Machine Learning for Hand Gesture Recognition Using Bag-of-words

1st Marouane Benmoussa
LIMIARF, Faculty of Science,
Mohammed V University
Rabat, Morocco
benmoussa.marouanee@gmail.com

2nd Abdelhak Mahmoudi
LIMIARF, Ecole Normale Supérieure,
Mohammed V University
Rabat, Morocco
abdelhak.mahmoudi@um5.ac.ma

***Abstract*—Human Computer Interaction received a great deal of attention this last decade. Last researches has turned to more natural interaction systems like gestural human machine interfaces. Recent works are attempting to solve the problem of hand gestures recognition using machine learning methods. Some of them are pretending to achieve very high performance. However, few of them are taking into account mandatory requirements to apply the workflow of a learning model, mainly data unbalance, model selection and generalization performance metric choice. In this work, we proposed a machine learning method for real time recognition of 16 gestures of user hands using the Kinect sensor that respects such requirements. The recognition is triggered only when there is a moving hand gesture. The method is based on the training of a Support Vector Machine model on hand depth data from which bag of words of SIFT and SURF descriptors are extracted. The data was kept balanced and the model kernel and parameters were selected using cross validation procedure. The method achieved 98% overall performance using the area under the ROC curve measure.**

***Index Terms*—Hand Gestures Recognition, SIFT, SURF, Bag of visual words, Kinect.**

I. Introduction

Gesture recognition has been a subject of much study lately, especially with the advance of Human-Computer Interaction (HCI) technology that led to a better experience. This last can be improved by familiarizing computers more with human ways of communications. The idea is to make computers understand human language (speech, facial expressions, and human gestures).

This paper focuses on gestures since it is a rich communicative way that is used to usually express ideas, needs or feelings. This guarantees a more friendly human-computer interface that is essentially due to the effectiveness of gestures to communicate precisely between a sender and a receiver. This can be easily shown by stating some of the very decisive examples; many people travel to foreign countries even if they dont know the official language of the visited country and still manage to communicate using gestures and sign language. More importantly, the gestures are the main language used for people who suffer from hearing and speech impairment. This demonstrates that gestures can be considered international and employed almost everywhere in the world. Additionally, there are situations when silent communication is preferred. For example, during a surgery, a doctor may gesture to the nurse for assistance. Thus the interest to make a real-time hand gesture recognition (HGR) system is inevitable.

Several works have been done in computer vision about HGR, but just few systems were done based on Kinect camera [8], [9]. Further more, limited number of gestures were recognized [10], [11]. Recent works are attempting to solve the problem of hand gestures recognition using machine learning methods. Some of them are pretending to achieve very high performance [4], [13]. However, few of them are taking into account mandatory requirements to apply the workflow of a learning model, mainly data unbalance, model selection and generalization performance metric choice. That is the main motive behind this paper which is based on a novel method for contact-less HGR using both bare hands or just one hand to recognize 16 gestures. The system is able to detect hand gestures, track them using Skeletal Tracking of kinect [12] and recognize their meanings. The real power of the system is its ability to give high accuracy despite the lighting conditions, skin color clothing and background. The model is also optimized to make no calculation unless there is a body movement.

The approache that was implemented in this paper is based on features extraction and machine learning methods. Uses Speeded Up Robust Features (SURF), and Scale Invariant Features Transform (SIFT), trained with both K-means and Support Vector Machine (SVM) classifiers.

II. Related Work

In order to choose the most suitable descriptor for our HGR system we must choose wisely what kind of descriptor can give the most promising results, as such HOG (Histograms of Oriented Gradients) feature that has been introduced into gesture recognition in several works such as [1], but the HOG features is not ideal for gesture recognition for its computation in dense grids at some single scale without orientation alignment. In the other hand, other descriptors such as SIFT and SURF have shown good achievements over the years in object recognition and thus, are a good candidates for our system.

A. The Scale Invariant Feature Transform (SIFT)

SIFT it the most forceful descriptor allowing detection, description of image interest points and identifying similar

keypoints in different images. The extracted features are invariant against scale, rotation, noise and illumination. D.G. Lowe et al. [2] urges on concatenating multiple SIFT descriptors which gives additional invariance to occlusion and clutter.
To get a SIFT descriptor, an image pursues the following steps: First, Scale-Space Extrema Detection which consists of building a scale space, that is basically a progressive process of convolution between the images with different sizes and a Gaussian kernel. Second, the algorithm approximates Laplacian of Gaussian (LoG), by subtracting each successive images. Once it localizes Keypoints based on their stability, the algorithm assign to each detected Keypoint an orientation by collecting gradient directions and magnitudes around it. Finally, the algorithm generates a very characteristic imprint for each Keypoint.

