2012 International Conference on Information and Network Technology (ICINT 2012)
IPCSIT vol. 37 (2012) © (2012) IACSIT Press, Singapore

A New Framework for Real-time Hand Gesture Detection and Recognition

J. Geethapriya [1 +] and A. Srinivasan [2]

[1] Department of Computer Science and Engineering, MNM Jain Engineering College, Thoraipakkam, Chennai.

[2] Head of the Department of Information Technology, MNM Jain Engineering College, Thoraipakkam, Chennai.

Abstract.A real-time system to interact withan application or video game via hand gestures is being developed. The proposed work results in a more efficient hand gesture detection and recognition system under real-world scenarios. The system has two stages, offline training stage and online testing stage. In the training stage, keypoints are extracted using rotation and scale invariant matching (RASIM) technique. After clustering the keypoints using K-Harmonic Means technique, the keypoints are mapped into a unified histogram vector(Bag-of-words vector), which is given as input to the multiconlitron classifier to build the training classifier. In the testing stage, the images are captured from the webcam or video file. The face is detected and subtracted before detecting the hand gesture using skin detection and contour comparison. The keypoints extracted using RASIM technique from every image is clustered to map them into a bag-of-words vector. This could be finally given as input to the multiconlitron training classifier to recognize the hand gestures.

Keywords: bag-of-words, hand gesture detection, rasim, k-harmonic means, multiconlitron,human computer interaction.

1. Introduction

Hand gesture is an effective way of communication between the human and computers. Efficient vision based hand tracking is a challenging task because of the high Degrees of Freedom (DoF) involved by human hand. The hand gesture recognition systems have to meet real-time performance, recognition accuracy and should be robust against transformations and cluttered background.

Color of the skin can be used as a significant image feature to detect and track human hands. But using skin color based methods for hand gesture detection have some challenges such as removing face and other skin-like objects from the image.To overcome this problem, a system is proposed that eliminates face region by detecting and replacing the face region with a black circle. Then the skin region is detected and contours of the skin area is compared with the template representing the hand gesture and the hand gesture is saved as a small image before applying feature extraction technique. The Bag-of-features approach [5]is used to reduce the dimensionality of the feature space.

2. Related Work

Vision-based hand gesture recognition systems are categorized into two approaches, Appearance-based approach and 3D hand model based approach [4]. Appearance-based approach uses image features to model the visual appearance of the hand and compare these parameters with the image features extracted from

[+] Corresponding author. Tel.: +9791070052.
E-mail address: jgpriya21@gmail.com.

theinput video frames. This approach has real-time performance, becausethe 2D image features can be easily used and hence this approach is used in this framework.

3D hand model-based approach [1] relies on a 3D kinematic hand model with considerable degrees of freedom and tries to estimate the hand parameters by comparison between the input frames and possible 2D appearance, projected by the 3D hand model. This approach has some constraints, as it requires a huge image database to deal with the entire characteristic shapes under several views, matching the query image frames of video input with the test databases are time-consuming and computationally expensive. It is difficult to extract features and unable to handle the similarities occurring from unclear views.To achieve real-time performance, some hand gesture recognition systems used data gloves and colored markers to make the gesture recognition task easier [6]. However, this affects user's convenience. Our approach focuses on developing a bare hand gesture recognition system without any markers or data gloves. In [3], Sebastien Marcel implemented the neural network approach is to recognize the hand postures. To segment the hand postures, a space discretization based on face location and body anthropometry was used.

This paper is organised as follows. Section 3 describes the system design of the real-time hand gesture detection and recognition system. Section 4 deals with the training stage of the system. Section 5 deals with the processes in the testing stage.

3. System design

The main contribution of this paper is to develop a more efficient bare hand gesture recognition system that provides very high gesture recognition accuracy in real-time situations. By using more efficient techniques in all the stages, the real-time requirements of the system can be met effectively. The complete detail of the system design is shown in Fig 1. The system includes two stages, offline training stage and online testing stage. In the training stage (pre-processing stage), the cluster model and the multiconlitron training model is built. In the testing stage of the system, the face is subtracted and the hand region is detected. After extracting the features these keypoints are given as input to the cluster model and the bag-of-words vector is generated. This is finally given to the multiconlitron training classifier to recognize the hand gestures.

