Hand Gesture Recognition from Surveillance Video Using Surf Feature

Dona Merin Joy[1], Dr. M.V Rajesh[2]

[1]M.tech-student, College of Engineering Poonjar, Kerala, India
[2]Department of Electronics & Communication, College of Engineering Poonjar, Kerala, India

Abstract— *The Sign language is very important for people who have hearing and speaking deficiency. It is the only mode of communication for such people to convey their messages. In this paper, a feature detection scheme is introduced called SURF, which stands for Speeded Up Robust Features. The real time images will be captured first and then stored in directory and on recently captured image and feature extraction will take place to identify which sign has been articulated by the user through SURF algorithm. A Text to Speech conversion is also included in this project. An experimental result shows that the proposed approach performs well with low time.*

Keywords— speeded up robust feature, Convolutional neural network, Text to speech, Deep dynamic neural network.

I. INTRODUCTION

In recent years, human action recognition has drawn increasing attention of researchers, primarily due to its potential in areas such as video surveillance, robotics, human computer interaction, user interface design, and multimedia video retrieval.

The works on video based action recognition focused mainly on adapting hand-crafted features [1],[2],[3] .These methods usually have two stages, an optional feature detection stage followed by a feature description stage. Well known feature detection methods are Harris3D[4],Cuboids[5],and Hessian3D[6].For descriptors, popular methods are Cuboids[7] , HOG/HOF[4] , HOG3D[8] and Ex-tended SURF[4] . In the recent work of Wang et al. [9], dense trajectories with improved motion based descriptors and other hand crafted features achieved state of the art results on a variety of datasets. Based on the current trends, challenges and interests with in the action recognition community, it is to be expected that many successes will follow.

However, the very high dimensional and dense trajectory features usually require the use of advanced dimensionality reduction methods to make them computationally feasible. The best performing feature descriptor is dataset dependent and no universal hand engineered feature that outperforming all others exists. This clearly indicates that the ability to learn dataset specific feature extractors can be highly beneficial and further improve the current state of the art. For this reason, even though hand crafted features have dominated image recognition in previous years, there has been a growing interest in learning low level and mid level features, and either in supervised, unsupervised, or semi supervised settings. Since the recent resurgence of neural networks invoked by Hinton et al. [14], deep neural architectures have become an effective approach for extracting high level features from data.

In the last few years deep artificial neural networks have won numerous contests in pattern recognition and representation learning. Schmidhuber [15] compiled a historical survey compactly summarizing relevant works with more than 850 entries of credited papers. From this overview we see that these models have been successfully applied to aplethora of different do-mains, the GPU based cudaconvnet implementation [16], also known as AlexNet, classifies 1.2 million high resolution images into 1,000 different classes. Multi column deep neural networks [17] achieve near human performance on the handwritten digits and traffic signs recognition benchmarks. 3D convolutional neural networks recognize human actions in surveillance videos. Deep belief networks combined with hidden Markov models for acoustic and skeletal joints modeling outperform the decade dominating paradigm of Gaussian mixture models (GMM) in conjunction with hidden Markov models. Multimodal deep learning techniques were also investigated to learn cross modality representation, for instance in the context of audio visual speech recognition. Recently Baidu Research proposed the Deep Speech system that combines a well optimized re-current neural network (RNN) training system, achieving the lowest error rate on a noisy speech dataset. Across the mentioned research fields, deep architectures have shown great capacity to discover and extract higher level relevant features.

Gestures are basically the physical action forms performed by Gestures are basically the physical action form performed by a per-son to convey some meaningful information. Gestures are a powerful means of communication among humans. In fact

gesturing is so deeply rooted in our communication that people often continue gesturing when speaking on the telephone. There are various signs which express complex meanings and recognizing them is a challenging task for people who have no understanding for that language. Sign language is categorized in accordance to regions like Indian, American, Chinese, Arabic and researches on hand gesture recognition, pattern recognitions, image processing have been carried by supposedly countries as well to improve the applications and bring them to the best levels. This Gesture recognition system can be further used in many application like home automation, banking, Robotics like autonomous robot to test its usability in the context of a realistic service task, smart watches. We can control application like Air Conditioner, TV, Mixer, Fan, Lights, etc. through gesture without use of switch. Even by using special image compression method, we can reduce the recognition time and design the specific chip for the same.

