Vision based Traffic Police Hand Signal Recognition in Surveillance Video - A Survey

R.Sathya
Research Scholar
Dept. of Computer Science and Engg.,
Annamalai University

M. Kalaiselvi Geetha
Associate Professor
Dept. of Computer Science and Engg.,
Annamalai University

ABSTRACT

Human gesture recognition has become a very important topic in computer vision. The purpose of this survey is to provide a detailed overview and categories of current issues and trends. The recognition of human hand gesture movement can be performed at various level of abstraction. This survey concentrate on approaches that aim on recognizing traffic police hand signals. Many application and algorithms were discussed with the recognition framework. General overview of an traffic control gestures and its various applications where discussed in this paper. Most of the recognition system uses the benchmark datasets like KTH, Weizmann. some other datasets were used by the action recognition system. In this paper image representation,action representation, human action detection,feature extraction and human action recognition were also discussed.

Keywords:

Computer vision, traffic control gesture, hand action, feature extraction, Activity recognition.

1. INTRODUCTION

Human action recognition is an significant research area in computer vision. It is the interpretation of action of human hands, body parts or arms that gives a semantic meaning. Human hand action can be very useful mainly at a long distance, where speech information is not available. It has a wide range of applications, this paper presented the current state of the art research in human action recognition. Hand action understanding signs of traffic person is required for traffic surveillance and automated vehicles etc., Human hand gestures are a natural form of a human action.

Action recognition is challenging due to significant variations in the video data that are caused by varying factors which include clothing and the subjects appearance, view point and scale, personal style and action length, self occlusion, background clutter, multiple video objects. In a human body parts the hand is the most effective, general purpose interaction tool due to its smart functionality in communication.

Technically, an action is a sequence of movements generated by a human agent during the performance of a task. Hand action recognition can be viewed at two levels: hand gesture or hand posture recognition. Hand posture is the static pose of palm and finger, for example, pointing, thumbs up etc. Whereas a hand gesture is the dynamic movement, that involves transformation of hand position and orientation such as stopping, waving, calling, etc.,

Most of the current research focus on human action recognition, human behavior analysis, hand action detection, gesture recognition. This paper concentrates on traffic police hand signals. Human gesture recognition for traffic control can be related used for human robot interaction. An human action is done normally with a number of successive actions, which gives an interpretation of the action carried out. Recent literature focus in the area of vision based analysis of human actions and poses from video sequence.

1.1 Traffic Control Gestures

In a human traffic control environment drivers, must follow the directions given by the traffic police officer in the form of human body gestures. To improve the safety of the drivers, research is developing a novel method to automatically recognize traffic control hand signals.

Traffic police control system, a human traffic controller is able to evaluate the traffic within visual range around the traffic intersection. Based on their observations theyuse the human arm directions for classifying the traffic control commands. The traffic control commands are categorized into three types such as, stop all vehicles in every road direction, stop all vehicles in front of and behind the traffic police officer and stop all vehicles on the right of and behind the traffic police officer. Each traffic hand signal is a combination of the arms directions. Twelve Indian traffic hand signal can be constructed from these control command types. The twelve traffic police hand signals are shown in Table 1.
Common actions such as left and right hand raises straight up, left and right hand raises for the left and right and left and right hand raises to the front. However, it is necessary to recognize all the hand signals.

First, the police officer gestures like 'slow down' can be expressed in some emergency traffic lights. Second, in the long term for the intelligent vehicle, an auto-driving vehicles or unmanned vehicles, it is necessary to have them under some special conditions to observe the commands of traffic police. There are two possible solutions to this recognition: active way or passive way. The passive way is to use body sensors to recognize the traffic police gestures. This method is called glove based approaches. Their object may

Table 1. Traffic police hand signals.

Gestures	Traffic police hand signals
1	To start one side vehicles
2	To stop vehicles coming from front
3	To stop vehicles approaching from behind
4	To stop vehicles approaching simultaneously from front and behind
5	To stop vehicles approaching simultaneously from right and left
6	To start vehicle approaching from left
7	To start vehicles coming from right
8	To change sign
9	To start one side vehicles
10	To start vehicles on T-point
11	To give VIP salute
12	To manage vehicles on T-point

be animated. The active way is to use cameras on unmanned vehicles to recognize the traffic hand signals. Due to the illumination problems mentioned before, to recognize the computer vision system. This method is called vision based approaches. Video camera is used to capture the image of hands, which are then processed and analyzed using computer vision techniques. Vision based hand gesture recognition is simple, natural and convenient for users. The active way is the better choice than the passive way.

To cover uniform hand signals and gestures for manual traffic direction and control, the officer should apply a portion where they can be seen clearly by all. These twelve hand signals used by police officers control the flow of vehicles at an intersection.

1.2 General Overview

The area of human action recognition is closely related to other lines of research that analyze human motion from videos. Human action recognition approaches are normally discussed as data glove based approaches and vision based approaches in the literature. The focus of this paper is limited to vision based human action recognition approaches. This paper addresses the issues at different levels as seen in Fig. 1. Moreover this paper point out the challenges involved and discusses the future directions.

The input video is broken down into a set of features taking individual frames into account. The human hands are isolated from their body parts as well as other background objects. The human hand directions for classifying the traffic control gesture commands. Appearance based approach method uses image features to model the visual appearance of the human hand and compare these parameters with the extracted image features from the video input.

