ISSN 2277-3061
Volume 15 Number 7
International Journal of Computer and Technology

COLOURED IMAGE SEGMENTATION USING K-MEANS ALGORITHM

Ishita Vishnoi[1], Nikunj Khetan[2], Dr. S. Indu[3]
[12]Department of Computer Engineering, Delhi Technological University, New Delhi, India
[3]Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi, Indiia
[1]ishita95.2010@gmail.com, [2]nikunjkhetan123@gmail.com, [3]s.indu@dce.ac.in

ABSTRACT

Hand gestures are natural means of communication for human beings and even more so for hearing and speech impaired people who communicate through sign language. Unfortunately, most people are not familiar with sign language and an interpreter is required to translate dialogues. Hence, there is a need to develop a low cost, easily implementable and efficient means to recognize sign language gestures to eliminate the interpreter and facilitate easier communication. The proposed work achieves a satisfactory recognition accuracy using in-built laptop webcam using combination of 3 skin color models(HSV,RGB,YCbCr) and background subtraction to eliminate noise from webcam low quality images to recognize sign language for helping the hearing and speech impaired in real-time without requiring too much computational power or any other device as it can be implemented in any laptop with a webcam.

Keywords

Background subtraction, Hand gesture recognition, HSV, RGB, YCbCr, Multi-class SVM

Academic Discipline And Sub-Disciplines

Computer Science, Computer Engineering, Human Computer Interaction

SUBJECT CLASSIFICATION

Image Processing, Machine Learning, Computer Vision

TYPE (METHOD/APPROACH)

Experimentation, Empirical studies, Computer Simulation

1. Introduction

Hand Gesture Recognition system is an active research area in computer vision due to its broad application in human computer interaction(HCI),sign language, augmented reality,gaming,etc. Hand gesture is a natural means of communication between humans and vision-based systems to provide a user interface that can help people convey their feelings and information satisfactorily.

In genereal,there are two types of gestures-static and dynamic. Static gestures are expressed in single frames whereas dynamic gestures span over multiple frames incorporating temporal variability. To recognize these gestures there are three basic processing stages- hand segmentation, feature extraction and classification. Hand segmentation contributes greatly to the accuracy of gesture recognition which poses many difficulties because hand segmentation is easily affected by illumination changes, skin color differences between humans, presence of other background elements having similar skin color and low camera quality. To solve this issue of effective hand segmentation many methods have been proposed. In [1-2] gloves and colored markers are used to segment the hand and also obtain additional information such as orientation and position of palm and fingers.

The color space that is used plays a significant role in the successful segmentation but, color spaces are sensitive to lighting changes, hence, there is a tendency to use chrominance components while neglecting the luminance components. We use a combination of three color models-HSV,BGR,YCbCr[11], to segment skin pixels and background subtraction[12] to negate effects of illumination changes and complex background.

Feature extraction is done using SIFT [9] and used Bag of Features model to map the space incompatible SIFT keypoints using K-means clustering as in [10].

We have used a multiclass SVM classifier to minimize computational complexity, dataset and show real-time behavior without delayed response.

The paper is organized as: Section A describes the image preprocessing, Section B illustrates the training stage and Section C illustrates the testing stage.

2. State Of The Art

In [3] infrared camera and in [4] Microsoft Kinect depth sensor was used. Both these methods give good accuracy but are expensive methods because both the infrared camera and Kinect sensor are costly to acquire. [5] uses two fixed cameras to real-time calculate the hand position and recognize gesture information, this method doesn't achieve good accuracy results and also increases complexity due to usage of two cameras. [6] uses 3D information provided by a range camera. This method can be used only for very simple gestures and has low segmentation and tracking speed.

For classification many methods have been used,[7] uses the Dynamic Time Approach(DTW) approach which requires large number of templates for a range of variations and cannot handle undefined patterns, [8] uses Hidden Markov Model(HMM)to recognize dynamic gestures with a high accuracy but requires high computational power and has a delayed response to detection of gestures. Neural Network classifier has been applied for gestures classification [14][15] but it is time consuming and when the dataset increases, the time needed for classification is increased too. Another type of feed-forward NNs were used called the Convolutional Neural Networks (CNN, or ConvNet) [16] to recognize hand gestures. The method also used Kinect Sensor along with GPU acceleration to achieve high accuracy but, requires high computational power and is costly. [17] also uses CNNs to achieve high classification rate under varying conditions of intensity and illumination changes but, also involves high computational power and is expensive.

