Using HMM Technique for Gesture/Action Recognition

Arpitha Y P [#1], Dr Usha Saktivel[*2]
[# 1]P.G.Student, Departmen Computer Science, RRCE
Vishveshwaraya Technological University, Karnataka, India
[#2]Professor and HOD, Department of Computer Science,
Rajarajeswari College of Engineering, Karnataka, India

***Abstract*— CSMMI: Class Specific Maximization of Mutual Information is the approach which is being discussed in the further proceeding. CSMMI provides different and separate wordbook for each and every category. The specified approach i.e., CSMMI has two main goals. They are: 1.Maximizing the mutual data between wordbook inside an intrinsic structure.2.Minimizing the mutual data between the wordbook of extrinsic structure. The main aim of the project is action and gesture recognition which follows the following four main steps: they are Feature extraction, learning initial wordbook, CSMMI and Classification. In CSMMI we are going to use HMM: Hidden Markov Model. HMM involves evaluation, estimation and decoding processes. In the proposal of CSMMI each and every class will have its own discriminative dictionary which leads better performance when compared to other shared dictionary methods.**

I. INTRODUCTION

CSMMI: Class Specific Maximization of Mutual Information is a new method which is being used to learn a compact and discriminative dictionary for each class is introduced here. CSMMI not only discovers the latent class-specific dictionary items that best discriminates different actions, but also captures unique dictionary items for a specific class. One of the common approaches for dictionary optimization is to use information theory, (e.g. maximization of entropy (ME), maximization of mutual information (MMI)) and it shows promising results for action and gesture recognition. Accordingly, MMI rule is adopted to optimize the class-specific dictionaries. However, the approach varies from the shared dictionary learning methods.

Sparse representation is usually done using a Gaussian Process (GP). This is the model which is used to optimize an objective function. This indeed will maximize the mutual information for appearance information and distribution of class.

CSMMI is majorly used to find out the class-specific dictionary items that best differentiates between different actions. This also identifies very unique items of dictionary for every specific class. There are many approaches for dictionary optimization and one of the common approaches or dictionary optimization is to use information theory. ME: Maximization of entropy, maximization of MMI: mutual information belongs to such information theory and by default it shows best results for recognition of action and gesture.

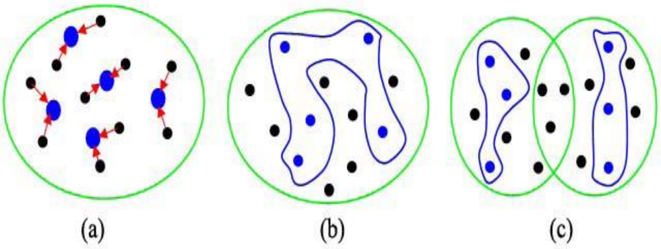

Figure 1: Dictionary item sets

Here, Figure1 (a) represents is LiuShah; Figure1 (b) represents QiuJiang and Figure1 (c) represents CSMMI . Each green circle in the figure represents the region of an initial dictionary. The black points represent the initial dictionary items and the blue points denote the selected dictionary items. In the methods of LiuShah and QiuJiang, the shared dictionary is the one which makes it difficult to distinguish about which dictionary item is important to a specific class. The one can only find the dictionary items that have the minimum loss of MI. In CSMMI, each class has one specific dictionary and some dictionary items shared between classes can be filtered out (Figure1 (c)).

CSMMI considers the global information and also unifies the inter-class and intra-class MI in a single objective function. Inter-class and intra-class information is more specific and useful than the class distribution used in QiuJiang since CSMMI captures discriminative dictionary items for a specific class. Our experimental results on public action and gesture recognition databases demonstrate that CSMMI compares favorably to the shared dictionary methods and other state-of-the-art approaches.

K-SVD (Scalar vector dictionary) proposed three dictionary learning frameworks: they are: 1. shared dictionary (classes contain only one dictionary), 2. Class-specific dictionary (one dictionary per class) and 3.concatenated dictionary (concatenation of the class-specific dictionaries). However, K-SVD only focuses on minimizing the reconstruction error and it is not clear about how to optimize the learned dictionaries. The learned dictionary obtained via K-SVD may not be compact and discriminative.

II. SYSTEM DESCRIPTION

Description regarding the existing system, disadvantages of existing system and proposed system are discussed below. Existing framework named CSSRC: Class Specific Sparse Representation Classification for action and gesture recognition. CSSRC includes four steps: 1. feature extraction and representation, 2.Learning initial class specific dictionaries, 3. CSMMI and 4. Classification. This work is inspired and only focuses on the shared dictionary. While this work explores the relationship between intra-class and inter-class MI for video-based recognition.

