Hand posture recognition using HOW and homogenous kernel

Reconnaissance de la posture de la main à l'aide du HOW et noyau homogène

Wassila Abadi*, Rachid Hamdi & Mohamed Fezari

* Electronics Departement, Badji Mokhtar University, PO Box 12, Annaba, 23000, Algeria.

Article Info

Article history:

Received 06/03/2018
Revised 02/02/2019
Accepted 03/02/2019

Keywords

arabic sign language - static hand gesture – how - one-vs-all support vector machine (svm).

ABSTRACT

In this paper, we present a static hand gestures recognition system of Arabic Sign Language alphabets. The proposed method uses Histogram Of visual Words (HOW) descriptor and Support Vector Machine (SVM). First, the images of static hand gestures are converted into HOW features and grouped using k-means clustering to create histograms. Then they are converted from non linear space into linear space using Chi-squared kernel. The result is fed into One-vs-All SVM classifier to build signs models. Training and test stages of this technique are implemented on hand postures images using cluttered backgrounds for different lighting conditions, scales and rotations. The proposed method shows a satisfactory recognition rate and achieves good real-time performance regardless of the image resolution.

Mots clés

langue des signes arabe - geste de la main statique - how - séparateur à Vaste Marges (svm).

RESUME

Dans cet article, nous présentons un système de reconnaissance des gestes statique de la main du Langage des Signes Arabe. La méthode proposée utilise un descripteur des histogrammes en mots visuels (HOW), et la classification en utilisant séparateur à vaste marges (SVM) avec un noyau homogène. Tout d'abord, les images des gestes statiques de la main sont converties en paramètres HOW et groupées en utilisant la classification k-means pour créer des histogrammes. Ensuite, ils sont convertis à partir d'un espace non linéaire en espace linéaire en utilisant le noyau Chi-carré (X^2). Le résultat est introduit dans le classificateur SVM pour créer des modèles de signes. Les étapes d'apprentissage et de test de cette technique sont implémentées sur des postures d'images en utilisant des fonds complexes pour différentes conditions d'éclairage, échelles et rotations. La méthode proposée montre un taux de reconnaissance satisfaisant et obtient de bonnes performances en temps réel quelle que soit la résolution de l'image.

Corresponding Author
Wassila Abadi
Electronics Departement, Badji Mokhtar University, PO Box 12, Annaba, 23000, Algeria.
Email: wassila.abadi@gmail.com