3. Problem Formulation

This paper proposes a method to recognize static hand gestures in complex background images that represent alphabets and numbers of the standard Polish Sign Language. The Polish Sign Language has been chosen as it uses only single handed gestures as compared to the Indian Sign Language that uses both hands for sign language gestures. The individual gestures once recognized accurately, can be further used to identify meaningful words and sentences.

4. Materials And Methods

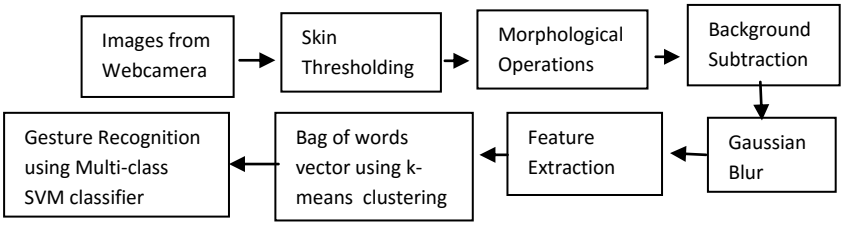


Figure 1. Method used in this paper for hand gesture recognition

Image Capture:

The following two devices were used to capture images for this work

-Canon EOS Rebel T3 DSLR Camera

-Usb 2.0 UVC HD Webcam

The images are taken under various illumination and intensity conditions. One input image(Figure 2(a)) is shown.

4.1 Image Preprocessing

Preprocessing phase of the proposed system consists of different operations: Skin Thresholding, Morphological Operations, Background Subtraction, Gaussian Blur.

I. **Skin Thresholding**: The objective of skin thresholding is to segment skin color pixels in the webcam image. Three color models are used RGB,HSV,YCbCr and the bounding rules of the components of these color models are derived from [11].See Figure 2(b).

II. **Morphological Operations**: Morphological filtering operations-dilation and erosion are used to reduce object noise from the binary image and as a result get a smooth, complete and closed segmented hand gesture after skin thresholding. See Figure 2(c)

III. **Background Subtraction**: It is a Gaussian Mixture-based Background/Foreground Segmentation Algorithm[12] that selects the appropriate number of gaussian distribution for each pixel. It provides better adaptibility to varying scenes due illumination changes etc. It eliminates background objects having similar skin color. See Figure 2(d)

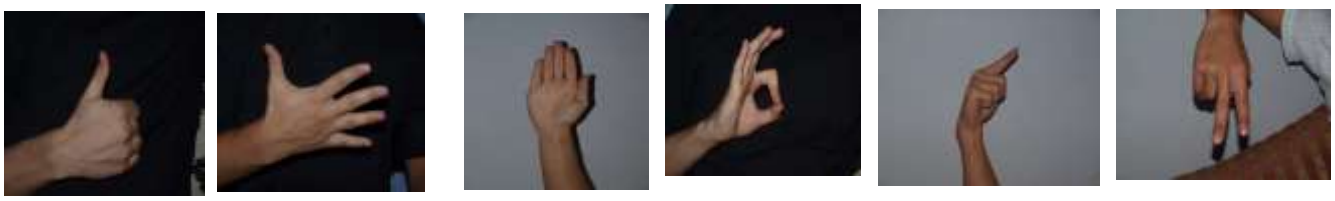
IV. **Gaussian Blur**: It is a very useful filter to eliminate noise from the image. Gaussian filter convolves each point in input array with Gaussian kernel and thereafter sums them all to give the output array. See Figure 2(e)

2(a)Input Image 2(b)Skin Thresholding 2(c)Dilation-Erosion 2(d)Background Subtraction 2(e)Gaussian Blur

Figure 2(a)-2(e)

4.2 Training Stage

For building the training set, a total of 120 images of different hand gestures of different people under various illumination conditions are taken to provide a more realistic and accurate training to the system. For training the system, we have taken 6 different gestures and assigned the labels as shown in Fig 3(a)-3(f).