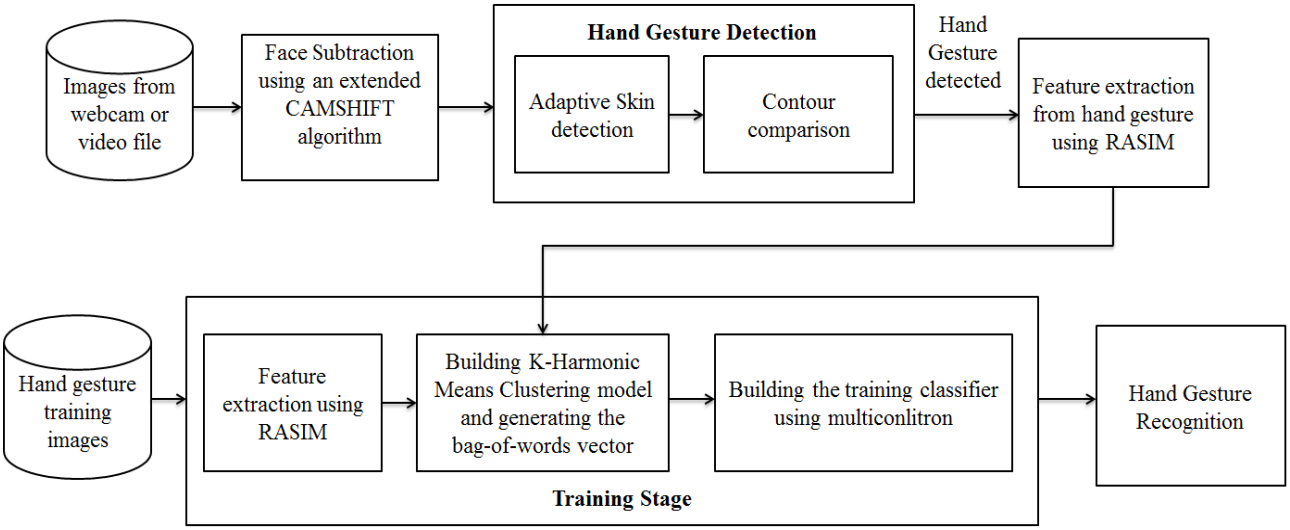

Fig. 1: System Design

4. Training Stage

Training stage is the pre-processing stage in which training images of hand gestures such as fist, index, palm and little finger gestures are captured from different people in varying scale, orientation and illumination conditions. The training images contain only the hand gestures without any other objects and the background is simply a white wall. The image processing time can be reduced by decreasing the number of keypoints extracted by RASIM technique. This can be done by reducing the image resolution and converting the training images into Portable Gray Map (PGM) format. The image processing time is not much important

in the training stage as in the testing stage, since training is the offline process. The bag-of-words model and the training classifier are built by the following process.

4.1. Feature extraction using robust and scale invariant matching (RASIM)

RASIM [8] provides resistance to partial occlusion and relatively insensitive to the changes in viewpoint. RASIM works with two stages. In the offline stage, Scale Invariant Feature Transform (SIFT) [2] is applied to the training image for finding reference object template keypoints (OKPs) and each keypoint is assigned with proper orientation and location. The number of keypoints is denoted asn_{ko}. Then image patch around each OKP is taken and the gradient magnitude and gradient orientation is computed and pseudo log-polar sampling is applied to gradient sub images and the set of log-polar pixels belonging to the same ring is taken as a separate reference object data vector (ODV). For each n_{ko}OKP, n_r ODVs from gradient magnitude sub image and n_r ODVs from gradient orientation sub image are acquired. The prediction filters $P_{i,r}$ for each keypoint and ring are saved to be used for matching process in the online step.