We use four types of features here. The first type is the space-time interest points (STIP) feature. We use STIP features to represent a video, and then histograms of oriented gradients (HOG) and histograms of optic flow (HOF) to describe each interest point. The second type is 3D enhanced motion scale invariant feature transform (EMoSIFT) feature which fuses the RGB data and depth information into the feature descriptors. The third type is Histograms of 3D Joints (HOJ3D) feature computed from skeleton information. The ast type is shape-motion feature, which is used to extract shape and motion features from video sequences. For different datasets, we may use different features based on the experimental result.

A. Disadvantages of Existing System:

CSSRC for action and gesture recognition. CSSRC includes four steps first being Feature extraction and representation, second is Learning initial class specific dictionaries, third is CSMMI and finally its classification. Work here is inspired and only focuses on the shared dictionary. While this work explores the relationship between intra-class and inter-class MI for video-based recognition. Drawbacks are we cannot expect accurate results from CSSRC and also is not compact and discriminative. To overcome the problem of existing system, we planned to propose a in HMM model.

B. Proposed Method:

Markov property is demonstrated by a time-domain process. First Markov process is defined as gif all present and past events are given then the conditional probability of a current event depends only on the most recent event. When considering the positions and orientations of the hands of a gesturer through time then this is a useful assumption to be made. The HMM is a governed by a). An underlying Markov chain with a finite number of states and, b).A set of random functions, each associated with one state.

An observation symbol is generated based on the random function from the current state. Each transition between the states has a pair of probabilities, they are Transition probability, and Output probability. Transition probability is the one which provides the probability for undergoing the transition and output probability is one which defines the conditional probability of emitting an output symbol from a finite alphabet when given a state.

The HMM model is a rich mathematical structure and efficiently deals with spatio-temporal information in a natural way. Only a sequence of observation can only be seen hence it is termed as hidden. It also involves efficient algorithms, such as Baum–Welch and Viterbi for evaluation, learning, and decoding. An HMM is expressed as a). Evaluation process for determining the probability that the observed sequence was generated by the model which uses Forward–Backward algorithm; b).Training or estimation for adjusting the model to maximize the probabilities which will be using Baum–Welch algorithm; c).Decoding to recover the state sequence using Viterbi Algorithm.

III. MAIN DESIGN

The basic architecture of Gesture/Action recognition is as below.

A. Architectural Representation

It first contains user who is responsible in providing the right input. Once the input is given next the process performs frame conversion which is very important for easy recognition of the required data.

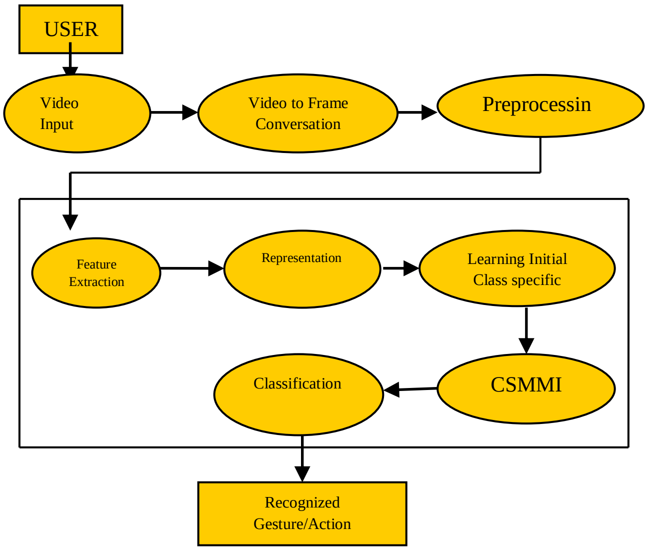
USER
Video Input
Video to Frame Conversation
Preprocessin
Feature Extraction
Representation
Learning Initial Class specific
CSMMI
Classification
Recognized Gesture/Action

Figure 2: Architecture diagram

Step by step procedure for the proposed model is:

- The input for the project is a video input which contains few set of different gestures and actions.
- Next the concentration will be for conversion of the input video to frames. Based on the frames obtained the detection of the dictionaries is done in third module.
- Obtained dictionaries are further processed for the detection of features and that will be done.
- Moving ahead it is the CSSRC for feature extraction, class specification, and classification process.

CSSRC is one of the process that involves the main CSMMI technology for the purpose of classification of objects. This is used to simplify the task of object classification which will be based on certain criteria. At the end what matters is the recognition of the right action and gesture and its proper presentation. For any given input histogram for every class, specific dictionary will be calculated. The obtained histograms are then tested to find out whether there are any reconstruction errors. The histogram which has a little reconstruction error will be compared with a reference histogram.