3(a)One 3(b)Five 3(c)B 3(d)O 3(e)U 3(f)D

Figure 3(a)-3(f)

The training dataset consists both of images taken against uniform background and images taken by a webcam agianst a more noisy background. The training images are used to build the bag of words vocabulary and to train the SVM classifier. Table 1 below shows the class labels assigned to gestures.

Table 1. Some Sign Language Gestures and their class labels

(Polish Sign Language)

Gesture (Sign Language)	Corresponding Class Label
One	Class 1
Five	Class 2
B	Class 3
O	Class 4
U	Class 5
D	Class 6

I. Extracting SIFT Keypoints

Keypoints are salient image patches that include rich local information of an image. Keypoints are detected by a robust feature detection method,namely SIFT, in this paper. SIFT, gives is found to give accurate results for the gesture recognition task at hand as it is invariant to changes in scale, orientation and illumination.

The steps that are involved to generate the set of image features are as follows:

1. **Scale-space extrema detection**: This first stage consists of carrying out computation that searches all scales and image locations. It is implemented efficiently by utilizing a difference-of-Gaussian function to identify potential interest points that are invariant to orientation and scale.

2. **Keypoint localization**: At each candidate location, a detailed model is apt to resolve scale and location. Keypoints are selected based on measures of their stability.

3. **Orientation assignment**: One or more orientations are appointed to each keypoint location which is based on local image gradient directions. All future operations that are performed on image data has been transformed relative to the appointed orientation, location, and scale for each feature, providing invariance to these transformations.

4**. Keypoint descriptor**: The local image gradients are measured by selecting scale in the region around each keypoint. These are then transformed into a representation which then allows for significant levels of local shape distortion and change in illumination.

A training images with its keypoints is shown in figure 4.

Figure 4. Keypoints in an image

II. K-Means Clustering And Bag Of Words Vocabulary

A clustering procedure is applied to group key points from all training images into a large number of clusters, with the center of each cluster corresponding to a different visual word. The number of clusters is to be pre-defined as the vocabulary size and depends on the structure of data used. Here it is estimated by the maximum number of keypoints extracted for an image that may also include other objects in the background.In this work, the maximum number of keypoints are for the palm gesture ,around 85,but for a webcam image with a cluttered background, this number increases due to keypoints from other objects too.We have taken the size of clusters to build the vocabulary as 700.

In k-means clustering, the vector space is divided into k randomly located centroids and the keypoints extracted from each of the images are assigned to the nearest cluster.The centroids are then shifted to the mean position of its keypoints and this process is continued till the assignments become constant. Each of the keypoints(feature vectors) are then assigned to the nearest cluster center based on Euclidian distance.Therefore, the keypoints extracted from images are used to build the clusters which are then used to build the vectors of the vocabulary used for training.The cluster model builds k vectors,equal to the number of centroids, each of which has 128 components ie. the length of each keypoint.These keypoint vectors of each of the training images are then fed into the k-means clustering model where every image with n keypoints (n x 128) is reduced to 1 x k (bag of words) where k is the number of clusters.

Each keypoint (vector) that is extracted from a training image is represented by one component in the generated vector (bag-of-words) and is assigned the index value of the cluster centroid with the nearest Euclidean distance. The generated vector is grouped along with all the generated vectors of other training images that have the same hand gesture and that are also labeled with the same class number.

Visual representation of clusters and centroids are shown below in Figure 5.

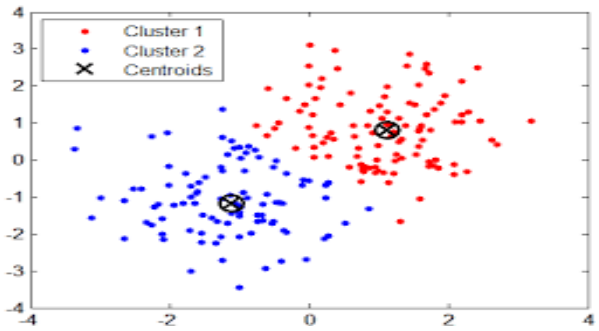


Figure 5. K-means Clustering

(http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/cluster/cluster.html)