Many researches have used SIFT descriptor to implement an HGR system. [4], [3]. C.C.Wang et al. improved the acccuracy of Adaboost learning model from 90% to 95% by training SIFT features on three postures classses: "fist", "palm" and "six". They accomplished hand posture recognition in real time, studying and applying different features such as contrast context histogram [3].

N.Dardas et al. used Bag-of-Words representation (BoW) and SIFT features. Basically, the idea is to use a dictionary of local interest points (keypoint) of an image or set of images, where the object is represented as a bag of words. These points are features and they are discriminative and very stable to rotations and scale changes. The authors used BoW model for gestures classification of four postures: "Fist", "Index", "Little", "Palm". They have built a vocabulary codebook of many visual words and represent every image as a histogram of the frequency words that are in the image.
These histograms are then fed into multi class SVM for classification which gave an accuracy above 90% [4]. Although their system gave efficient results, it can do really bad during scale changing, especially with the poor segmentation of hand region. Moreover, tracking information is not available so that the subject cannot move while performing gesture which is not practical in real world use.

B. Speeded Up Robust Features (SURF)

SURF is mainly known for its fast calculations [5]. It is inspired from SIFT algorithm which is the precursor in the extraction domain of keypoints. The main difference is in the dimension of feature descriptor that is 64 against 128 for SIFT. SURF uses Haar wavelet response to compute gradient orientation and gradient magnitude unlike SIFT that uses gaussien kernels.
Many research has been done using SURF descriptor. U.Zhang et al. presents SURF feature modified by gray threshold segmentation for detecting gesture features. Their method improve the accuracy of recognition especially with SVM classifier [7].
Guoming et al. tried to uses both SIFT and SURF with the concept of BoW and classify the histograms using an SVM classifier. However, the collected data is unbalanced which gives an overly optimistic results especially with the use of accuracy as an evaluation metric. Furthermore, their method cannot detect the hand outside a small range which force signer to stay in a closed range, and finally, their method is unable to track the hand movements due to the poor segmentation [13] .

Unlike the state-of-the-art, the main contribution of this research work is using the significance of bag of words vector based on SIFT and SURF features that are invariant against scale and rotation in order to build a recognition model of 16 gestures using both hands with the possibility to add complex gestures. To meet real time requirements, the features are extracted from depth data of the users hands only. This last is detected and tracked based on the minimum depth in a range of 4 meters long. The classification of gestures of the detected hand is triggered only when there is a moving body performing in front of the sensor.

III. HGR Recognition System

A. Data Acquisition

To build our HGR, we first used Microsoft Kinect sensor to acquire depth images for 16 different gestures and build a datasets of 8000 images, 500 for each gestures (Figure 1). The depth sensor of the Kinect camera can give distance data until 4 meters long. Using depth data can eliminates all sorts of confusion related to background. This also filters the cluttered background or overlapped images and illumination changes. In addition, the images of depth does not show enough contrast variations in the hand region in order to get more discriminative keypoints to differentiate close gestures.

For a better resolution, the user must be in an optimal distance which is practically between [0.8m, 2,5m]. Of course, the closer the better. We chose maximal resolution of 640 $\times$ 480 to have clear details of the hand's contour. This choice greatly reduce the number of operations, and consequently will improve the efficiency of the HGR system.

B. Hand Detection and Tracking

To detect and track the hands we followed three steps, as shown in Figure 3:

1) Decide which pixels are going to be taken into the account to carry on the tracking. The Kinect can catch the distance of the points which are visible to the camera, between the values D_{min} and D_{max} (see Figure2). The minimum and maximum depth are calculated. The depth pixels between these two values are maintained. This method allows to have greater mobility, because it does not force the user to stay in the same position while performing gestures. Second, this tolerance value T will allow user to capture the entire hand rather than the closer parts.
2) Apply the floodfill algorithm to find all the connected image regions [6]. First we notice that there are regions that do not belong to the hand, such as what appears to be one of user arms or some furniture or user body when hand is closer to its body. These objects just happen to