In image processing and computer vision the concept of detection of features usually deals with the methods whose main aim is to compute abstractions of the information of the image. Then local or instant decisions are made at every point of the image. The decisions are made about the features of the image of a given type is present at that point or not. The output obtained will be the subsets of the image domain, which will be often in the form of isolated points, continuous curves or connected regions. Learning initial class is to extract the samples from the class. On the basis of similarities observed further classifications will be made. Here we are going to extract the features based on the training given.

CSMMI will be capturing discriminative dictionary items for any specific class. The experimental results on the public action and on gesture recognition databases will reveal that CSMMI compares to the methods on shared dictionary and other approaches such as state-of-the-art. Each class owns a dictionary item classification and is made very easy and also it facilitates parallel execution by increasing speed of execution at each and every step.

IV. FLOW CHART

Flow chart here illustrates the two main phases. They are training phase and the testing phase for gesture recognition.

A. Training phase

In the training phase, training for individual and most occurring shapes will be feeded. Once the image is read, classification is done based on the criteria's and the shape feature is extracted. The obtained feature is thus stored in classification.

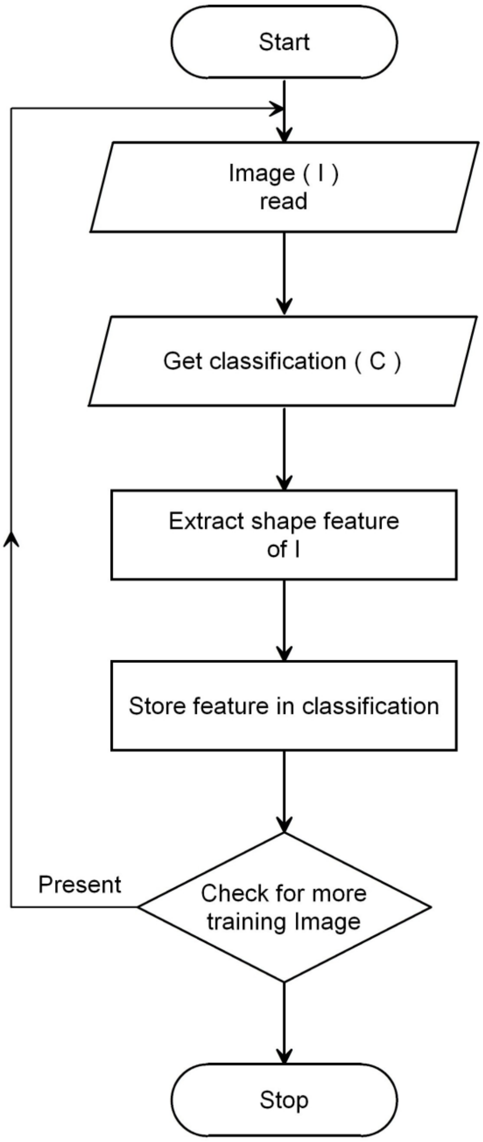
Start
Image (I) read
Get classification (C)
Extract shape feature of I
Store feature in classification
Check for more training Image
Present
Stop

Figure3: Flow chart for training phase

Similarly maximum number of images will be trained and stored in the form of cluster. Thus, when any sequence is given gestures or shapes or actions can be easily identified.

B.Testing phase for gesture recognition

In this testing phase for gesture recognition, an image is fed as input. From the input the shape feature is extracted. Classification is the one where the trained features are stored. It contains n number of clusters. Cluster is formed based on the similarity of the shapes.

The given input is sequentially checked for the match. The distance between the stored or trained image and the image in the input is calculated. The distance thus calculated is stored. The distance is calculated for every similar image. Then the average of the distance in the classification is found.