©UBMA 2019

1. INTRODUCTION

Sign language is the fundamental communication method between people who suffer from hearing defects [1, 2]. A translator is usually needed during communication between about 70 million deaf people and ordinary person to translate sign language into natural language and vice versa [3, 4, 5]. In addition, sign language is not universal. Different countries have different sign languages, for example: American Sign Language (ASL), Arabic Sign Language (ArSL), French Sign Language (FSL) and German Sign Language (GSL) have different alphabets and word sets [6, 7, 8]. The similarities among signs in a sign language are created by complex body movements, i.e., using the right hand, the left hand, or both. When signs are created using both hands, the right hand is more active than the left hand. Sign language speakers (signers) also support their signs with their heads, eyes, and facial expressions, movements, collection of gestures [3]. Many researches have been developed for hand gesture recognition of various sign languages. In this domain, there are two main approaches: glove based technique and computer vision based technique [9]. The glove based technique mainly utilizes sensory gloves to retrieve the joint angles for gesture features. According to Cemil & Leu [6], an electronic sensory gloves is used to recognize and translate ASL words into English and a neural network as a classifier. For Computer vision based technique, Hough transform and neural network have been used for ASL recognition [1]. Furthermore, Mohandes et al. [10] have designed a system for ArSL recognition using Hidden Markov Models (HMMs). A system for recognizing static gestures of alphabets in Persian Sign Language (PSL) using wavelet transform and neural network is presented in [2]. Zaki & Shaheen [11] have also proposed a combination of vision based features for ASL recognition tested on RWTH-50 databse. Al-Jarrah et al. [4], have developed a system for automatic translation of gestures in the ArSL using Adaptive Neuro-Fuzzy Inference System. Yang Quan [5] has proposed Chinese sign language recognition system based on video sequence appearance and Support Vector Machine (SVM) classifier, using spatio-temporal characteristics and image features (SIFT, Hu moments). Pahlevanzadeh et al. [12], have used Generic Cosine descriptor (GCD) as feature extraction method for the interpretation of the Taiwanese sign language. Nuwan Gamage et al., [13] have proposed Gaussian Process Dynamical Model as an alternative machine learning method for hand gesture recognition. They used 66 hand gestures from the Malaysian sign language to present the experimental results. Sinith et al. [14] have designed a system to recognize only six gestures using Sobel filter for edge detection. The edge coordinates are given as the input to SVM classifier. According to Chung et al. [7], a real time hand gesture recognition interface based on Haar wavelet is proposed. Zhao et al. [15], have used Histograms of Oriented Gradients (HOG) features and PCA-LDA to reduce the dimensionality of extracted features. The result is obtained by nearest neighbor classifier. Otiniano-Rodriguez et al. [8] have used Hu and Zenike moments and SVM classifier for sign language recognition. According to Dardas and Georganas [16], a videogame application is developed via hand gestures, where bag of features and multiclass SVM were used. Moment invariants (Hu moments) is used by Premaratne et al. [17] as an input of the neural network classifier to recognize numerals gestures in Australian sign language. The objective of this work is to propose an efficient and accurate system for sign language recognition using Histogram Of visual Words and Chi-squared Support Vector Machine. We use complex backgrounds with different: lighting conditions, scales, rotations and positions as shown in figure 1. The rest of this paper is organized as follows: in section 2 we present system overview, while section 3 gives details on the classification phase. The experimental results are summarized in section 4, then followed by a discussion in section 5. Section 6 concludes future work directions.

Figure 1. Selected signs against different complex backgrounds and different light conditions

©UBMA 2019

2. SYSTEM OVERVIEW

The proposed hand gesture system consists of two stages: training and recognition. The training framework is shown in figure 2.

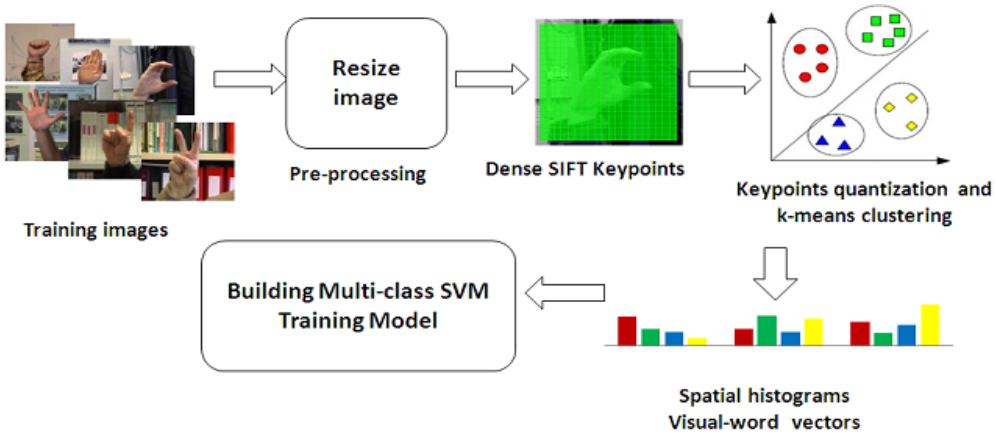

Figure 2 : Training stage diagram

The multiclass SVM models are built in the training stage to be used in the recognition one. For each test hand gesture, a model is built and compared with SVM training models in the classification phase as shown in figure 3.