FIGURE 1: THE DIFFERENT SIGN LANGUAGES

In this paper, we have SURF feature that is speeded up robust features which is a patented local feature detector and descriptor. It can be used for tasks such as object recognition, image registration, classification or 3D reconstruction. It is partly inspired by the scale invariant feature transform (SIFT) descriptor.

To detect interest points, SURF uses an integer approximation of the determinant of Hessian blob detector. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. We are also converting the detected gesture in to audio using Text To Speech.

II. METHODOLOGY

The Hand Gesture Recognition using surf feature consists of Training and Testing videos. The training videos extracted as training frames and Testing video is extracted as testing frame. After the preprocessing step we are finding the best match of test image with training images. For that we have to identify the interest points or the important points. By using SURF Feature we detect the interest points, by using Hessian Matrix. Here the interesting points are blob like structures. The following block diagram explains the working of the proposed system.

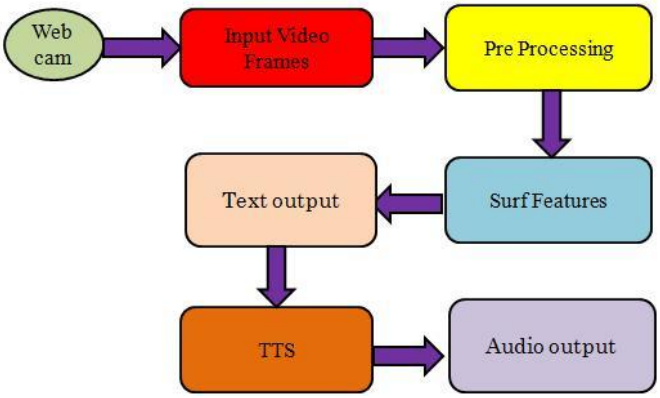

FIGURE 2: THE BLOCK DIAGRAM OF PROPOSED SYSTEM

2.1 Webcam &Input Frames

A webcam is essentially just a camera that is connected to a computer either directly or wirelessly and gathers a series of images. Here the train images are extracted from training video and the test images are extracted from testing video by using snapshot.

FIGURE 3: THE TRAINING VIDEO FRAMES

FIGURE 4: THE TESTING VIDEO FRAME

The above figures represent the training frames and testing frames. We are recognizing the hand gestures here. Therefore we have to crop the particular part of hand in pre-processing.

2.2 Pre-Processing

The aim of preprocessing is an improvement of the image data that enhance some image features important for further processing. We are cropping the input image frame to get the gesture from it, and thereby we reduce the high dimensionality.

2.3 Surf Feature

The main working of the gesture recognition is based on SURF. Speeded up robust features (SURF) is a patented local feature detector and descriptor. It can be used for tasks such as object recognition, image registration, classification or 3D reconstruction. It is partly inspired by the scale-invariant feature transform (SIFT) descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. There are SURF Detector, SURF Descriptor, and SURF Comparator.

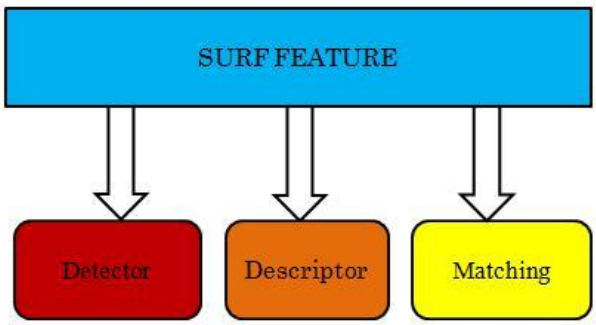

FIGURE 5: SURF DETECTOR, SURF DESCRIPTOR, AND SURF COMPARATOR.

2.3.1 Surf Detector

The SURF detector focuses its attention on blob like structures in the image. These structures can be found at corners of objects, but also at locations where the reflection of light on specular surfaces is maximal blob detection methods are aimed at detecting regions in a digital image that differ in properties, such as brightness or color, compared to surrounding regions. Informally, a blob is a region of an image in which some properties are constant or approximately constant.

Given some property of interest expressed as a function of position on the image, there are two main classes of blob detectors: (i) differential methods, which are based on derivatives of the function with respect to position, and (ii) methods based on local extrema, which are based on finding the local maxima and minima of the function. With the more recent terminology used in the field, these detectors can also be referred to as interest point operators, or alternatively interest region operators.