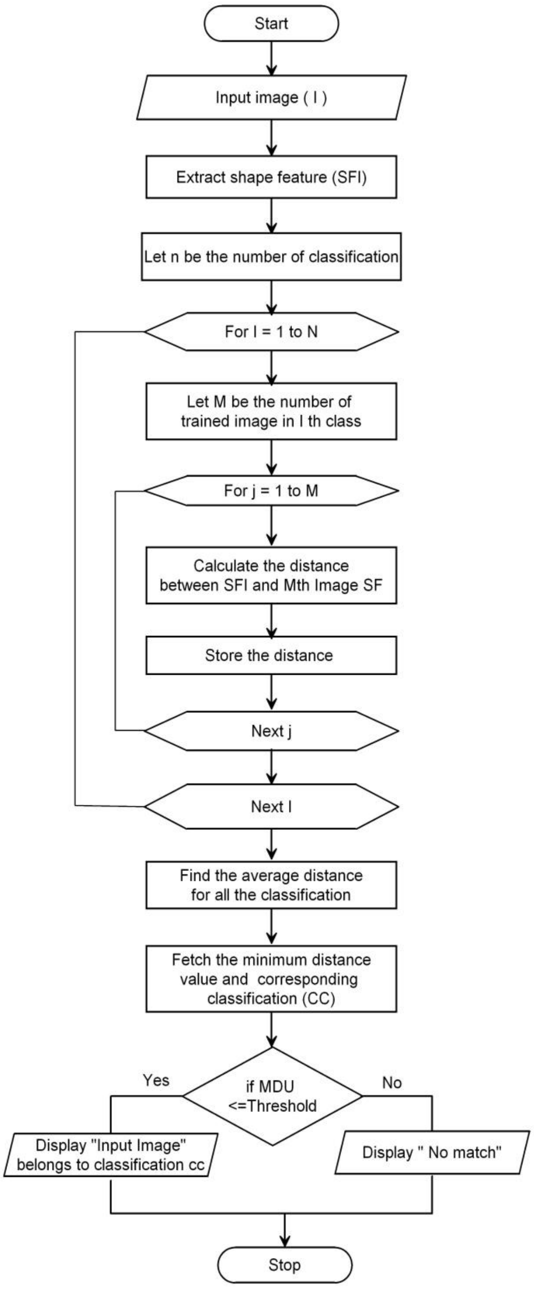
Start
Input image (I)
Extract shape feature (SFI)
Let n be the number of classification
For I = 1 to N
Let M be the number of trained image in I th class
For j = 1 to M
Calculate the distance between SFI and Mth Image SF
Store the distance
Next j
Next I
Find the average distance for all the classification
Fetch the minimum distance value and corresponding classification (CC)
Yes
if MDU <=Threshold
No
Display "Input Image" belongs to classification cc
Display " No match"
Stop

Figure4: Gesture Recognition

The minimum distance is pointed along with the corresponding classification (CC). If the minimum distance value (MDV) is less than or equal to the threshold value then the display would be something like the input image belongs to the classification CC. If the minimum distance value (MDV) is greater than the threshold value then it displays as no match.

V. CONCLUSION

An approach called as class specific dictionary learning for gesture and action recognition is done using information theory. CSMMI's goal is to choose items from dictionary that are more related to any specific class and reject the one that is less related. By this method the speed can be boosted and also complexity can be reduced.

REFERENCES

1. T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 8, pp. 1576–1588, Aug. 2012.
2. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
3. Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in Proc. IEEE ICCV, Nov. 2011, pp. 707–714.
4. S.Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 7, pp. 1294–1309, Jul. 2009.
5. H.Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
6. J. Liu and M. Shah, "Learning human actions via information maximization," in Proc. IEEE Conf. CVPR, Jun. 2008, pp. 1–8.
7. J.K.Aggarwal and M. S. Ryoo, "Human activity analysis: A review," ACM Comput. Surv., vol. 43, no. 3, p. 16, 2011.
8. R.Poppe, "A survey on vision-based human action recognition," Image Vis. Comput., vol. 28, no. 6, pp. 976–990, 2010.
9. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image sesarch," in Proc. IEEE ICCV, vol. 2. Oct. 2005, pp. 1816–1823.
10. H. Jégou, M. Douze, and C. Schmid, "Packing bag-of-features," in Proc. 12th IEEE Int. Conf. Comput. Vis., Sep./Oct. 2009, pp. 2357–2364.
11. H.Su, M. Sun, L. Fei-Fei, and S. Savarese, "Learning a dense multiview representation for detection, viewpoint classification and synthesis of object categories," in Proc. 12th IEEE Int. Conf. Comput. Vis., Sep/Oct. 2009, pp. 213–220.
12. J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in Proc. IEEE Int. Conf. CVPR, Jun. 2007, pp. 1–8.
13. B.Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in Proc. IEEE ICCV, Nov. 2011, pp. 1331–1338.
14. Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in Proc. IEEE Int. Conf. CVPR, Jun. 2011, pp. 1697–1704.
15. W.Dong, X. Li, D. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in Proc. IEEE Int. Conf. CVPR, Jun. 2011, pp.
16. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in Proc. IEEE Int. Conf. CVPR, Jun. 2008, pp.1–8.
17. J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in Proc. IEEE ICCV, Dec. 2013, pp. 533–547.