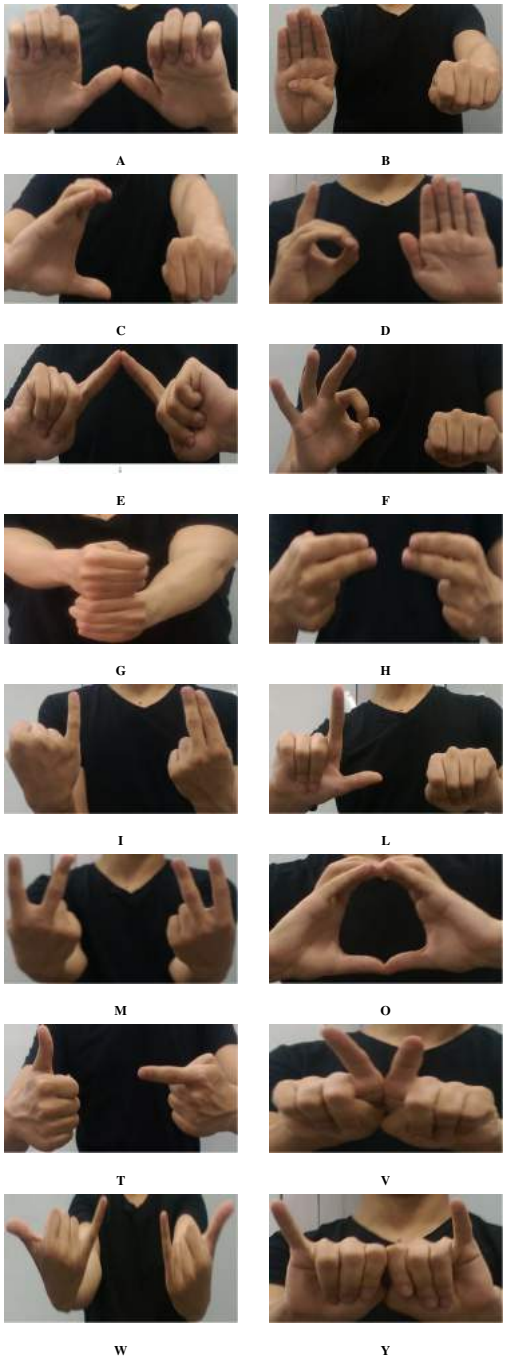

Fig. 1. Vocabulary dataset

be on the same depth layer as user arm and hand. This means that the distance between these objects and the Kinect lies between the range :

$$[D_{min}, D_{min} + T] \pm E$$

where E is the error commited by the kinect, and T is the tolerance value.

3) The same process is done with the other hand. Making an OR operation to the two images, we end up having two hands isolated (Figure 3).

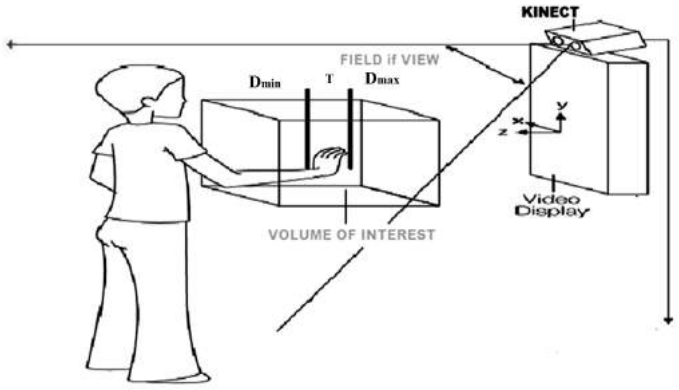

Fig. 2. Relative depth thresholding

Motion capturing of hand articulation is a challenging task from the moment that the hand can move with very high freedom degrees this cause a rotation variance. Additionally, our system is based on relative depth which allow user to move around the camera, this induces a problem of scale variance. To solve these issues, we chose the best feature descriptor that are most suitable for such a task, which are SIFT and SURF descriptors both scale and rotation invariant.

C. Gesture Recognition Using SURF and SIFT

Hand gestures training images can be represented by a sets of keypoints generated by SIFT and SURF, but the number of keypoints from the images are different and lack meaningful ordering. To address this problem, we use the bag-of-words approach. Bag-of-words (Bow) which is one of the most famous approaches in computer vision. Basically, the important points in images are the visual words. These points are features and they are discriminative and very stable to rotation and scale changes.