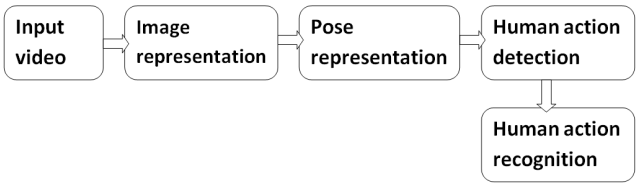

Fig. 1. General architecture

In this overview, many papers have been focused with research detailing several techniques for the segmentation, activity detection and activity recognition. The remainder of this paper is organized as follows. The next section discusses benchmark datasets. Section 3 explains the image representation. Section 4 details human action detection procedure. Section 5 explains the action representation. The human action recognition techniques are presented in Section 6. Application areas of the research area is discussed in Section 7. Finally, Section 8 concludes the paper with future directions.

2. BENCHMARK DATASETS

The benchmark data introduced so far focus more on ”activity recognition” that is video sequences. These datasets are typically pre-segmented and contain only one activity. Some of the benchmark datasets are discussed here.

2.1 Weizmann Benchmark Dataset

Weizmann human action Datasets contains ten actions such as walking, running, jumping-jack, jumping-forward-on-to-legs, jumping-in-place-on-two-legs, waving-two-hands, waving-one-hand, bending, galloping-sideways, performed by nine different actors. The view point is static. Two separate sets are available. One set shows walking movement viewed from different angles. The second set shows front to parallel walking actions with slight variations (carrying object, different clothing, and different styles). The background is static permitting simple background subtraction and foreground silhouettes are included in the dataset. Video clips are at 25 fps.

Fig. 2. Example frame for weizmann Dataset

2.2 KTH Benchmark Dataset

The KTH human action dataset is a mostly exploited while very challenging dataset. It contains six actions such as walking, jogging, boxing, hand waving and hand clapping. They are performed by 25 actors. Video clips are at 25 fps. It contains six hundred video sequences. Each video has only one action performed by twenty five different actors. Four different scenarios are used: Outdoors, outdoors with scale variations, outdoors with different cloths, and indoors. In total, the data consists of 2391 video samples.The backgrounds are relatively static (no background noise). Apart from the zooming scenarios, there is only slight camera movement.

2.3 IXMAS Action Dataset

The INRIA XMAS dataset contains eleven daily-life actions. Action such as chuckle watch, cross arms, scratch head, sit-down, getup, turn around, walk, wave, punch, kick, pick-up, performed each three times by eleven non professional actors. It contains four

Fig. 3. Example frames for KTH Dataset

twenty nine multi view sequences. If the views are considered individually then it consists of 2145 sequences. The actions were filmed with five carefully calibrated and synchronized cameras.

2.4 UCF Sport Action Dataset

Broadcast television UCF Sport action dataset consists of one fifty video sequences performed by thirteen actions type. Ten main actions such as dive, golf, kick, lift, ride, run, skateboard, swing bond, swing side and walk. Most action classes there is considerable variation in action performance, human appearance, camera movement, view point, illumination and background.

2.5 Other Action Datasets

Action datasets containing still images, figure skating, base ball and basket ball are performed in [1]. In [2] Presented a set of still images collected from the web. The crowded videos dataset introduced in [3]. The HOHA dataset [4] are a large collection of short segments of real Hollywood movies annoted with 12 action classes. Mono pedestrian detection study on pedestrian classification in [5] benchmark dataset. Occluded pedestrian classification benchmark dataset described in [6]. The Hollywood human action data set contains eight actions extracted from movies and performed by a verity of actors. Stereo-based pedestrian detection in [7] benchmark dataset. And many action images or action videos collected from the web.

3. IMAGE REPRESENTATION

The features extracted from the image sequences should generalize over the small variations in human appearance, action execution, view point and background. Image representation can be classified as two categories:

1. Global representation

2. Local representation.

3.1 Global Representation

The global representation is obtained by a top-down fashion. Initially, human object localization in the image is analyzed by using background subtraction or tracking method and region of interest (ROI) is prearranged which results in image description. In general, ROI is obtained using background subtraction or tracking.

ROI is divided into a fixed spatial or temporal grid small variations due to noise, partial occlusion and changes in view point can be partly overcome by global grid based representation and space time volumes. Space Time Volumes are often called as spatio-temporal volume (STV). By stacking frames over a given sequences 3 dimension spatio-temporal volume is formed. To drive local space time saliency and oriented features the solutions poisson equation is used. In 3D spatio-temporal volume background subtraction is not necessary, where as 3-D super pixels are obtained from segmenting the STV. Global features for a given temporal range are obtained by calculating weighted moments over these local features. To construct an STV of flow and sample the horizontal and vertical components in space time [8] uses a 3-D variance of the rectangle features.

3.2 Local Representation

Local image representation contains a collection of local descriptor or patches to describe the observation. In this local representation the accurate localization and background subtraction are not essential. Local representation differs from spatial representation, these are somewhat invariant to changes in view point, human appearance and partial occlusion.

The spatio temporal size of a patch is usually determined by the scale of the interest point. Patches can also be described by local grid-based descriptors. Position of the human body and size of the head is analyzed based on the grid spans. A subset of all possible blocks within the grid is selected using AdaBoost. Space time interest point detectors locates the sudden changes of action in the video sequnces. And these locations are informative for human action recognition. In space time interest point the local neighborhood has significant variation in both spatial temporal domains. Space time interest point detects subspace of correlated movement instead of detecting interest point over the entire volume. Correlation between local descriptors can also be obtained by tracking features.

