Object recognition and localization

Badri Narayana Patro
Dept. of Electrical Engineering
badri@iitk.ac.in

Ganesh Boddupally
Dept. of Electrical Engineering
boddgane@iitk.ac.in

***Abstract*—Aim of this project is to detect and recognize a particular object in a video and then find out corresponding timing details of that object present in the video sequence, i.e, what are frames that object is available or the different time interval this object is available in the video scene. we proposed an algorithm which will take care of the object recognition and localization in a video. Object detection and classification are the major part of the object recognition. Object detection is carried out with the help of GMM and object localization using method of object proposal calculation. also we have used Bag of word model to training data set.**

***Index Terms*—Object Detection, Bag of Word, Object Recognition, Localization, Classification..**

I. Introduction

In the modern era computer vision,Object recognition and localization in a video are the two primary research problem. Object recognition is concern with [11] object representation, detection, training and classification. Most of literature is dealing with location recognition , which is quite different from object localization.localization means finding the location of the selected object in the video sequence. Video Object Recognition and localization. First, input frame is extracted from the video sequence. Then pre-processed input image frame in order to remove noise, if required, then process for object recognition step. In object recognition process involves interest point detection, extract feature descriptor from interest point then matching with feature data set to in order to labeling the object belongs to one of the category present in the dataset. Finally the labeled object is searched for perfect match in all the frames of the video and note down the matched frames and time durations.

II. Literature Survey

There are various object recognition and localization technique is proposed by different authors as follows. Lowe, et al. 1999[1], has explained a method for single object recognition using set of local feature templates which can use with corner detector and filters. he has verified planar object recognition using SIFT features[2] which can provide match for affine geometric alignment. Also Lowe, et al. 2001[1], extract object outlines with background subtraction which can recognition 3D objects which is more accurate then affine model and also robust to occlusion and illumination invariant. Another well know Feature-based object recognition technique is Bag of key points given by Csurka et al. 2004[4]. This is based on the analogy to learning methods using bag-of-words representation for text categorization. Here,features are quantized into words. This method discusses on three main issues, one is representation, in which object category is represented based on appearance only or location and appearance. Second one is Learning, i.e., how to form the classifier in order to train given training data. Finally, Recognition, i.e., how the classifier is to be used the compared all the stored objects/features with test features or words and take category decision. The Problem with BOW is all word/object have equal probability. Secondly, it doesn't provides Location information of the words, which is important for recognition. In order to reduce quantization error in local descriptor, vocabulary tree is used for scalable recognition. The vocabulary tree, which uses multi-path based on visual words is presented by Nistr et al. 2006[6]. Csurka et al. 2006[5], object recognition problem is divided into following parts, Object representation, learning and classification. Another well know object recognition is techniques pixel-based techniques, which is combination of segmentation and recognition technique. Other recognition methods are not mentioned in the literature, which is not in the scope of the project.

Object Localization based on detection window, which is proposed by Ramanan et al. 2007[17], which incorporates prior positions and color models. Dai et al. 2012, is discussed about learning to predict bounding box localization, which is based on structural learning for sliding window detection approaches by Blaschko et al. 2008[16]. Li et al. [18], tells about location recognition algorithm, not discussed about location present in the image or video. Object localization using background and foreground separation method is proposed by Kalirajan et al. 2015[14]. in this method foreground feature points are localized in video frame depend upon the maximum likelihood feature points over the input video frames.

III. Assumption

Following assumption required for implement our project

- Camera is fixed, means back ground is fixed and foreground is moving.
- Here we assume that selected object is a rigid body means no deformable part. .
- Camera is placed in front of the object means taking front view of the object.

IV. Proposed Algorithm:

Object detection is a computer technology. It is related to computer vision and image processing. It deals with detecting instances of semantic objects of a certain class (such as humans, tigers, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. The main applications of Object detection is in many areas of computer vision, that including video surveillance and image retrieval. The major steps for object object recognitions and localization is as shown in Figure.