In the online step of RASIM, the data vector TDV and n_r values are acquired for test image.The adaptive transform will be constructed using the prediction filter $P_{i,r}$ saved in the offline step. Both of the non-adaptive wavelet transform and the adaptive wavelet transform are translation invariant so these wavelet transforms are applied to $TDV_{j,r}$ and wavelet coefficients are compared for each of the reference image keypoint and the test image keypoint. The similarity value $\rho_{i,j,r}$ between ith reference object keypoint, OKP_i and jth test image keypoint, TKP_j can be represented as

$$\sum_{r=1}^{2*n_r} w_{r,}\rho_{i,j,r}$$

Where

$$w_r = e^{-((r-1)mod\, nr^2)}/nr^2$$

The operation ‘mod’ denotes modulus after division.

4.2. K-Harmonic means clustering

The keypoints are clustered using K-Harmonic means (KHM) clustering [9], which uses the harmonic averages of distances from each data point to the prototypes. This algorithm converges to a better solution compared to both K-means and EM clustering. The number of clusters and the values of the starting clusters are determined. The membership function $m(C_j|x_i)$ of the keypoint x_i to the cluster center C_j has the following properties.

$$m(C_j|x_i) \geq 0,$$

$$\sum_{j=1}^{k} m(C_j|x_i) = 1$$

K-Harmonic Means applies soft-membership between the cluster and the data point.

$$0 \leq m(C_j|x_i) \leq 1$$

The keypoints of each of the training image is given to the KHM clustering model to build the bag-of-words vector with components equal to the number of clusters [12]. The bag-of-words approach is used because the keypoints extracted are different and lack meaningful ordering. This creates problem during classification. So, it is necessary to map the keypoints into fixed dimensional bag-of-words vector before providing it as input to the multiconlitron classifier. The bag-of-words vectors are grouped and the same hand gestures are given the same class number.

4.3. Building the training classifier using multiconlitron

The bag-of-words vector along with their class number is given as input to the multiconlitron classifier [10] to build the training model. The multiconlitron is the union of multiple conlitrons and comprises a set of hyperplane or linear functions surrounding a convex region to separate two convexly separable datasets. A new iterative algorithm called Cross Distance Minimization Algorithm (CDMA) is used to compute the hard margin non-kernel support vector machines by nearest point pair. Using CDMA, two algorithms are derived, the Support Conlitron Algorithm (SCA) to construct the support conlitrons and the Support Multiconlitron Algorithm (SMA) to construct the support multiconlitrons. Two complicated non-intersecting classes are separable by using a set of linear functions without using kernels. The multiconlitron may consist of many redundant conlitrons, which can be deleted to make the multiconlitron smaller. The classifier model is thus built using multiconlitron classifier.

5. Testing stage

In the testing stage, for each frame captured from the webcam, applying skin detection for the whole image makes the contours of fist hand gesture to be recognized as human face, as both of them has similar contours. So, it is necessary to subtract the face before applying skin detection technique to detect hand gesture. The detected hand gesture is saved as a small image. The keypoints are extracted from the small image and given to the cluster model to map into bag-of-words vector, which is given as input to the multiconlitron training classifier model to recognize the hand gestures.

5.1. Face detection and subtraction

Face is detected and segmented using a combination of haar-like features and skin color model based on the adaboost algorithm [11]. This face detection system includes three steps, classifying skin and non-skin pixels, identifying different skin regions using morphological operations and blob analysis and using adaboost to decide whether skin regions identified is a face or not.

To achieve real-time performance, face tracking is performed using extended Continuous Adaptive Mean SHIFT (CAMSHIFT) [11]. This algorithm has advantages such as providing robust real-time tracking of human faces, occlusion handling and the use of multi-dimensional histogram with hue and saturation component in HSV model to deal with objects that are similar to the background color. An arbitrary number of histograms are modelled for different appearances of target objects.