III. Multi-Class SVM Training Classifier

The bag of words vector constructed for every training image, along with its class label is used as input for training the multiclass SVM training classifier model. A Support Vector Machine (SVM) is a discriminative classifier that is defined by a separating hyperplane i.e., given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. The optimal hyperplane is computed as shown below:

Let's introduce the notation used to define formally a hyperplane:

$$F(x)=\beta_0 + \beta^T x,$$

where β is known as the *weight vector* and β0 as the *bias.*

The optimal hyperplane can be represented in an infinite number of different ways by scaling of β and β0. Conventionally among all the possible representations of the hyperplane, the one chosen is

$$\beta_0 + \beta^T x=1$$

where x symbolizes the training examples closest to the hyperplane. Training examples that are closest to the hyperplane are called **support vectors**. This representation is known as the **canonical** hyperplane.

Now, we use the result of geometry that gives the distance between a point $\mathbf{x}$ and a hyperplane (β, β0)

$$\textbf{Distance}=|\beta_0+\beta^T x|/||\beta||$$

In particular, for the canonical hyperplane, the numerator is equal to one and the distance to the support vectors is

$$\text{Distance}_{\text{support vectors}}=|\beta_0+\beta^T x|/||\beta||=1/||\beta||$$

Recall that the margin introduced in the previous section, here denoted as M, is twice the distance to the closest examples:

$$M=2/||\beta||$$

Finally, the problem of maximizing M is equivalent to the problem of minimizing a function **L(β)** subject to some constraints. The constraints model the requirement for the hyperplane to classify correctly all the training examples $\mathbf{x_i}$. Formally,

Min $\mathbf{L(\beta)=0.5||\beta||^2}$ **subject to** $y_i(\beta_0+\beta^T x_i)>=1$ ∀i,

β, β_0

where $\mathbf{y_i}$ represents each of the labels of the training examples.

This is a problem of Lagrangian optimization that can be solved using Lagrange multipliers to obtain the weight vector **β** and the bias β_0 of the optimal hyperplane.

The SVM once trained, can predict with some accuracy the label of the class for a test image.

4.3 Testing Stage

In this part of the system, the six hand gestures for which the multi class SVM classifier was trained, are tested. The images are taken from a webcam (50 images) and from the DSLR camera (50 images) to simulate realistic behaviour, are taken under various light and background conditions. The keypoints are extracted from the testing images and after applying the k-means clustering model on them, are mapped to the bag of words vocabulary created in the testing stage. Each feature vector in the keypoints is represented by one component in the generated vector (bag of words) with value equal to the index of centroid in the cluster model with the nearest Euclidean distance. Then, the generated vector (Bag-of-Words) is used as input to the multi-class SVM training classifier model that was built in the training stage which predicts the most accurate matching class label for each testing image. Thus, the pre-defined hand gestures can be recognized from the class labels. The average time taken for testing and predicting the class of 50 images is 10 seconds.

5. Results

The training dataset used to train the classifier consists of a total of 100 images of which 50 images are taken by the DSLR camera with a good resolution and uniform background, and other 50 images are taken using a webcam. The training images taken under different scale, orientation and illumination conditions for each gesture make the system robust to changes in these factors. Moreover, the skin thresholding and pre-processing eliminates the extra objects and noise in the background.

The testing dataset is made up of 100 images of each hand gesture, 50 taken from the camera and 50 from webcam. The accuracy of the present system is an average 90% for images taken from the DSLR camera in uniform background and 80% for images taken from the web camera as shown in Table 2.

The accuracy level is high when images are taken from a good resolution camera as compared to webcam images where it suffers due to low quality of images with a lot of background noise. Moreover, it was observed during the experiment that in the present system, the accuracy levels drop when the hand gestures are somewhat similar in the number of fingers or shape. Keeping in mind these constraints, the system can still be trained for more gestures, provided there is sufficient training data set for each of those hand gestures. The model described in this paper, however, responds fairly well to the specific distinct hand gestures in different conditions and is time efficient.