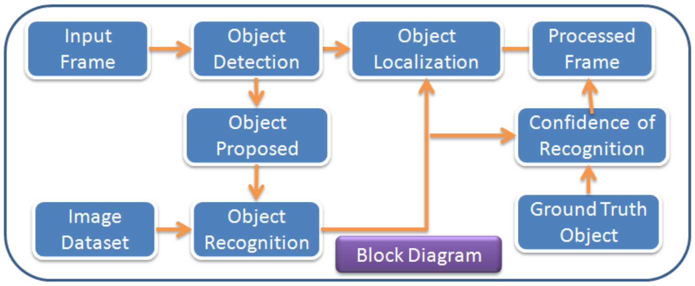

Fig. 1: Block diagram of Proposed Algorithm

Object recognition is a process for identifying a particular object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using feature-based techniques or appearance-based. Proposed Algorithm consist of three parts Object Detection, Object Recognition and Object Localization

A. *Object Detection methods*

*1) **Method 1: Frame Difference:**:* A motion detection algorithm begins with the segmentation part where foreground or moving objects are segmented from the background. The simplest way to implement this is to take an image as background and take the frames obtained at the time t, denoted by I(t) to compare with the background image denoted by B. Here using simple arithmetic calculations, we can segment out the objects simply by using image subtraction technique of computer vision meaning for each pixels in I(t), take the pixel value denoted by P[I(t)] and subtract it with the corresponding pixels at the same position on the background image denoted as P[B]. In mathematical equation, it is written as:

$$P[F(t)] = P[I(t)] - P[B]$$

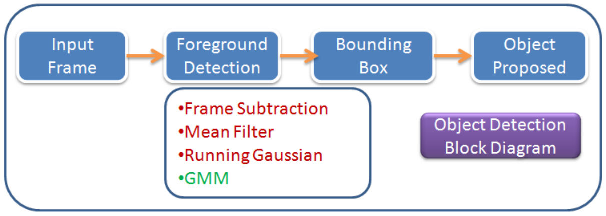

Fig. 2: Block diagram of Object Detection method

*2) **Method 2: Mean Filter:**:* For calculating the image containing only the background, a series of preceding images are averaged. For calculating the background image at the instant t,

$$B(x, y, t) = \frac{1}{N} \sum_{i=1}^{N} V(x, y, t - i)$$

Where N is the number of preceding images

After calculating the background *B(x,y,t)* we can then subtract it from the image *V(x,y,t)* at time *t*=t and threshold it. Thus the foreground is

$$|V(x, y, t) - B(x, y, t)| > Th$$

where *Th* is threshold

*3) **Method 2: Running Gaussian:**:* The pdf of every pixel is characterized by mean μ_t and variance σ_t^2. We are assuming some initial mean and variance.Background may change over time (e.g. due to illumination changes or non-static background objects). To accommodate for that change, at every frame t, every pixel's mean and variance must be updated, as follows:

$$\mu_t = \rho I_t + (1 - \rho)_{(t-1)}$$
$$\sigma_t^2 = d^2 \rho + (1 - \rho)\sigma_{(t-1)}^2$$
$$d = |(I_t - \mu_t)|$$

where I_t is the value of the pixel's intensity at time t_t is the mean of the pixel at time t. Where ρ determines the size of the temporal window that is used to fit the pdf and d is the Euclidean distance between the mean and the value of the pixel. We can now classify a pixel as background if its current intensity lies within some confidence interval of its distribution's mean:

$$|(I_t - \mu_t)|/\sigma_t > k, \quad \text{This is Foreground}$$
$$|(I_t - \mu_t)|/\sigma_t >> k, \quad \text{This is Background}$$

where the parameter k is a free threshold

*4) **Method 2: GMM:**:* Mixture of Gaussians method approaches by modelling each pixel as a mixture of Gaussians and uses an on-line approximation to update the model. In this technique, it is assumed that every pixel's intensity values in the video can be modeled using a Gaussian mixture model. A simple heuristic determines which intensities are most probably of the background. Then the pixels which do not match to these are called the foreground pixels. Foreground pixels are grouped using 2D connected component analysis. At any time t, a particular pixel (x_0, y_0)'s history is