Gaussian derivative filters could be used to locate features. Specifically, we can detect blobs by convolving the source image with the determinant of the Hessian (DoH) matrix, which contains different 2D Gaussian second order derivatives. This metric is then divided by the Gaussians variance, σ2, to normalize its response.

$$DoH(x,y,\sigma) = \frac{G_{xx}(x,y,\sigma)\cdot G_{yy}(x,y,\sigma) - G(x,y,\sigma)^2}{\sigma^2} \tag{1}$$

$$G_{ij}(x,y,\sigma) = \frac{\partial N(0,\sigma)^2}{\partial i \cdot \partial j} * image(x,y) \tag{2}$$

The local maxima of this filter response occur in regions where both G_{xx} & G_{yy} are strongly positive, and where G_{xy} is strongly negative. Therefore, these extrema occur in regions in the image with large intensity gradient variations in multiple directions, as well as at saddle points. Interest points can be found at different scales, partly because the search for correspondences often requires comparison images where they are seen at different scales. In other feature detection algorithms, the scale space is usually realized as an image pyramid. Images are repeatedly smoothed with a Gaussian filter, and then they are sub-sampled to get the next higher level of the pyramid. Therefore, several floors or stairs with various measures of the masks are calculated:

$$\sigma_{approx} = CurrentFilterSize * \frac{BaseFilterScale}{BaseFilterSize} \tag{3}$$

The scale space is divided into a number of octaves, where an octave refers to a series of response maps of covering a doubling of scale. In SURF, the lowest level of the scale space is obtained from the output of the 9x9 filters. The output of the above 99 filter considered as the initial scale layer at scale s=1.2.The following layers are obtained by filtering the image with gradually bigger masks, taking into account the discrete nature of integral images and the specific filter structure. This results in filters of size 99, 15x15, 21x21, 27x27.Non-maximum suppression in a 3x3x3 neighborhood is applied to localize interest points in the image and over scales. The maxima of the determinant of the Hessian matrix are then interpolated in scale and image space. SURF detector interpolates the coordinates of any local maxima found into the subpixel and subscale range. Finally, SURFs propose to make the distinction between bright blobs found on a dark can be represented by the sign of the Laplacian, as shown below:

$$\operatorname{sgn}\{G_{xx}(x,y,\sigma) + G_{yy}(x,y,\sigma)\} = \begin{cases} +1 \Rightarrow \text{bright blob over dark background} \\ -1 \Rightarrow \text{dark blob over bright background} \end{cases} \tag{4}$$

2.3.2 Surf Descriptor

To describe each feature, SURF summarizes the pixel information within a local neighborhood. The first step is determining an orientation for each feature, by convolving pixels in its neighborhood with the horizontal and the vertical Haar wavelet filters. These filters can be thought of as block based methods to compute directional derivatives of the images intensity. By using intensity changes to characterize orientation, this descriptor is able to describe features in the same manner regardless

of the specific orientation of objects or of the camera. This rotational invariance property allows SURF features to accurately identify objects within images taken from different perspectives.

2.3.3 Surf Comparator

The SURF comparator is nothing but we are finding the matching. Here the matching is done with the help of testing and training images. By comparing the descriptors obtained from different images, matching pairs can be found.

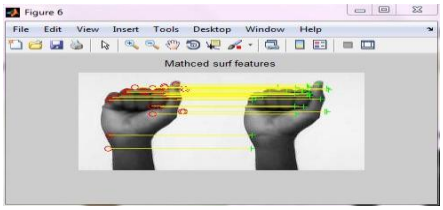

FIGURE 6: THE MATCHED SURF FEATURE

2.4 Text To Speech

TTS stands for Text-to-Speech a form of speech synthesis that converts text into voice output. There are numerous ways you can create audio from text.