4. HUMAN ACTION DETECTION

Human action detection is an important component of computer vision system in video sequnces. The essential step is to identify the feature set that separate the human from the background even in cluttered scenes for identifying the human action performed.

4.1 Feature Extraction for Recognition

Feature extractions is the main vision task in human action recognition and consist in extracting hand gesture, posture, facial expression, behavior, gait and motion cues from the video that are discriminative with respect to human action. The features are the useful information that can be extracted from the segmented human object by which the machine can understand the meaning of that posture. The features are extracted from foreground images such as texture, type of cloths or color. Such features are usually extracted directly from video.

4.2 Spatio Temporal Feature

The Spatio Temporal (ST) features have recently become a popular video representation for human action recognition and content-based video. The ST feature normally captures the strong variation of the data in spatio and temporal direction that are caused by motion of the actor. ST features are extracted several methods to extend two dimensional features to the temporal dimension. The most representative method is cuboids where many local cubic spatio-temporal regions are extracted. However, spatio temporal features contain only the appearance and motion information and ignore the shape of the information. Local space-time features capture characteristic shape and motion in video and provide relatively indepen-

dent representation of events with respect to their spatio-temporal shifts and scales as well as background clutter and multiple motions in the scene. Several different space-time feature detector [9] and descriptors [10] have been proposed in the past years. Space-time feature detectors usually select spatio temporal features. Feature descriptors usually capture shape and subject in the neighborhoods of selected points using image measurements such as spatio or spatio-temporal feature gradients.

4.3 Motion History Image

Motion History Image (MHI) is a static image template where pixel intensity is a function of the recency of motion in a sequence. Two types of MHI features such as projection profile based features and centroid based feature. Projection profile based feature indicates the bias of the MHI along horizontal and vertical direction the centroid of MEI. This indirectly will convey the temporal information of motion along horizontal and vertical direction. The centroid of MHI is different because it is computed using grey-level time stamp values as weights in the summation. It gives the approximate direction of the movement of centroid for the corresponding action. MHI for obtaining the spatial location and the temporal properties of human action from raw video sequences is seen in [4].

4.4 Motion Flow History

Motion Flow History (MFH) is a static image. Three types of the common features are extracted from MFH. It comprises of the entire history of spatial motion information, many useful features are extracted from MFH. Three types of features such as affine features, projected 1D feature and 2D polar feature can be extracted. The affine parameters are estimated by standard linear regression techniques. The regression is applied separately on each motion vector component since the x affine parameter depends only on horizontal component of motion vector and y parameter depends only on vertical component of motion vector. [11] proposed a method for constructing MFH and MHI from compressed video with minimal decoding.

4.5 Spin-Image Feature

Spin-images have been successfully used for object recognition. For actions, the spin-images can provide a richer representation of how the local shape of the actor is changing with-respect to different reference point is seen in [12]. These reference points may correspond to different limbs of the human body. Instead of attempting pair wise matching of spin-images to match two actions, they use the bag of spin-image strategy. [13] first apply PCA to compress the dimensionality of the spin-image, and then use K-means to quantize them. Then call the group of spin-images as a video-word. Finally, the action is represented by the bag of video-words model.

4.6 Silhouette Feature

A silhouette is the image of a person where motion is represented as a solid figure of a single color, usually black, it gives outline of the subject. [14] silhouette based method aims recognize actions by characterizing the shape of the motion silhouette through space time. The interior of a silhouette is featureless. It is usually extracted by finding the difference between background and current image. Using only the contour points and not the whole silhouette is motivated by getting rid of the redundancy that introduces the inside part of the human silhouette, leading therefore to a less expensive feature extraction.

4.7 Mid-Level Motion Feature

The mid-level motion features are focused on local regions of the image sequence. These features are tuned to discriminate between different classes of action, and are efficient to compute at run-time. Mid-level motion features are weighted combinations of threshold low-level features. Each mid-level feature covers small spatiotemporal cuboids, part of the whole figure-centric volume, from which its low-level features are chosen. Low level feature corresponds to a location in the figure-centric volume. For some small cuboids inside the figure-centric volume, adaboost algorithm [15] is used to select a subset of the weak classifiers inside each figure-centric volume to construct better classifiers.

4.8 Low-Level Features

The low-level motion features are calculated by a figure centric spatio-temporal volume for each person. For each frame, low-level motion features are extracted from optical flow channels at pixel locations in that frame and a temporal window of frames adjacent to it. Two types of low-level features such as optical flow and histogram of oriented gradients 3D (HOG3D) are seen in the literature. Optical flow is calculated from the entire human figure to capture global motion information. HOG3D descriptor is extracted from each cuboid to characterize the local motion and appearance information. The local spatial-temporal feature can deal with noise and partial occlusion. Low level features is some geometric features which can extracted quickly and considered robust to noise [16].

4.9 Haar-Like Feature

Haar-like feature have been successfully used in face detection. Every Haar feature can be regarded as a template of several white and black rectangles interconnected. Three types of Haar like features are seen. These features are used in learning the characteristics of human arm posture, such as edge features, center surround features and line features [17]. The adaBoost learning algorithm can considerably improve the overall accuracy stage by stage by using a linear combination of these individually weak classifier. [**?**] Navie bayes method is an effective and fast method for static hand gesture recognition.