These features are extracted using SURF or SIFT descriptor forming a database of visual words then every new image is going to be represented as a histogram of the visual words that appears in the image.

This is done using the vector quantization clustering technique Where each cluster represent a visual word that corresponds to a unique local pattern shared by the keypoints in that cluster. Figure 4 represents a flowchart for the structure of the algorithm that uses BoW approach to classify gestures. In block A, the extracted keypoints are clustered and generate a codewords dictionary, then compute a histogram using visual words (vocabulary) for all the training samples. These histograms are then fed to an SVM classifier. In block B, test images are converted to histogram representation by counting how many keypoints appeared in each cluster, and then classify the results using the trained SVM model.

IV. Results and Discussion of SIFT and SURF descriptors

In this section we are interested in maximizing the performances of SVM using k-Fold cross validation to make model selection. With k=10, it has ensure a great tuning for the

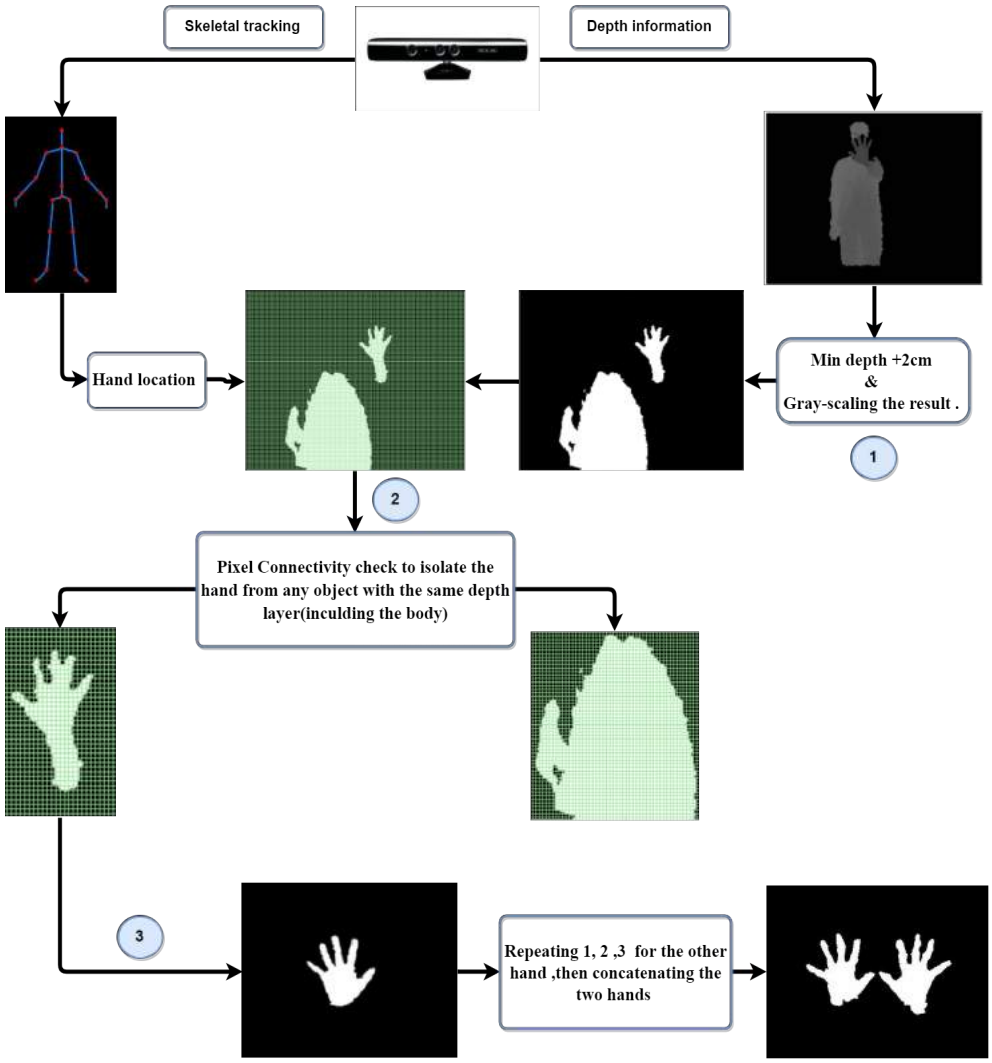

Fig. 3. The three steps of hand segmentation process

model parameters, we have chose Area Under The Receiver Operating Characteristic Curve (AUC ROC) to be our main evaluation metric. Since the method can be used for binary classification, therefore, we used *one versus all* method to convert the problem to a binary task.