Table 2. Testing Results of Images

Gesture (Sign Language)	Corresponding Class Label	ACCURACY(%)	
		Canon EOS RebelT3 Images	Webcam Images
One	Class 1	90	81
Five	Class 2	92	85
B	Class 3	93	79
O	Class 4	95	83
U	Class 5	85	75
D	Class 6	87	76

6. Conclusion

Our proposed work has a good accuracy of recognition of sign language gestures of the Polish Sign Language, 90% for uniform background images and 80% in the case of webcam images. The satisfactory recognition accuracy using in-built laptop webcam makes it easily implementable and useful to recognize hand gestures hence, sign language recognition for helping the hearing and speech impaired can be done in real-time without requiring too much computational power or any other device as it can be implemented in any laptop with a webcam. This work can further be extended to form and identify meaningful words and statements from the individual alphabets and numbers of the sign language, to be displayed as subtitled text for the sign language in real time to aid the physically impaired.

7. Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

REFERENCES

[1] D. J. Sturman and D. Zeltzer, “A survey of glove-based input,” IEEE Computer Graphics and Applications, vol. 14, no. 1, pp. 30–39, Jan. 1994

[2] N. Ibraheem, M. Hasan, R. Khan, P. Mishra, (2012). “comparative study of skin color based segmentation techniques”, Aligarh Muslim University, A.M.U., Aligarh, India.

[3] Xingyan Li. (2003). “Gesture Recognition Based on Fuzzy C-Means Clustering Algorithm”, Department of Computer Science. The University of Tennessee Knoxville.

[4] Yingying She, Qian Wang, Yunzhe Jia, Ting Gu, Qun He, Baorong Yang[2014], “A Real-time Hand Gesture Recognition Approach Based on Motion Features of Feature Points”,Software School ,Xiamen University Xiamen, China..

[5] Tong-de Tan, Zhi-min Guo ,“Research of Hand Positioning and Gesture Recognition Based on Binocular Vision “,Information and Engineering School, Zhengzhou University ,Zhengzhou, China

[6] Hervé Lahamy ,Derek Litchi ,” REAL-TIME HAND GESTURE RECOGNITION USING RANGE CAMERAS “, Department of Geomatics Engineering, University of Calgary,Calgary,Alberta.

[7] Hong Cheng, Jun Luo,Xuewen Chen,” A WINDOWED DYNAMIC TIME WARPING APPROACH FOR 3DCONTINUOUS HAND GESTURERE COGNITION”.

[8] Zhong Yang,Yi Li,Weidong Chen,Yang Zheng(2012),” Dynamic Hand Gesture Recognition Using Hidden Markov Models”.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91-110, 2004.

[10] Nasser Dardas, Qing Chen, Nicolas D. Georganas,Emil M. Petriu, Fellow(2010),"Hand Gesture Recognition Using Bag-of-Features and Multi-Class Support Vector Machine".

[11] Nusirwan Anwar bin Abdul Rahman, Kit Chong Wei and John See," RGB-H-CbCr Skin Colour Model for Human Face Detection".

[12] Z.Zivkovic, "Improved adaptive Gausian mixture model for background subtraction" in 2004 and "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction" in 2006.

[13] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines,2001.

[14] Simei G. Wysoski, Marcus V. Lamar, Susumu Kuroyanagi, Akira Iwata, (2002). "A Rotation Invariant Approach On Static-Gesture Recognition Using Boundary Histograms And Neural Networks," IEEE Proceedings of the 9th International Conference on Neural Information Processing, Singapura.

[15] Kouichi M., Hitomi T. (1999) "Gesture Recognition using Recurrent Neural Networks" ACM conference on Human factors in computing systems: Reaching through technology (CHI '91), pp. 237-242. doi: 10.1145/108844.108900

[16] L.Pigou,S.Dieleman,P.-J.Kindermans and B.Schrauwen, Sign language recognition using convolutional neural networks. In ECCVW, 2014.

[17]Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz Hand Gesture Recognition with 3D Convolutional Neural Networks. In CVPRW, 2015.

Author' biography with Photo

Ishita Vishnoi,

Undergraduate,

Department of Computer Science,

Delhi Technological University

Nikunj Khetan,

Undergraduate,

Department of Computer Science,

Delhi Technological University

Dr. S. Indu,

Associate Professor,

Department of Electronics & Communication,

Delhi Technological University