4.10 Skeletal Feature

The skeletal feature are used for separating human body model into several human body parts like face, torso, hand and limbs. Human action recognition system based on segmented skeletal features which are separated into several human body parts. [18] the size and position of neck, lower or upper arms and hands can be expressed as a function of heads size and location as follows. The neck space is of a head length. The hand is of a head length. The lower arm is 5/4 heads length. The upper arm is 3/2 heads length.
A fixed size feature vector, which is invarient to translation, rotation and scaling must be extracted for each skeleton. The human body are 15 joints which are used to denote 15 corresponding human body parts.

Skeleton based object recognition systems generally perform better than shape based object recognition approaches [19]. Human skeleton is extracted using normalized gradient vector flow in the space of diffusion tensor fields, using the eigen values and eigen vectors of the segmented skeletal features.

4.11 Contour-Based Feature

Contour-based feature representations have a long history in object recogniiton and computer vision. In contour-based approaches, often the first step is detected from edges. To extract the contours, the document images first need to binarized. In this method, instead of tracking the whole set of pixels comprising an object, the algorithm tracks only the contour of the object. [20] proposed contour based nonridge object tracking method via the contour energy function. Trackd the complete region of the nonridge objects and recovered the occluded object parts.

5. ACTION REPRESENTATION

Human action is characterized by a spatial element, which is the body poses at each time step and a temporal element which is the evaluation of the body poses. An alternative of including body poses in all frames have a more compact representation of a human action. Human action representation has two major advantages. First, actions are recognized by comparison with known action models, a small number of key poses reduces the computational complexity. Second, focusing on the key poses only, that can capture the essence of an action class even if there is variance in execution styles of the same action.

Human action representations can be classified as two categories:

1. Spatial action representation

2. Temporal action representation

5.1 Spatial Action Representation

Spatial action representation is used to discriminate actions from visual data. Various representations have been suggested. They mainly contrast by the amount of high level information they represent versus how efficient they are in practice. The spatial representation can be categorized into three main groups such as body model, image model and spatial statistics.

5.1.1 Body Model. Represent the spatial structure of the human body. Video streams are observed each frame, the pose of human body is recovered from a variety of available features. [21] human action recognition is performed based on such pose estimates. Recognition divides the task of action recognition in two separate stages. A motion captures stage which estimates a 3D model of the human body. [22] human body typically represented as a kinematic joint model and an action recognition stage which operates on joint trajectories.

5.1.2 Image Model. Global image based representations of actions also called holistic representations, which do not require the detection and labeling of individual body parts. Image models can be much simpler than parametric body models. In most cases, features are then computed densely one regular grid bounded by the detected regions. [23] presented a typical image model, where images of hand gestures are directly correlated without feature extractions. An important class of image models uses silhouettes and contours of the human agent performing the actions.

5.1.3 Spatial Statistics. Local representations of action which decompose the image or video into smaller regions, not linked to body parts or image coordinates. [24] action recognized based on the statistics of local features from all regions. Such approaches are typically based on bottom up strategies, which first detect interest point in the image. Mostly at corner or blob like structures and then assign each region to a set of preselected vocabulary features.

5.2 Temporal Action Representation

Temporal action representation is classified under three main categories, viz grammars, templates and temporal statistics.

5.2.1 Action Grammars. Action grammars represent an action as a sequence of moments, each with their own dynamic and appearance. A common way to approximate a dynamical system over feature observation is to group features into similar configuration. Models for generally into the class of graphical models, which are best described as probabilistic grammars. [25] probabilistic grammars used for action recognition the most prominent is certainly the hidden Markov model. Evaluation and learning of grammar-based action recognition remains an outstanding problem with large numbers of actions classes.

5.2.2 Action Templates. Templates are typically computed over long sequences of frames, and should not be confused with spatio-temporal features or optical flow, which are computed over small time windows and serve as components of other action classifiers. Template based representations of very different kind have been proposed. Generally they are effective and discriminative action representations, and in particular attractive for action classification. [26] introduce templates of body feature trajectories after tracking over extended time sequences.

5.2.3 Temporal Statistics. Temporal statistics looks at the data for each time step and computers some statistical information of how a point or variable changes over time. The single static image sequences can be encoded without taking temporal relations into account. Temporal bag of features used to represent sequences simply base on the frequency of feature occurrence over time [27]. Examples are methods that learn an appearance model of action from a single characteristic key-frame as in a photograph or the histogram of a features.

6. HUMAN ACTION RECOGNITION

Human action recognition or classification of hand gestures are the last phase of the action recognition system. Human hand gestures can be classified using two approaches. These approaches are rule based approaches and machine learning based approaches.

6.1 Rule based Approaches

Rule based approaches are represents the input features as manually encoded rule, and the winner gesture is the one that matched with the encoded rules after thier features has been extracted. The main problem of this technique is that the human ability is encoding the rules limits the successfulness of the recognition process [28].

6.2 Machine Learning based Approaches

Machine learning based algorithm can be divided into two approaches. These two approaches are supervised and unsupervised approaches. Supervised learning is the machine learning task of inferring a function from labeled training data. In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Unsupervised learning is closely related to the problem of density estimation in statics.

6.2.1 Action Recognition using the SVM. Support vector machine (SVM) is a most popular technique for classification in visual pattern recognition [29]. SVM is a kernel-based technique which is based on the principle of structured risk minimization (SRM). SVM constructs a linear class boundaries based on support vectors.