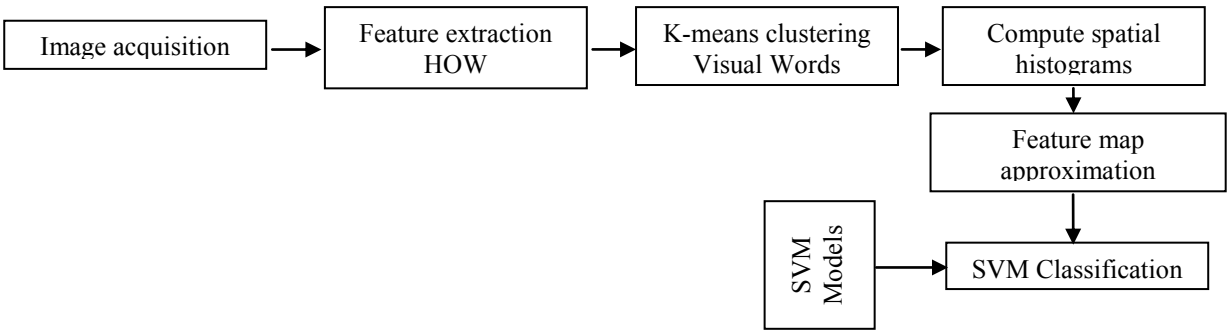

Figure3. Recognition Stage diagram

We chose Marcel Train images from Marcel dataset [18] as well as our Arabic sign language database for both training and testing stages. The recognition rate and the elapsed time for recognition per image are calculated. The bag-of-features model is built using HOW in feature extraction phase, learning a visual vocabulary by k-means clustering, quantizing features using visual vocabulary, and finally representing images by frequencies of visual words (visual histograms).

2.1. Features Extraction

Histogram Of visual Words (HOW) descriptor for appearance has been used to extract features. This technique works by partitioning the image into increasingly fine subregions and computing histograms of local features found inside each sub-regionthese local features are based on SIFT method [19]. The result is a simple and computationally efficient extension of an orderless bag of features image representation [20].

2.2. Key points Quantization

First, we need to establish a vocabulary of visual words. For this, we have sampled many local features from our training set and then clustering them with k-means. The number of k-means clusters is the size of our vocabulary. Vocabulary sizes for our different experiment scenarios are: 100, 200, 300, 500 and 700, based on related work and biography search we can say that these numbers are enough to validate the results.

©UBMA 2019

2.3. Spatial Histograms

It is a simple histogram obtained by taking a region of an image, assigning a label to each pixel (somehow; via some mapping function), and then computing a histogram of the labels. The histogram counts how many pixels received that label. We obtain at the end a feature vector: a vector of counts, one count per possible label. At this stage, training and testing images are presented with histograms of visual words. We simply count how many of each centroid (HOW descriptors) occurred in the visual words vocabulary for each image.

2.4. Building SVM Models

The last task is to train 1-vs-all linear SVMS. Linear classifiers are one of the simplest possible learning models. Chi-squared distance is calculated between histograms, and then the result is treated to one-versus-all SVM (support vector machine). From the histogram using SVM a set of one-versus-all classifiers can be generated.

3. CLASSIFICATION PHASE

We have used one-versus-all linear SVM, and treated the histograms representing each image of the selected signs as normalized vectors to a unit Euclidean norm. The homogeneous Chi-squared (X^2) kernel is used, because of its ability in multi-class classification such as in object and scene categorization, but instead of computing kernel values we will explicitly compute the feature map, so that the classifier remains linear (in the new feature space) as presented in equation 1.

$$K(x,y) = e^{-\gamma X^2(x,y)} \qquad (1)$$

where:

x,y : Finite dimensional histograms.

and:

$$X^2(x,y) = \sum_j \frac{(x_j - y_j)^2}{x_j + y_j} \qquad (2)$$

The X^2 kernel, which has been found yields the best performance in most applications, and the most compact feature representation [21]. Since non-linear SVMs can be seen as linear SVMs operating in an appropriate feature space, there is at least the theoretical possibility of extending such efficient learning methods to a much more general class of models. The success of this idea requires that the feature map can be efficiently computed, and the corresponding feature space is sufficiently low dimensional [22, 23]. The Chi-squared distance calculated between histograms of each class to create linear feature map kernel for Support Vector Machine (SVM) classifier.