A. Training, Validation and Testing process

Before the training process, in order to keep the data balanced, we collected 500 images for each of the 16 gestures. Next, we divided our dataset into three sets; the training set consists of 300 images for each gesture with a total of 4800 images, the validation set has 100 images with a total of 1600 images and finally, the test set of 100 images with total of 1600 images.

SURF and SIFT are extracted from training set with different scales and orientations. K-means clustering is then applied in order to build the visual words. Each one is considered as centroid of a subgroup of the same features. A histogram of visual words is then generated, this basically means getting frequencies of visual words in the current image (Figure5).

The extracted features are then fed to a Support Vector Machine (SVM) classifier. This last uses a non-linear mapping to transform original training dataset into higher dimension and searches for a linear optimal separating hyperplane. Since our gestures can have the same shape and certainly often have the same features for two different gestures, this will make linear separation hard and can lead to false results. Alternatively, the kernel method can be used to transform data into higher dimension space in which they are separable.

k-fold cross validation is applied to the validation set to tune the SVM parameters, using both linear and RBF kernels. Also, K, that represents the number of clusters (codebook$'s$ size), plays a role in tuning our model, the result will be shown and discussed in the next sections. Finally the test set is used to measure the generalization performance of the trained SVM model.

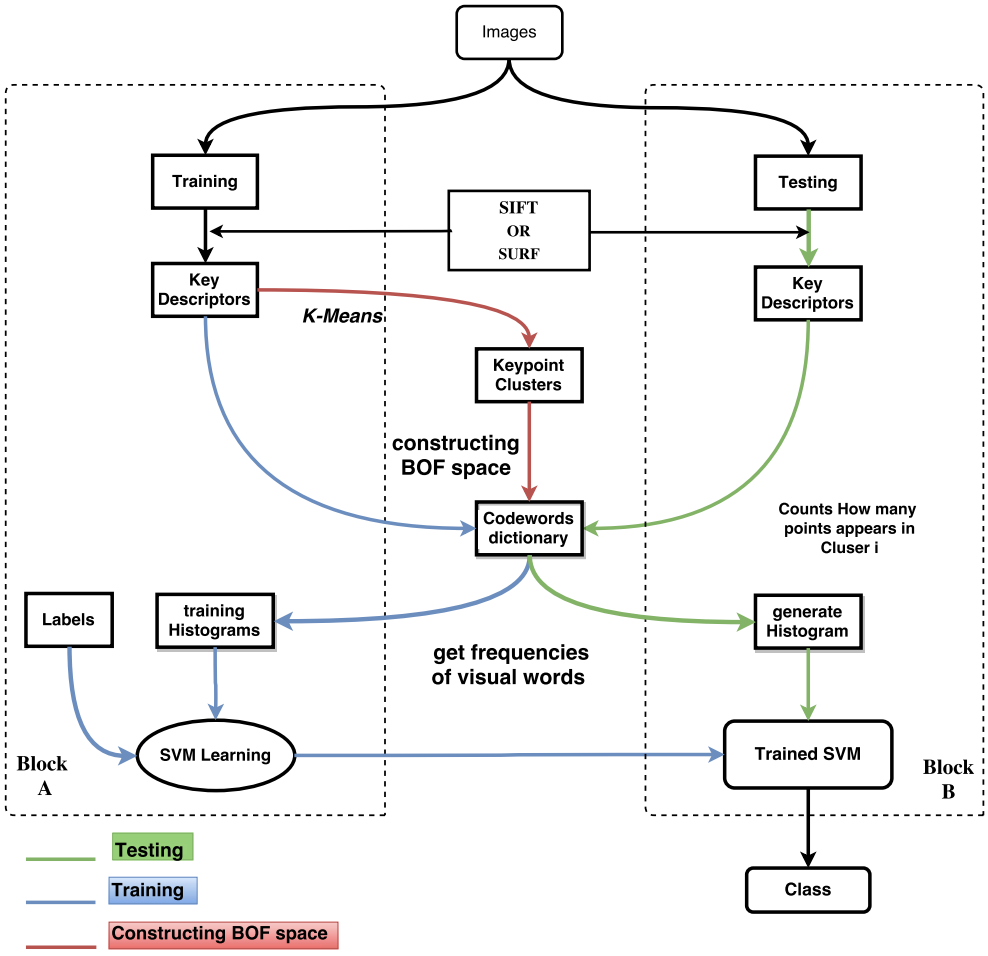

Fig. 4. Flowchart for the structure of the classification algorithm

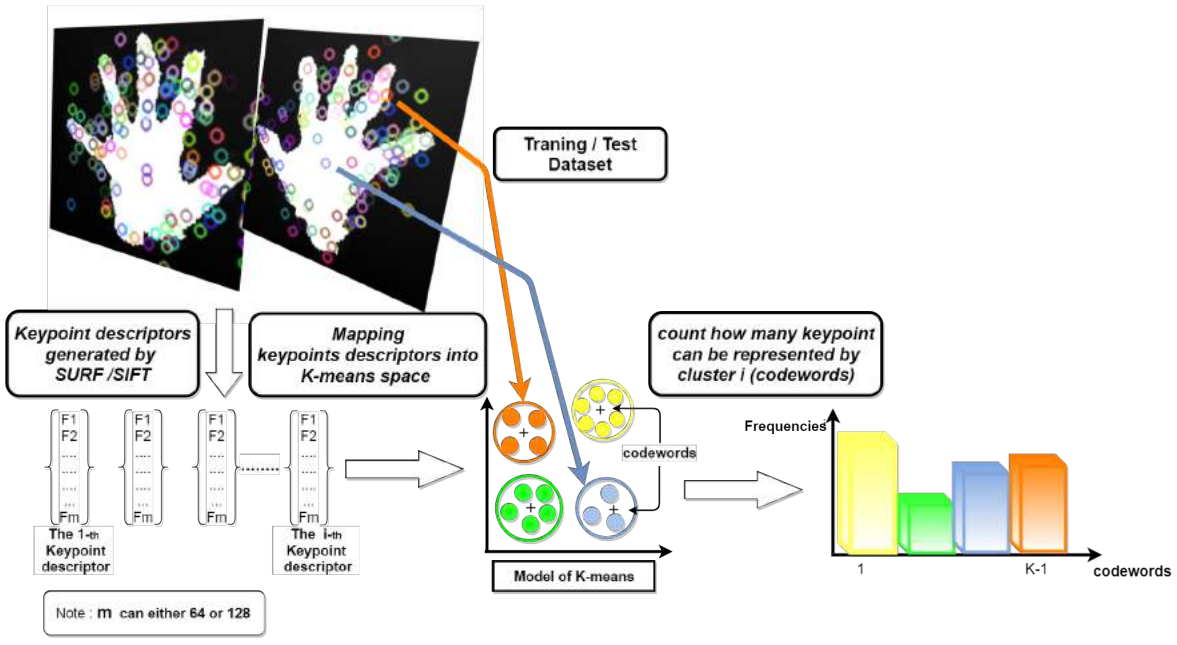

Fig. 5. Generating histogram for training image and test image by Mapping keypoint into Bag-of-Features space (bag of visual words space)