SVM are set of related supervised learning model with associated learning algorithm that analyze data and recognize patterns used for classification and recognition.They belong to a family of generalized linear classifiers.

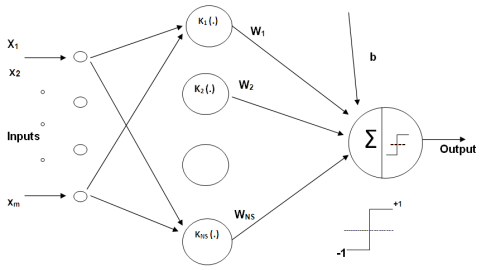

Fig. 4. Achitecture of the SVM (Ns is the number of support vectors).

SVM is linear separable data. A linear SVM is used to classify data sets which are linearly separable. The SVM linear classifier tries to maximize the margin between the separating hyperplane. SVM is linear non-separable data. It maps the data in the input space into a high dimension space with the kernel function to find the separating hyperplane. SVM inner product Kernels are three types.

1. Polynomial kernel

2. Gaussian kernel

3. Sigmoid kernel

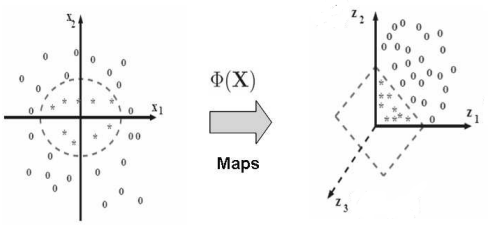

Fig. 5. Non-leaner, Linear structure

Vision based gesture recognition for alphabetical hand gesture recognition using SVM [30]. These gesture recognition system for alphabetical hand gesture is build. The system is designed using the Support Vector Machines classifier which is widely used for classification and regression testing.

6.2.2 Action Recognition using the HMM. In Markov Model (MM) the state is directly visible to the observer, so the state transition probability is the only parameter. The Hidden Markov Model (HMM) models are sequence of observations as a piecewise stationary process. In hidden markov model the states are not directly accessible to the observer. Each state has probability distribution over output tokens. The sequence of tokens generated by an HMM gives some information about the sequence of state. Hidden variable controls the components to be selected for each observation. The HMM is stochastic approach which models the given problem as a ”doubly stochastic process” in which the observed data are thought to be the result of having passed the hidden processes are to be characterized using only the one that could be observed.

Two types of HMM model.

1. Ergodic model

2. Left-Right model.

Every state of the ergodic model can be reached from every other state in a finite number of steps. Left-Right model has the property that as time increase the states index increases that states proceed from left to right. The HMM will be useful in real world applications, if the following three basic problems of HMM are solved [35].

Three basic problem of HMM:

1. Evaluation problem (forward algorithm)

2. Decoding problem (Viterbi algorithm)

3. Learning problem. (Forward-Backward algorithm)

The Hidden Markov Models are a popular technique for recognizing human gesture in a varity of applicatisons and sensor configuration. Hidden Markov Models are double stochastic process as governed by an underlying Markov chain with a finite number of states, and a set of random functions each of which is associated with one state [31].

The following is an illustrative list of applications of HMM:

Speech recognition

Gait recognition

Optical character recognition

Lip-reading (visual speech to text mapping)

Gesture and body motion analysis

Several hand gesture recognition systems have been developed using various features computed from static images or image sequences [32]. The vision-based method selects the input data as the feature vectors for the HMM input and other HMM-based [33, 34] hand gesture recognition systems have also been development. Gesture recognition system using HMM models has been developed.

6.2.3 Action Recognition using the KNN. K-nearest neighbors (KNN) classifiers have a good performance when the attributes of a system are linearly separable. The class which has the most vectors in those K neighbors is chosen to be the class of the input vector. A cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering is an unsupervised learning method which deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way.

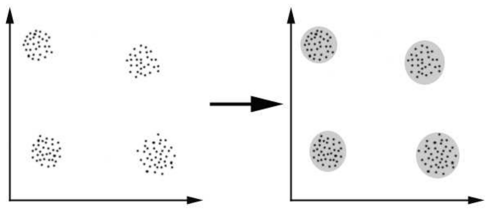

Fig. 6. Clustering

k-means clustering [35, 36] is an algorithm to group objects based on attributes/features into k number of groups where k is a positive integer. The rouping (clustering) is done by minimizing the

Euclidean distance between data and the corresponding cluster centroid. Thus the purpose of k-means clustering is to cluster the data. They calculate the distance between cluster centroid to each object using Euclidean distance measure. E-M algorithm [35, 36] finds out maximum likelihood estimates of parameters in probabilistic models. This algorithm iterates between the E- step and the M-step until convergence. Expectation step computes an expectation of the likelihood assuming parameters. Maximization step computes maximum likelihood estimates of parameters by maximizing the expected likelihood found in E-step.

K-nearest Neighbors With Distance Weighting (KNNDW) is an improvement which has been proved to perform better than KNN in many cases [37]. In this method, the contribution of each neighbor to the overall classification is weighted by its distance from the point being classified.