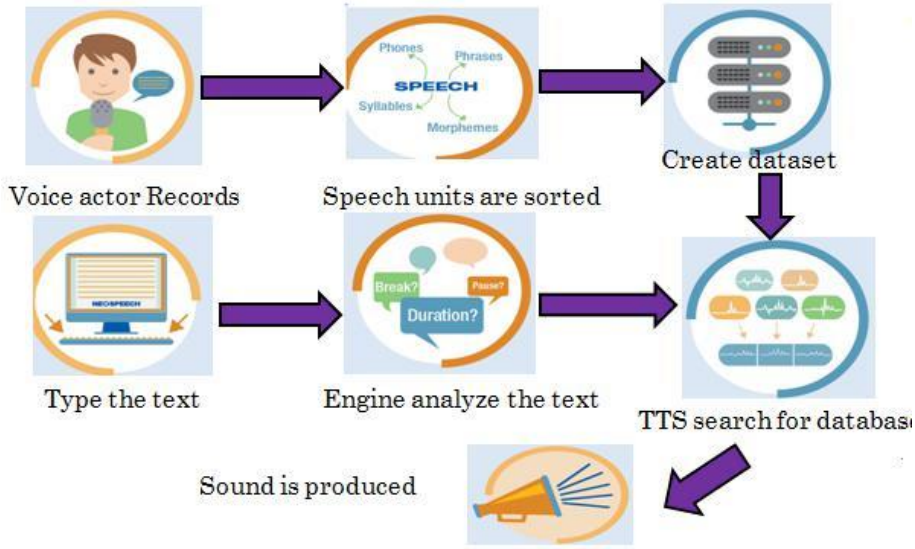

FIGURE 7: WAY TO CREATE AUDIO FROM TEXT

First, choose a voice actor with a great sounding voice who is frequent in any language. Record the voice actor speech units, from whole sentences to syllables. There by we get the natural sound of the actor. Now there are thousands of recorded sound files, we need to sort them and organize them. The speech units are labeled and segmented by phones, words, phrases, and sentences. These speech units are used to build a large voice database.

Type the text that we have to convert to audio. Here the text we consider for the signs are superb, nice, good, etc. The next process is language processing .We are considering these text for audio creation. From the language processing end, the text is normalized and broken down into phonetic sounds before going through a series of analyses to understand the structure of the sentences as well as to determine the context of the word for pronunciation. Through these processes, we are able to establish prosodoy rhythm, stress, and produce natural sounding speech. the Natural Language Processing (NLP) Part and the Voice Database come together to start producing speech.

III. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the proposed method, all the algorithms are implemented in the MatlabR2013a environment on a I3 -3.30GHz PC with 4GB RAM. Our algorithm is mainly based on the concept of SURF. Surf detects the blobs and finding the best matches .First of all we have to take the testing video and training video with the help of webcam.

FIGURE 8: THE TESTING VIDEO AND TRAINING VIDEO WITH THE HELP OF WEBCAM

The testing video is extracted as testing frame and the training video extracts the training frames. We have now 25 frames of training images and one testing image. The figures shows the trained images and test images.

FIGURE 9: THE EXTRACTED TEST IMAGE

FIGURE 10: THE MATCHED TESTING AND TRAINING SURF IMAGES

The next step is to checking the training images with our test image. If the image 1 (trained frame) having best match the output =a. If the image 2 has best match output =b. Here the testing image is 18 then we have to get the output as 'r'. Finally we get the output corresponding to our test image with some blobs. We compared all the train images with our test image. The output is also displayed as the corresponding alphabet.

The figure shows the corresponding letter and matched train image. We are giving each gestures with different names, such as superb, good, nice, etc. it will also displayed in the window corresponds to the letter. Finally, we are converting the text SUPERB to an audio by using a TTS function.

FIGURE 11: THE MATCHED LETTER FOR TEST IMAGE

Therefore using the surf algorithm we get better fast performance compare with other algorithms. Thus we can identify our test image of 18 by the letter 'r'. The text is also converted as audio by using text to speech conversion. Compare with other methods hand gesture recognition using surf feature has good performance. They take less time to recognize a gesture.

IV. PERFORMANCE EVALUATION

The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

A is the binary image of train image and B is binary image of test image. And jaccard index value should be in between 0 and 1 (0 J(A; B) 1).It calculates the similarity points in between trained and testing image. And this similarity is taken as jaccard

index. The numerator portion represents similar points between images and the denominator portion represents total points between trained and tested image. The ratio between these values represents jaccard index.