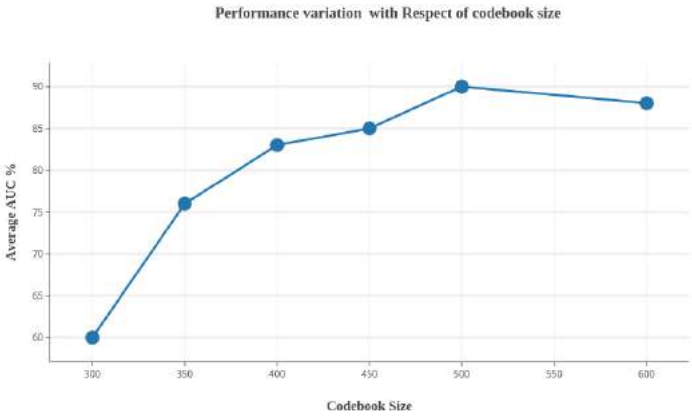

Fig. 6. Performance variations with respect to the size of codebook

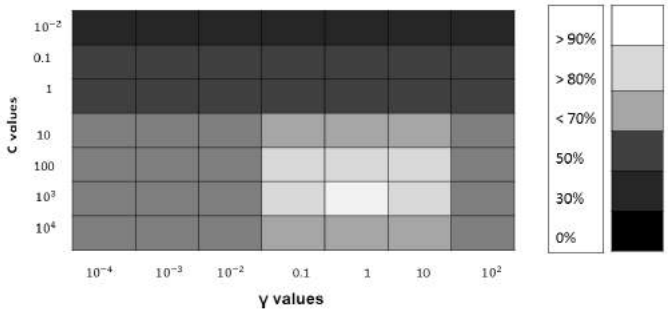

Fig. 7. Gridsearch for the optimal C and γ values that maximize the performance of SURF descriptor

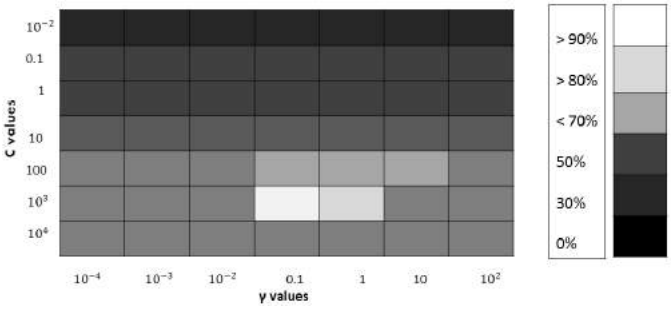

Fig. 8. Gridsearch for the optimal C and γ values that maximize the performance of SIFT descriptor

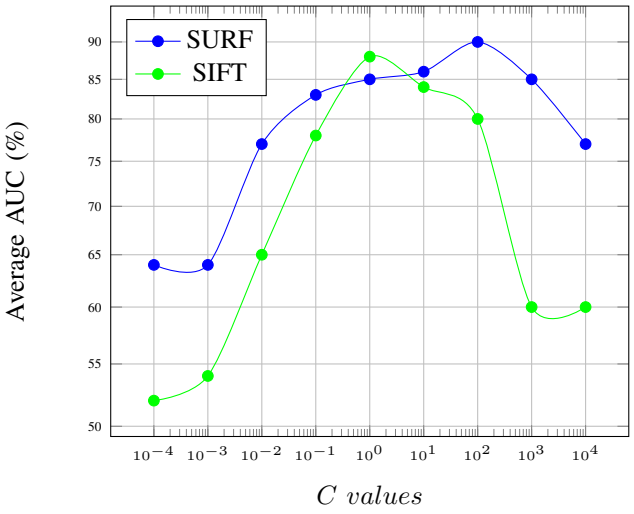

Fig. 9. The choice of hyper-parameter C for SVM Linear

B. Generalization Performance Using Area Under ROC

Receiver Operating Characteristics (ROC) is a method that visualizes, organizes and selects models based on their performance. This method is for binary classification, but can be adapted to multi-classification problem using "One Versus All" approach. It basically visualizes the classifier performance through a set of thresholds. The area which quantifies how well the classifier perform is the area under the curve (AUC). The AUC is actually the probability that a classifier will class a randomly chosen positive instance greater than a randomly chosen negative one. Generally, if AUC is lower than 0.5 (straight line)) it is similar to a random guessing.

C. Performance variation in function of the codebook's size K

The codebook size will be determined depending on the type of the data processed. There will be a sort of compromise for choosing the number of clusters. If it is too large, then there will be an overfitting because of insufficient samples of the keypoints extracted from the training image. If it is too small, then each bag of words vector will not represent all the keypoints extracted from its related image.
Since the structure of our data has not a lot of keypoint variations, as a matter of fact, the hands are the only object detected in all the scenes. We will then increase K until finding the best AUC.
In the figure 6, we show that K = 500 is the best choice (AUC 90%).