4. EXPERIMENTS AND RESULTS

We report results on two diverse datasets: Arabic Sign Language dataset and Marcel Static hand posture Database [18]. We perform all processing in grayscale. All experiments are repeated with different randomly selected training and test images, and the recognition rate is recorded for each class. Multi-class classification is done with a support vector machine (SVM) trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response. The system was implemented in MATLAB version 7.12 under windows 64 bits and CPU characteristics with: 2.93 GHz Core2 Duo, and RAM 4 Giga-bytes with Windows7 system.

4.1. Data Collection

To improve the efficiency of our method, we conducted experiments on three different databases: Arabic sign language database, Marcel Static hand postures database and ArSL hand posture database.

4.1.1. Arabic Sign Language dataset

Our dataset of the Arabic sign Language alphabet contains 30 hand signs from the unified Arabic Sign Language Alphabet. The hand postures were performed by 5 different subjects against: uniform background (Dark) and different complex (cluttered) backgrounds. The subjects did not wear gloves and they performed the 30 letters at average of 6 times. The hand postures were taken on different light conditions, different scales and different orientations to increase the robustness of our system.

©UBMA 2019

4.1.2. Marcel Static Hand Posture Database

Marcel Static Hand Posture Database [18] contains: 6 hand postures (A (fist), B(palm), C, Five, Point(index), V (two), performed by 10 persons against different: backgrounds, scales, rotations and illuminations conditions; to increase the robustness of the SVM classifier model.

4.2. Results on Marcel Dataset

We have used Marcel Static Hand Posture Database presented in [18] for training and testing phases. The training stage was turned with 250 training images with different resolutions such as 384 X 288 pixels, 240X320 pixels, 155X155 pixels, 70X82 pixels, 66X76 pixels. The test stage was done with the rest of images of each class as presented in Table 1.

Training and test stages of this technique are implemented on hand postures images under differnt conditions as mentioned before as expressed by Abadi et al. [24].

Table 1. Performance of the system with cluttered background and different size images

Posture Name	Number of images	Recognition rate (%)	Average Recognition Time (second/image)
A	1044	84.09	0.015
B	202	92.07	0.025
C	287	90.24	0.017
Five	369	97.28	0.026
Point	1110	84.59	0.017
V	150	82.00	0.017

4.3. Results on our Dataset

We developed our own dababase for hand gestures. It is a challenging dataset, not only because of the large number of classes, but also because it contains images with highly variable poses and different cluttered backgrounds. The images were acquired with a diverse set of cameras. Histogram of visual words features have been extracted from images with uniform and black background for training and test phases, we took six classes only as: Alif, Ayn, Ba, Dal, Dhad, dhal; to be confronted with the problem of similarity of letters: Dal and Dhal. The results are shown in the table 2.

Table 2. Recognition rate of 6 Arabic sign Language letters(50x50 pixels) with uniform background

Letters	Recognition Rate (%)
Alif (أ)	100
Ayn (ع)	87.5
Ba (ب)	71 .42
Dal (د)	100
Dhad (ض)	100
Dhal (ذ)	60

We tested six hand gestures as shown in table 2. Our experiments are turned on different size of images: 170X170, 180X180, 160X120, 320X240. For images of size 50X50 pixels, we have used 10 images for training stage of each letter. Complex background was used for test and training phases. The PGM (Portable Gray Map) format of each letter image is used to reduce time of features extraction. The obtained results are presented as shown in table 3 for different number of words: 300, 500 and 700.