6.2.4 Action Recognition using the NN. Artificial Neural Network (ANN) is an information processing system that is inspired by biological nervous system like the brain process information. A neural network is a machine that is designed to model the way in which the brain performs a particular task or functions. The network is usually implemented by using electronic components or simulated in software on a digital computer. To achieve good performance, neural networks employs massive interconnections of simple computing cells referred to as ”neurons” or ”processing units”. Gesture recognition is an important for developing alternative human-computer interaction modalities using ANN.

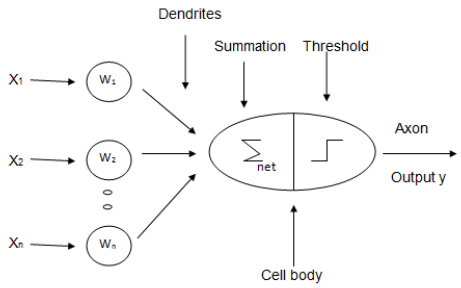

Fig. 7. Structure of artificial neuron

Three types of Activation function.

1. Linear/Threshold function

2. Sigmoid/Squashing function

3. Hyperbolic tangent function

Backpropogation Neural Network (BPNN) is a multi layer feed forward neural network. BPNN structure is input layer, Hidden layer and output layer. Probabilistic neural network (PNN) is a feed-forward neural network that implements a Bayesian decision strategy for classifying input vectors. Artificial Neural networks are flexible in a changing environment [38] and it also describes the process of gesture recognition using ANN [39]. Probabilistic Neural Network (PNN) is a feed-forward neural network that implements a Bayesian decision strategy for classifying input vectors. Neural network models such as Backpropagation Neural Network (BPNN) [40] and radial basis function neural network (RBFNN) are used for pattern classification because of their ability to capture the nonlinear hyperspace separating the classes in the feature space.

A special kind of backpropagation neural network called Autoassociative Neural Network (AANN) [41] can be used to capture the distribution of feature vectors in the feature space.

7. APPLICATION AREAS OF HAND ACTION

Vision based hand actions are modeled as sequences of multiple events. Every event is matched independently with its own event model and linear time scaling. Human hand action recognition constitutes matching the appropriate events sequentially. Recognition of hand gestures is performed using a probabilistic finite state machine. Human hand has abundant joints. It is able to form several distinct hand shapes. Human hand shapes are quite useful to represent different communication signs, which are defined to execute specific tasks.

7.1 Robotics

Gestures are used for controlling robots, corresponding to virtual reality interaction system. For virtual reality application gestures are considered as one of the effective spreading stages in computing area [42]. Gesture-to-speech application allow hearing-impaired people to communicate with their surrounding environments through computers. It converts hand gestures into speech. [43] introduced hand gestures to speech using neural networks for data glove device.

Computer vision and hand gesture recognition techniques developed to control the VLC media player. [44] developed a vision based low cost input device for controlling the VLC player through human hand gestures using recognition techniques. Human hand postures and hand gestures controlling television is seen in [45].

7.2 Sign Language Gestures

In vision based hand gesture recognition, hand shape segmentation is one of the toughest problems under a dynamic environment. It can be simplified by using visual marking on the hands.

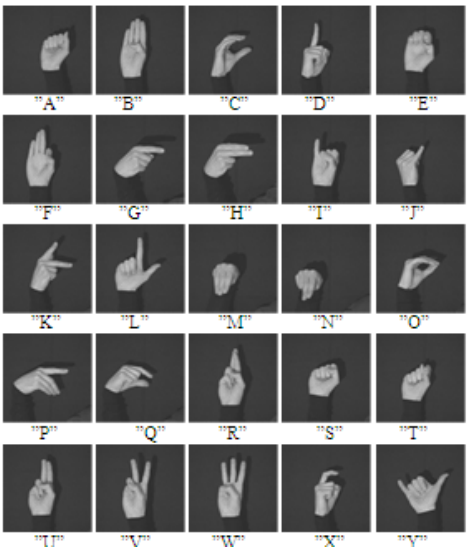

Fig. 8. American sign language

Some researchers have implemented sign language and pointing gesture recognition based on different marking modes [46]. American sign language gestures are recognized [47]. It is one example that has received significant attention in the gesture literature. Korean sign languages [48] are recognized and a new gesture recognition algorithm for Korean scripts. [49] Taiwanese sign language (TWL) introduced lexicon of 250 vocabularies. Arabic sign language (ArSL) are regonized and classified [50]. Japanese sign language [51] recognized for words and alphabets, they could recognize 10 words and 42 alphabets using two types of neural network algorithms.

7.3 Traffic Police Hand Signals

The traffic police gesture systems are mainly expressed by arms. According to Indian traffic rules there are 12 hand gestures. The Chinese traffic police gesture system is defined and regulated by Chinese ministry of public security [52] Gesture of traffic police officers are captured in the form of depth images. In road traffic control system [53] consider only the arm directions for classifying the traffic control commands. Traffic gesture using three types of control commands like six defined traffic gestures are used.

Fig. 9. Traffic police hand signals

The Indian traffic police officer hand signal consists of twelve signals. These hand signals are discussed in section 1.1. Police officer stands straight with weight evenly distributed on both feet and allowing their hands and arms to hang easily on the officer sides except when gesturing.

8. FUTURE DIRECTIONS

Real time processing of human action recognition in the field of traffic surveillance is very essential. Benchmark data sets focus on particular application domain. This paper proposes current issues and state of the art research in human action recognition. Action recognitions are applicable in wide range of application. Major challenges in this area are illumination effects, various poses of the police officer, viewing directions, occlusions etc. If the challenges ahead of this research are fulfilled, this would be a great step towards achieving a robust interpretation and recognition of actions in rear future.

