et al.* [6]	95.3
	Proposed method	95.0

C. *Recognition performance comparison*

The comparison of recognition performance on dataset I with different techniques is shown in Table II. In Rasines *et al.* [8], the obtained recognition accuracy is 97.5%. In this method 25% of gesture-7 images are misclassified with gesture-8. This misclassification is effectively overcomes by proposed method.

TABLE II: Recognition performance on dataset I

S. NO	Methods	Accuracy (%)
1	Avram *et al.* [13]	83.1
2	Rasines *et al.* [8]	97.5
3	Proposed method	99.5

Table III shows the Recognition performance of proposed method on dataset II for block size 16 × 16 pixels. The recognition accuracy for Protocol-1 is 91.5% and the same for protocol-2 is 95%. Recognition performance for block size 32 × 32, based on protocol-1 is 87.6% and the same on protocol-2 is 92.5%. The recognition accuracy in the case of 32 × 32 is less than that in the case of 16 × 16, since the recognition accuracy is depending on the length of the feature vector. The length of the feature vector for block size 16 × 16 is 512 and the same for block size 32 × 32 is 128. The recognition performance is increases as the length of the feature vector is increases and vice versa. Because, a fine information can be confined, if the block size is 16 × 16 than that in the case of 32 × 32. Though Jochen Triesch *et al.* [4] achieved 95.2% accuracy for protocol P1, this method is facing more computational complexity. Proposed method is giving comparable recognition accuracy with less computational complexity. The recognition performance based on protocol-2 is better than existed techniques [5], [7] and comparable with Moghaddam *et al.* [6].

IV. Conclusion

This paper proposes a novel local histogram feature descriptor (LHFD). The Laplacian of Gaussian filter followed by zero crossing detector is used to detect the edges of hand region by removing shadow present in the background. The recognition accuracy of the proposed method for dataset I is 99.5%, which gives satisfactory performance when compared to the all viable techniques. Proposed method is efficiently classifying the gestures of different classes belong to the dataset I, which are having similar shapes. Proposed feature extraction technique is able to extract the discriminative features for the gestures of different classes. The recognition accuracy of the proposed method for dataset II is 91.96% for protocol-1 and 95% for protocol-2. Length of the feature vector can be increased by decreasing the block size, which improves the recognition accuracy but cost to the increase of computational complexity.

References

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[2] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, Jan 2015.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, apr 2002.

[4] J. Triesch and C. von der Malsburg, "Classification of hand postures against complex backgrounds using elastic graph matching," *Image and Vision Computing*, vol. 20, no. 13–14, pp. 937–943, dec 2002.

[5] D. Kelly, J. McDonald, and C. Markham, "A person independent system for recognition of hand postures used in sign language," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1359–1368, aug 2010.

[6] M. Moghaddam, M. Nahvi, and R. H. Pak, "Static persian sign language recognition using kernel-based feature extraction," in *2011 7th Iranian Conference on Machine Vision and Image Processing*. IEEE, nov 2011, pp. 1–5.

[7] A. Just, Y. Rodriguez, and S. Marcel, "Hand posture classification and recognition using the modified census transform," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, April 2006, pp. 351–356.

[8] I. Rasines, A. Remazeilles, and P. M. I. Bengoa, "Feature selection for hand pose recognition in human-robot object exchange scenario," in *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*. IEEE, sep 2014, pp. 1–8.

[9] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Jun 1998.

[10] A. L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2d static hand gesture colour image dataset for asl gestures," *Research Letters in the Information and Mathematical Sciences*, vol. 15, pp. 12–20, 2011.

[11] E. Y. Lam, "Combining gray world and retinex theory for automatic white balance in digital photography," in *Proceedings of the Ninth International Symposium on Consumer Electronics, 2005. (ISCE 2005).*, June 2005, pp. 134–139.

[12] H. Kong, H. C. Akakin, and S. E. Sarma, "A generalized laplacian of gaussian filter for blob detection and its applications," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1719–1733, Dec 2013.

[13] M. Avraam, "Static gesture recognition combining graph and appearance features," *International Journal of Advanced Research in Artificial Intelligence*, vol. 3, no. 2, pp. 1–4, 2014.