Table 3. Recognition rate of seven Arabic sign language letters (50x50 pixels) for different number of words

Arabic Sign Language	***Recognition Rate (300words) (%)***	***Recognition Rate (500words) (%)***	***Recognition Rate (700words) (%)***
Ba (ب)	90	83.33	80
Dal (د)	85.18	81.48	85.18
Lam (ل)	62.5	62.5	70.83

©UBMA 2019

Arabic Sign Language	***Recognition Rate (300words) (%)***	***Recognition Rate (500words) (%)***	***Recognition Rate (700words) (%)***
Miem (م)	35.48	38.71	32.23
Sad (ص)	54.16	50	62.5
Sien (س)	72.73	77.27	72.73
Ya (ي)	72	72	84

In this work, we have presented the recognition time including: the pre-processing time, feature extraction time, keypoints quantization time and building histograms time as presented in table 4.

Table 4. Average Recognition time of seven Arabic Sign Language letters (50X50 pixels) for 300 words

Letters	Average Recognition Time (second)
Ba (ب)	0.0238
Dal (د)	0.0112
Lam (ل)	0.0250
Miem (م)	0.0247
Sad (ص)	0.0217
Sien (س)	0.0369
Ya (ي)	0.0306

As presented in table 5, we perform tests on 30 signs of Arabic Sign Language against cited conditions, different rotations and different scales. The process were applied on 'jpeg' images but after conversion to grayscale and resize to 120x160 pixels. The selected parameters to achieve these results: number of clusters: 300, number of classes: 30, number of training images: 30images/class, Image size: 120x160, SVM Kernel: Humogenous Intersection Kernel.

Table 5. Recognition rate of 30 Arabic Sign Letters with cluttered background (120X160 pixels)

Letters	Recognition Rate (%)	Letters	Recognition Rate (%)
Alif (أ)	81.81	Tah (ط)	100
Ba (ب)	92.5	Thah (ظ)	97.22
Ta(ت)	86.84	Ayn (ع)	89.6
Tha (ث)	100	Ghayn (غ)	88.4
Jiem (ج)	97.62	Fa (ف)	92.86
Ha (ح)	100	Qaf (ق)	91.43
Kha (خ)	100	Kaf (ك)	86.84
Dal (د)	89.2	Lam (ل)	97.06
Dhal (ذ)	87.8	Miem (م)	90.24
Ra (ر)	94.44	Noon (ن)	88.6
Zay (ز)	94.87	He (هـ)	75
Sien (س)	100	Waw (و)	91.42
Shien (ش)	92.1	Ya (ي)	97.14
Sad (ص)	88.23	La (لا)	78.26
Dhad (ض)	97.06	T (ة)	84.21

According to the Table 6, the recognition rate increases when we increase the number of training images.

©UBMA 2019

Table 6. Recognition Rate Of Seven Arabic Sign Language Letters (90x90 Pixels) For Different Number Of Training Images For 200 Words

Posture Name	10 Training Images	20 Training Images
Ba (ب)	83.33	85
Dal (د)	92.59	94.11
Lam (ل)	75	78.57
Miem (م)	41.93	69.23
Sad (ص)	50	78.57
Sien (س)	81.81	83.33
Ya (ي)	68	86.66

5. DISCUSSION

According to the results shown in Table 1, the high classification rate 97.28% is obtained using HOW and X2 kernel SVM classifier and without skin detection compared with [16] where the high recognition rate was: 98.04% with skin detection, where they have extracted features only for hand posture, while we have extracted the features from image without skin detection and contour comparing. Also the obtained recognition time is the same with Marcel dataset compared to the obtained results in [16, 24].

According to the results shown in Table 5, the high classification rate 100% is obtained using HOW and SVM classifier and without skin detection.

For each sign, we can distinguish that the recognition rate varies from number of words to another and it depends on the background and objects that belong in the image, hence the high recognition rate of the letter 'Miem is obtained with the number of words 500, and the letters 'Ya', 'sad' and 'Lam' with the number of words: 700, and objects in the image have an influence in the feature extraction phase of the keypoints and the recognition rate.