8.1 Conclusion

This paper discussed the detailed overview and categories of current issues and trends in action recognition, focusing towards recognizing the police hand signals. Several applications of human action recognition is discussed in this paper. Benchmark datasets like KTH dataset, Weizmann dataset and other datasets are discussed.

In this survey basic concepts and techniques behind action recognition system has been studied. Vision based approaches is a widely used method in action recognition since, it is very realistic approach and also it provides better results while compared to data glove approaches. The study of the traffic control gestures and its various applications were discussed in this paper. The different types of techniques to recognize hand gesture are reviewed and analyzed. Human machine interaction can be achieved by different action recognition technique so that the hand action can be recognized even if they are not performed perfectly.

9. REFERENCES

[1] Yang Wang, Hao Jiang, Mark, S., Drew, Ze-Nian Li, and Greg Mori. june, 2006. Unsupervised discovery of action classes, *In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR06)*, Vol. 2, pp. 1654-1661.

[2] Nazli Ikizler, Ramazan, G., Cinbis, Selen Pehlivan, and Pinar Duygulu. december, 2008. Recognizing actions from still images, *In: Proceedings of the International Conference on Pattern Recognition (ICPR08), Tampa, FL*, pp. 1-4.

[3] Yan Ke, Rahul Sukthankar, and Martial Hebert. october, 2007. Event detection in crowded videos, *In: Proceedings of the International Conference On Computer Vision (ICCV07), Rio de Janeiro, Brazil*, pp. 1-8.

[4] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. 2008. Learning realistic human actions from movies, *In: Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

[5] Munder, S. and Gavrila, D. M. november, 2006. An Experimental Study on Pedestrian Classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, pp. 1863–1868.

[6] Enzweiler, M., Eigenstetter, A., Schiele, B. and Gavrila, D. M. 2010. Multi-Cue Pedestrian Classification with Partial Occlusion Handling, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Keller, C., Enzweiler, M. and Gavrila, D. M. 2011. A New Benchmark for Stereo-based Pedestrian Detection, *Proc. of the IEEE Intelligent Vehicles Symposium*, pp. 691–696.

[8] Paul, A. Viola, Michael, J. Jones. December, 2001. Rapid object detection using a boosted cascade of simple features, *In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR01)*, Vol. 1, pp. 511-518.

[9] Willems, G. Tuytelaars, T. and Van Gool, L. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector, *In ECCV 2, Volume 5305 of letecture notes in computer science*, pp. 650–663.

[10] Klser, A. Marszaek, M. and Schmid, C. September, 2008. A spatio-temporal descriptor based on 3Dgradients, *In BMVC*, pp. 995–1004.

[11] Babu, R. Venkatesh, and Ramakrishnan, K. R. November, 2003. Recognition of human actions using motion history information extracted from the compressed video, *Image Vision Computer*, Vol. 22, pp. 597–607.

[12] Johnson, A. and Hebert, M. May, 1999. Using Spin Images for Efficient Object Recognition in Cluttered 3D scenes, *IEEE Tran. On PAMI*, Vol. 21, No.5.

[13] Jingen Liu, Saad Ali, Mubarak Shah. 2008. Recognizing Human Actions Using Multiple Features, *Computer Vision and Patteren Recognition*.

[14] Alexandros Andre Chaaraoui, Pau Climent-Perez, and Francisco Florez-Revuelta. November, 2013. Silhouette-based human action recognition using sequences of key poses, *Pattern Recognition Letters*, Vol. 34(15), pp. 1799–1807.

[15] Viola, P. and Jones, M. 2001. Robust real-time object detection, *In 2nd Intl. Workshop on Statistical and Computational Theories of Vision*, vol. 2, 3, 4.

[16] Murthy, G. R. S. and Jadon, R. S. 2009. A review of vision based hand gestures recognition. *International Journal of Information Technology and Knowledge Management*, Vol. 2, No. 2, pp. 405–410.

[17] Oana-Mihaela Vultur, Ovidiu Gherman. 2010. Human Arm Detection using Haar-like Features, *DOCT-US, an II, nr*, Vol. 2(2).

[18] Quoc Khanh Le, Chinh Huu Pham and Thanh Ha LeRoad. July, 2012. Traffic Control Gesture Recognition using Depth Images, IEEK Transactions on Smart Processing and Computing, Vol. 1, No. 1.

[19] Sang Min Yoon, Kuijper, A. Aug, 2010. Human Action Recognition using Segmented Skeletal Features, *20th International Conference on Pattern Recognition (ICPR)*, pp. 3740–3743.

[20] Yilmaz, A. Li, X. and Shah, M. 2004. Contour-based object tracking with occlusion handling in video acquired using mobile cameras, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 26(11), pp. 1531-1536.

[21] Johansson, G. 1973. Visual perception of biological motion and a model for its analysis, *Perception and Psychophysics 1414*, Vol. 2.

[22] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Feb, 2011. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition, *Computer Vision and Image Understanding*, Vol. 115(2), pp. 224–241.

[23] Drrell, T. and Pentland. 1993. Space-time gesture, *In: Conference on Computer Vision and Patteren rcognition*, pp. 335–340.