D. SVM model selection

In order to compare SIFT and SURF recognition performances, we have tuned the parameters of two kernels of SVM classifier : The linear kernel and Radial Basis Function (RBF) kernel. To find the optimal value that maximizes the SVM performance with linear kernel, we varied the hyper-parameter $C \in [10^{-4}, 10^{4}]$ (Figure 9). $C = 100$ is proved to be the best for linear SVM using SURF (90%) and $C = 1$ using SIFT (88%). For RBF kernel, we have tuned using two parameters : the hyperparameter C in $[10^{-2}, 10^{4}]$ and the parameter γ in $[10^{-4}, 10^{2}]$. $C = 10^{3}$ and γ = 1 were proved to be the best for RBF kernel using SURF (98%) (Figure 7) . and $C = 10^{3}$ and γ = 0.1 using SIFT (91%) (Figure 8).

E. SURF and SIFT comparaison

In order to compare the speed of recognition using our trained model on bag of words of SIFT and SURF descriptors, we took 8000 images (500 image for each class) as shown in the table I. SIFT is three times slower than SURF, which gave lag during feature extraction. On the other hand, SURF performs well on depth and blurry images. SIFT consumed

TABLE I
COMPARAISON OF DIFFERENT DESCRIPTOR

	SIFT	**SURF**
Average time (in seconds) consumed for features detection and description Per Image	**0.30s**	**0.12s**
Speed	**Very Low (0.5s-0.6s)**	**Medium To Low (0.25s-0.3s)**
number of features extracted per image	**43**	**107**
Rotation Invariant	**Good**	**Very Good**
Scale Invariant	**Good**	**Good**
Translation Invariant	**Very Good**	**Very Good**

0.30s on detecting and extraction, with an average of 43 features per image, against SURF that spends 0.12s to detect and extract 107 features per images. Also SURF was capable of describing gesture keypoint better than SIFT by difference of 50 keypoints. It can gives an idea about the descriptor that suits best our data and can extract as much information as needed for recognition. Last but not least, SURF proved to be efficient against scale, rotation and translation variance.

V. CONCLUSION

In this work, we proposed a machine learning method for real time recognition of 16 gestures of user hands using the Kinect sensor. The method is based on the training of a Support Vector Machine model on hand depth data from which bag of words of SIFT and SURF descriptors are extracted. The recognition is triggered only when there is a moving hand gesture in front of the camera. The data was kept balanced and the SVM model kernel and parameters were tuned using cross validation procedure. The method achieved a performance of 98% for SURF, and 91% for SIFT computed using the area under the ROC curve measure. We can conclude that SURF is three times faster than SIFT, which make it more suitable for real time application.

REFERENCES

[1] Marouane Hamda and Abdelhak Mahmoudi. Hand gesture recognition using kinect′s geometric and hog features. In *Proceedings of the 2Nd International Conference on Big Data, Cloud and Applications*, BDCA'17, pages 48:1–48:5, New York, NY, USA, 2017. ACM.

[2] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[3] Chieh-Chih Wang and Ko-Chih Wang. Hand posture recognition using adaboost with sift for human robot interaction. In *Recent progress in robotics: viable robotic service to human*, pages 317–329. Springer, 2007.

[4] Nasser Dardas, Qing Chen, Nicolas D Georganas, and Emil M Petriu. Hand gesture recognition using bag-of-features and multi-class support vector machine. In *Haptic Audio-Visual Environments and Games (HAVE), 2010 IEEE International Symposium on*, pages 1–5. IEEE, 2010.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[6] Shane Torbert. *Applied computer science*. Springer, 2016.

[7] Runqing Zhang, Yue Ming, and Juanjuan Sun. Hand gesture recognition with surf-bof based on gray threshold segmentation. In *Signal Processing (ICSP), 2016 IEEE 13th International Conference on*, pages 118–122. IEEE, 2016.

[8] Mazlina Abdul Majid and Jasni Mohamad Zain. A review on the development of indonesian sign language recognition system. 2013.

[9] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.

[10] Thad Starner, Joshua Weaver, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[11] Paranjape Ketki Vijay, Naphade Nilakshi Suhas, Chafekar Suparna Chandrashekhar, and Deshpande Ketaki Dhananjay. Recent developments in sign language recognition: A review. *Int. J. Adv. Comput. Eng. Commun. Technol*, 1(2):21–26, 2012.

[12] Y. Choubik and A. Mahmoudi. Machine learning for real time poses classification using kinect skeleton data. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, pages 307–311, 2016.

[13] Guoming Chen, Qiang Chen, Yiqun Chen, and Xiongyong Zhu. Hand gesture recognition via bag of visual words. In *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on*, pages 1159–1163. IEEE, 2016.