The average recognition time varies depending on the surface of the sign in the image, and presence of objects in the complex background. The average recognition time presented includes: the pre-processing time, feature extraction time, keypoints quantization time, building histograms time and SVM classification time. Also lighting conditions and rotations are factors that affect recognition rate and recognition time as shown in Table 5.

As shown in Figure 4, the recognition rate increases when the vocabulary size increases. But it also depends on the image content and the complex background.

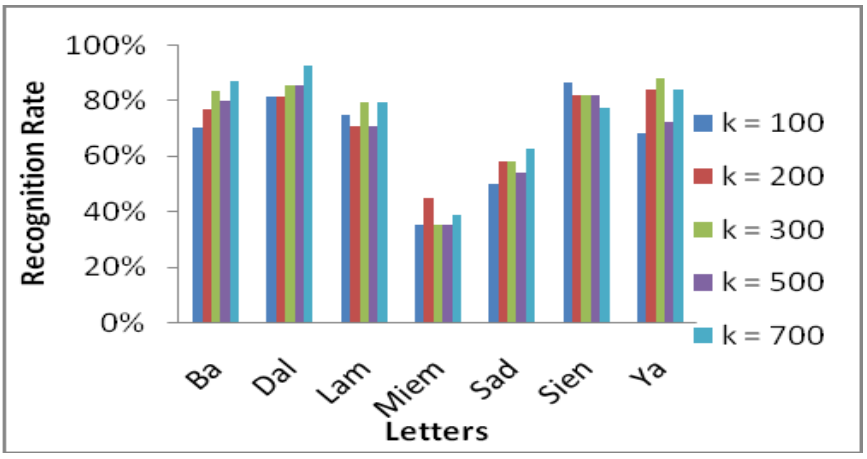

Figure 4 : Recognition Accuracy Of selected Arabic Sign Alphabets for different vocabulary sizes (80X80 pixels).

As presented in table 7, we use 30 train images per class only compared to other researches used 100 images / class for a number of classes less than 30classes; while the recognition rates obtained by our method is sufficient compared with [16].

©UBMA 2019

Table7. Performance Comparison with Other Approach

	Our Method	[16]
Recognition Time (Second/image)	0.0046	0.017
Number of postures	7	10
Recognition Rate	88.38%	96.23%
Number of Trainig images/posture	20	100
Scale	Invariant	Invariant
Rotation	Invariant	Invariant
Lighting Changes	Invariant	Invariant
Background	Cluttered	Cluttered
Feature extraction	From all image	From Hand only
Classifier	SVM One-Vs-All with Chi-Squared Kernel	SVM One-Vs-All
Features	Histogram Of visual Words	SIFT features

6. CONCLUSION

In this paper, we developed a sign language recognition system. The system has two stages: the feature extraction stage and the classification stage. In feature extraction phase, Dense-SIFT features are extracted from all image included the other objects in the cluttered background without skin detection, and BOVW are built to train the SVM classifier. The experimental results demonstrated that the BOVW histograms and Chi-squared kernel SVM leads to reach a high recognition rate of about 97.28%. The system were tested on two different dataset, one professional developed and used in [18], the other is our proper dataset with different background lights and different noises.

In the future work, further features extraction methods will be used to increase the recognition rate and other classification methods will be explored. The final objective of this work will be to implement it on portable device based on DSP or FPGA circuit and used as Human machine interaction.

REFERENCES

[1] Qutaishat M., Habeeb M., Takruri B. & Al-Malik H., 2007. American sign language (ASL) recognition based on Hough transform and neural networks, *Expert Systems with Applications*, vol. 32, pp. 24-37.

2] Karami A., Zanj B. & Sarkaleh A.K., Mar.2011. Persian sign language (PSL) recognition using wavelet transform and neural networks, *Expert Systems with Applications,* vol. 38, no. 3, pp. 2661-2667.