[24] Cuzzolin, F., Sarti, A. and Tubaro, S. 2004. Action modeling with volumetric data, *in: International Conference on Image Processing*, Vol. 2, pp. 881–884.

[25] Rabiner, L. R. 1990. A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, pp. 267–296.

[26] Lv, F., Nevatia, R. and Lee, M. 2005. 3d human action recognition using spatiotemporal motion templates, *in: ICCV Workshop on Human-Computer Interaction*, pp. 120.

[27] Schuldt, C., Laptev, I. and Caputo, B. 2004. Recognizing human actions: A local svm approach, *in: International Conference on Pattern Recognition*,pp. 32–36.

[28] Murakami, K. and Taguchi, H. 1999. Gesture recognition using recurrent neural networks. *ACM, Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technologyCHI*, vol. 91, pp. 237-242.

[29] Lewis, J.P. 2004. Tutorial on SVM, *CGIT Lab, USC*.

[30] Cortes, C. and Vapnik, V. 1995. Support vector networks, *Machine learning*, Vol. 20, pp. 53–60.

[31] Oren Boiman, Michal Irani. 2007. Detecting irregularities in images and in video, *International Journal of Computer Vision (IJCV)*, vol. 74 (1), pp. 17-31.

[32] Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected application in speech recognition, *Proc. IEEE 77*, pp. 267–293.

[33] Huang, T.S. and Pentland, A. Hand gesture modeling, analysis, and synthesis, June, 1995. *Proceedings of the International Workshop on Automatic Face-and Gesture-Recognition, Zurich, Switzerland*, pp. 73–79.

[34] Campbell, L.W., Becker, D.A., Azarbayejani, A., Bobick, A.F. and Plentland, A. 1996. Invariant features for 3-D gesture recognition, *Proceedings IEEE Second International Workshop on Automatic Face and Gesture Recognition*, pp. 157–162.

[35] Duda, R.O., Hart, P.E. and Stork, D.G. 2003. *Pattern Classification, John Wiley and Sons, Singapore*.

[36] Lawrence R Rabiner. Feb, 1989. A tutorial on hidden markov models and selected applications in speech recognition, Vol. 77(2), pp. 257-286.

[37] Aseema Sultana, T. and Rajapuspha. july, 2012. Vision Based Gesture Recognition for Alphabetical Hand Gestures Using the SVM Classifier, *International Journal of Computer Science and Engineering Technology (IJCSET)*, Vol. 3(7).

[38] Pujan Ziaie, Thomas Mller, Mary Ellen Foster, Alois Knoll. November, 2007. Using a Nave Bayes Classifier based on K-Nearest Neighbors with Distance Weighting for Static Hand-Gesture Recognition in a Human-Robot Dialog System.

[39] Miss. Shwetah, K. Yewale, MR. Pankaj, K. Bharne. April, 2011. Artificial neural network approach for hand gesture recognition,*International Journal of Engineering Science and Technology (IJEST)*, Vol. 3(4).

[40] Philip, D. Wasserman. 1989. Neural Computing Theory and Practice, *New York*.

[41] Yegnanarayana, B., Gangashetty, S.V. and Palanivel. December, 2002. S. Autoassociative neural network models for pattern recognition tasks in speech and image, *in Soft Computing Approach to Pattern Recognition and Image Processing, World Scientific publishing Co. Pte. Ltd, Singapore*, pp. 283-305.

[42] Murakami, K. and Taguchi, H. 1999. Gesture recognition using recurrent neural networks, *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technologyCHI, ACM*, pp. 237–242.

[43] Fels, S. S. and Hinton, G. E. 1998. Glove-talk IIA neural-network interface which maps gestures to parallel formant speech synthesizer controls, *IEEE transactions on neural networks*, Vol. 9(1), pp. 205–212.

[44] Rautaray, S. S. and Agrawal, A. 2010. A vision based hand gesture interface for controlling VLC media player, *International Journal of Computer Applications*, Vol. 10(7).

[45] Freeman, W. T. and Weissman, C. D. 1995. Television control by hand gestures, *IEEE International Workshop on Automatic Face and Gesture Recognition, Zurich*, pp. 179–183.

[46] Fan Guo, Zixing Cai, Jin Tang. 2010. Chinese Traffic Police Gesture Recognition in Complex Scene, *International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11*.

[47] Starner, T., Weaver, J. and Pentland, A. 2002. Real-time American Sign Language recognition using desk and wearable computer based video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20(12).

[48] Kim, J., Jang, W. and Bien, Z. 1996. A dynamic gesture recognition system for the Korean sign language (KSL), *IEEE transactions on systems, man, and cybernetics-part B: Cybernetics*, vol. 26(2), pp. 354–359.

[49] Liang, R. and Ouhyoung, M. 1998. A real-time continuous gesture recognition system for sign language, *IEEE Third International Conference on Automatic Face and Gesture Recognition Proceedings*, pp. 558–567.

[50] Maraqa, M. and Abu-Zaiter, R. 2008. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks, *IEEE First International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2008*, pp. 478–48.

[51] Murakami, K. and Taguchi, H. 1999. Gesture recognition using recurrent neural networks. *ACM, Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technologyCHI '91*, pp. 237–242.

[52] Pavlovic, V. I., Sharma, R. and Huang, T. s. 1997 Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19(7), pp. 677–695.

[53] Ben Wang, Tao Yuan. 2008. Traffic Police Gesture Recognition using Accelerometers, Vol. 1, pp. 4244–2581.