5.2. Hand gesture detection

Adaptive skin color model implemented in YCbCr color space is used to detect skin region in the image or frame to detect the hand region [7]. This skin color model works in four steps. Capturing the image, CbCr mapping, Shape enhancement which includes erosion and dilation, Borderline formation that includes edge detection. The pixel values of the hand is captured and converted into YCbCr color space. Then the CbCr color space is mapped to CbCr color plane and a clustered region of skin color is built. Then edge detection is applied to the cluster to create adaptive skin color boundaries for classification and the hand region is segmented.

For each frame captured, the skin like objects and other objects in the background are eliminated by comparing the contour of the detected skin area with any of the hand gesture contour [12]. If the skin contour complies with the contour of the hand gesture, the hand region alone is segmented and saved as a small image (50 x 50 pixels).

5.3. Hand gesture recognition

The small image that contains 50 x 50 pixels is converted into PGM format. This reduces the image processing time required to extract the keypoints. For every PGM image, the keypoint vectors are detected by applying RASIM offline and online stages. These data vectors are fed into K-harmonic means clustering model built in the training stage and the keypoints are mapped into bag-of-words vector with component equal to the number of clusters formed by the clustering model in the training stage. Finally the bag-of-words generated are given to multiconlitron training model built in the training stage to classify and recognize the hand gestures.

6. Results and discussion

The bag-of-words (unified histogram) vector is generated for the training images. The faces are detected and subtracted from the images captured from the webcam. The hand gestures will be detectedand recognized.The results will be shown with public image dataset containing hand gestures and the recognition accuracy will be highlighted.

7. Conclusion

A new framework for hand gesture detection and recognition system, which fulfils the real-time requirements such as high recognition accuracy, robustness in the presence of cluttered background and changing illumination conditions will be developed to interact with the computers in real-time. The efficiency of the proposed work will be compared with the existing system and the performance improvement will be highlighted.

8. References

[1] J. M. Rehg and T. Kanade, "Visual tracking of high DOF articulated structures: An application to human hand tracking", in Proc. European Conference on Computer Vision., 1994, pp. 35–46.

[2] D. G. Lowe, "Object recognition from local scale-invariant features", in Proc. IEEE International Conference on Computer Vision, Kerkyra, Greece, September 20-25 1999, pp. 1150–1157.

[3] S.Marcel, "Hand posture recognition in a body-face centered space", in Proc. Conference on Human Factors Computing systems (CHI), 1999, pp. 302-303.

[4] H. Zhou, T. S. Huang, "Tracking articulated hand motion with Eigen dynamics analysis", Proc. of International Conference on Computer Vision, Vol. 2, pp. 1102-1109, 2003.

[5] Y.Jiang, C.Ngo and J.Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", in Proc. of ACM International conference on Image video retrieval, 2007, pp. 494-501.

[6] A. El-Sawah, N.Georganas and E.Petriu, "A prototype for 3-D hand tracking and gesture estimation", IEEE Transactions on Instrumentation and Measurements, vol.57, no.8, pp. 1627-1636, Aug. 2008.

[7] Ahmad YahyaDawod, Junaidi Abdullah, Md. Jahangir Alam (2010), "Adaptive Skin Color Model for Hand Segmentation",2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2010).

[8] Mahdi Amiri and Hamid R. Rabiee, Senior Member, IEEE (2011), "RASIM: A Novel Rotation and Scale Invariant Matching of Local Image Interest Points", submitted to IEEE Transactions on Image Processing.

[9] MariuszFra¸ckiewicz, HenrykPalus (2011), "KHM Clustering Technique As A Segmentation Method For Endoscopic Colour Images", International Journal. Appl. Math. Computer Science.

[10] Li Yujian, Liu Bo, Yang Xinwu, Fu Yaozong, and Li Houjun (2011), "Multiconlitron: A General Piecewise Linear Classifier", IEEE Transactions on Neural Networks, Vol. 22, No. 2.

[11] Lae-Kyoung Lee, Su-Yong An, and Se-Young Oh (2011), "Efficient Face Detection and Tracking with Extended CAMSHIFT and Haar-Like Features", Proceedings of the 2011 IEEE International Conference on Mechatronics and Automation.

[12] Nasser H. Dardas, Nicolas D. Georganas, Fellow, IEEE (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.