[3] Mitra S. & Acharya T., 2007. Gesture Recognition: A Survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 37, no. 3.

[4] Al-jarrah O. & Halawani A., 2001. Recognition of gestures in Arabic sign language using neuro-fuzzy systems, vol. 133, pp. 117–138.

[5] Yang Q., June 2010. Chinese sign language recognition based on video sequence appearance modeling.*5th IEEE Conference on Industrial Electronics and Applications*, pp. 1537-1542.

[6] Oz C. & Leu C. M., Oct. 2011. American Sign Language word recognition with a sensory glove using artificial neural networks.*Engineering Applications of Artificial Intelligence*, vol. 24, no. 7, pp. 1204-1213.

[7] Chung W.K., Wu X., and Xu Y., Feb. 2009. A realtime hand gesture recognition based on Haar wavelet representation. *IEEE International Conference on Robotics and Biomimetics*, pp. 336-341.

[8] Otiniano-Rodrıguez K. C., Cámara-Chávez G., Menotti D., 2012. Hu and Zernike Moments for Sign Language Recognition. *The 2012 Internat. Conf. on Image Processing, Computer Vision, and Pattern Recognition*, pp. 1-5.

[9] Pavlovic V.I., Sharma R. & Huang T.S., 1997. Visual interpretation of hand gestures for human-computer interaction:

©UBMA 2019

a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7.

[10] Mohandes M., Quadri S.I. & Deriche M., 2007. Arabic Sign Language Recognition an Image-Based Approach. *21st International Conference on Advanced Information Networking and Applications Workshops*, pp. 272–276.

[11] Zaki M.M. & Shaheen S. I., March 2011.Sign language recognition using a combination of new vision based features. Pattern Recognition Letters, vol. 32, no. 4, pp. 572–577.

[12] Pahlevanzadeh M., Vafadoost M. & Shahnazi M., February 2007. Sign language recognition. *9th International Symposium on Signal Processing and Its Applications*, pp. 1-4.

[13] Gamage N., Kuang Y.C., Akmeliawati R. & Demidenko S., November 2011. Gaussian Process Dynamical Models for hand gesture interpretation in Sign Language. *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2009–2014.

[14] M.S. Sinith, Soorej G. Kamal, Nisha B., Nayana S., Kiran Surendran & Jith P.S., August 2012. Sign Gesture Recongnition Using Support Vector Machine. *International Conference on Advances in Computing and Communications*, pp. 122–125.

[15] Zhao Y., Wang W. & Wang Y., September 2011. A real-time hand gesture recognition method. *2011 International Conference on Electronics, Communications and Control*, pp. 2475-2478.

[16] Dardas N.H & Georganas N.D., November 2011. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 11, pp. 3592–3607.

[17] Premaratne P., Yang S., Zou Z.M. & Vial P., 2013. Australian Sign Language Recognition Using Moment Invariants. *in Intelligent Computing Theories and Technology SE - 59, of Lecture Notes in Computer Science*, vol. 7996 pp. 509-514. Springer Berlin Heidelberg.

[18] http://www.idiap.ch/resource/gestures/ visited: 01/10/2013

[19] Lowe D.G., November 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110.

[20] Dahmani D., Larabi S., 2014. User-independent system for sign language finger spelling recognition, *Journal of Vision Communication and Image Representation,*vol. 25 (5) 1240–1250

[21] Vedaldi A. & Zisserman A., 2011. Efficient Additive Kernels via Explicit Feature Maps, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. , pp. 480-492.

[22] Vempati S., "Generalized RBF feature maps for Efficient Detection," 2010.

[23] Lazebnik S., Schmid C., & Ponce J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2169-2178.

[24] Abadi W., Fezari M. & Hamdi R., Nov. 2015. Bag Of Visual Words and chi-squared kernel support vector machine : a way to improve hand gesture recognition, *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication,* Article No. 91.

©UBMA 2019