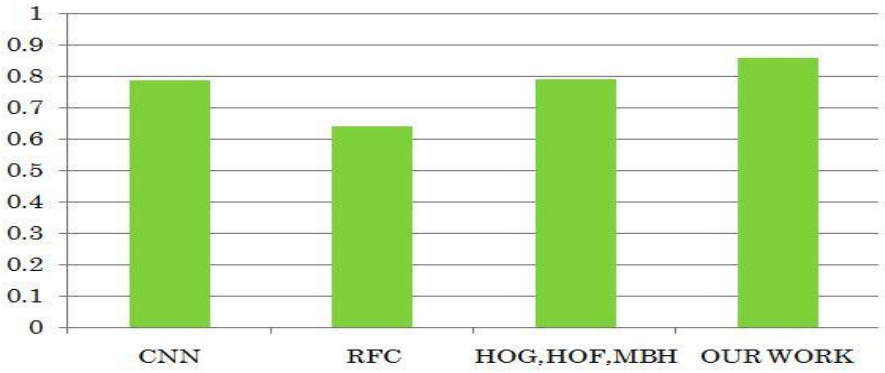

FIGURE 12: COMPARISON OF PREVIOUS WORKS OUR WORK

In the figure 5.1 our method shows jaccard index value greater than 0.8,which is larger than other algorithms.RFC method shows least jaccaed index value.HOG method have second largest jaccard index value. So from the graph we can clearly say that our method have better performance than previous methods. There is another performance evaluation which is done by accuracy index. In jaccard index section the convolutional neural network has the nearest performance with this work. So we consider the accuracy index with CNN. Let X_i is the accuracy index we are consider number of correctly classified samples and total number of samples.

TABLE 1
PERFORMANCE EVALUATION BASED ON ACCURACY INDEX

Methods	Total number of samples	Correctly classified samples	Accuracy (%)
CNN[9]	20	15	78.5
This Work	20	17	80

Thus the performance evaluation in accuracy index shows that this work consists of more accuracy than the CNN method. We considered total number of samples as 20 in both case. The number of correctly classified samples in CNN method is 15 and in this method we have 17 correctly classified samples. Therefore CNN has the accuracy of 78.5 % and we have 80 % of accuracy.

V. CONCLUSION

In this paper, we are finding the best match comparison with our train image and test image using SURF feature. The SURF feature detects the important points or interest point's .With the help of these interest points we can find the matching gestures and thereby we are converting a text to audio. Here we are only considering RGB images. An advantage of the SURF algorithm is that, it takes less time to find matching. By using jaccard index values, we can find that our performance is high compared to other algorithms. As a future work we can find the position of the gesture in the video without fixed position of gesture and also we can perform different mathematical algorithms by replacing surf feature.

REFERENCES

[1] L. Liu, L. Shao, F. Zheng, and X. Li, "Realistic action recognition via sparsely-constructed Gaussian processes," Pattern Recog., vol. 47, pp. 38193827, 2014, Doi: 10.1016/j.patcog. 2014.07.006.

[2] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacianpyra-mid coding for action recognition," IEEE Trans.Cybern., vol. 44, no. 6, pp. 817827, Jun. 2014.

[3] D. Wu and L. Shao, "Silhouette analysis-based action recognitionvia exploit-ing human poses," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 2, 236243, Feb. 2013.

[4] I. Laptev, "On space-time interest points,"Int. J. Comput. Vis.,vol. 64, pp. 107123, 2005.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features,"in Proc. 2nd Joint IEEE Int. Workshop Vis. Surveillance Perform. Eval. Tracking Surveillance, 2005, pp. 6572.

[6] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector,"in Proc. 10th Eur. Conf. Comput. Vis., 2008, pp. 650663.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in Proc. 15th Int. Conf. Multimedia, 2007, 357360

[8] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in Proc. Brit. Mach. Vis. Conf., 2008, pp. 2751.

[9] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks,"in Proc. Eur. Conf. Com-put.Vis. Pattern Recog.Workshops, 2014, pp. 16.

[10] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al., "Evaluation of local spatio-temporal features for action recognition," in Proc. Brit. Mach. Vis. Conf., 2009, pp. 111.

[11] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in Proc. 11th Eur. Conf. Comput. Vis., 2010, PP. 140153

[12] D. Wu and L. Shao, "Deep dynamic neural networks for gesture segmenta-tion and recognition,"in Proc. Eur. Conf. Comput. Vis. Pattern Recog. Work-shops, 2014, pp. 552571.

[13] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, PP: 33613368

[14] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification," in Proc. Brit. Mach. Vis. Conf., 2012, pp. 112.

[15] G. E.Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets,"NeuralComput., vol. 18, pp. 15271554, 2006.

[16] J. Schmidhuber, "Deep learning in neural networks: An overview," arXiv preprint arXiv:1404.7828, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Neural Inf. Process. Syst., 2012, pp. 11061114.