$$X_1, ..., X_t = \{V(x_0, y_0, i) : 1 \leq i \leq t\}$$

This history is modeled by a mixture of K Gaussian distributions:

$$P(X_t) = \sum_{i=1}^{K} W_{i,t} N(X_t|\mu_{i,t}, \Sigma_{i,t})$$

$$N(X_t|\mu_{i,t}, \Sigma_{i,t}) = 1/((2\pi)^{D/2}|\Sigma_{i,t}|^{1/2})$$

$$exp(1/2(X_t - \mu_{i,t})^T \Sigma_{i,t}^{-1}(X_t - \mu_{i,t}))$$

An on-line K-means approximation is used to update the Gaussian

B. Running Gaussian:

Background modeling and subtractions core component in motion analysis. The central idea behind such module is to create a probabilistic representation of the static scene that is compared with the current input to perform subtraction. Background modeling is at the heart of any background subtraction algorithm. Background modeling uses the new video frame to calculate and update a background model. Background modeling techniques can be classified in

C. Object Detection Method 2:

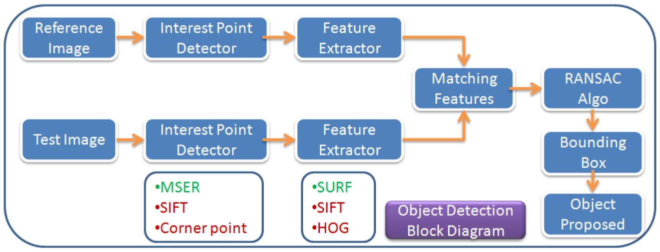

Fig. 3: Block diagram of Object Detection method

*1) **Interest point detector:**:* Interest point detector is used to detect the interest points for subsequent processing. Here we are using MSER(maximally stable extremal regions),SIFT and Corner point detector.

*2) **MSER** :* It is used as a method of blob detection in images. This technique was proposed by Matas et al.[1] to find correspondence between image elements from two images with different viewpoints. This method of extracting a comprehensive number of corresponding image elements contributes to the wide-baseline matching, and it has led to better stereo matching and object recognition algorithms.

3) Scale-invariant feature transform (or SIFT): is an algorithm in computer vision to detect and describe local features in images.

*4) **Corner detection**:* Corner detection is an approach used within computer vision systems to extract certain kinds of features and infer the contents of an image. Corner detection is frequently used in motion detection, image registration, video tracking, image mosaicing, panorama stitching, 3D modelling and object recognition. Corner detection overlaps with the topic of interest point.

*5) **Feature extractor**::* The SIFT approach, for image feature generation, takes an image and transforms it into a ”large collection of local feature vectors” (From ”Object Recognition from Local Scale-Invariant Features”, David G. Lowe). Each of these feature vectors is invariant to any scaling, rotation or translation of the image.

Speeded Up Robust Features (SURF) is a local feature detector and descriptor that can be used for tasks such as object recognition or registration or classification or 3D reconstruction. It is partly inspired by the scale-invariant feature transform (SIFT) descriptor. The standard version of SURF is several times faster than SIFT .

The histogram of oriented gradients (HOG) is counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy.

*6) **RANSAC Algorithm**::* Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. Therefore, it also can be interpreted as an outlier detection method. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed. RANSAC is used to solve the Location Determination Problem (LDP), where the goal is to determine the points in the space that project onto an image into a set of landmarks with known locations.

D. Block Diagram of Object Recognition :

Extract Key Points locations using the Grid method. Grid-Step is [8 8] and Block Width is [32 64 96 128] : to take care of Scale Information. Feature Descriptors: Extracting SURF features from the selected interest point. Strongest Features: 80 percent of the strongest features. Find the minimum no of strongest feature among all the data set. Lets say M is minimum no of feature among all N(14) Image data set, each is having 100,200, 300 images.

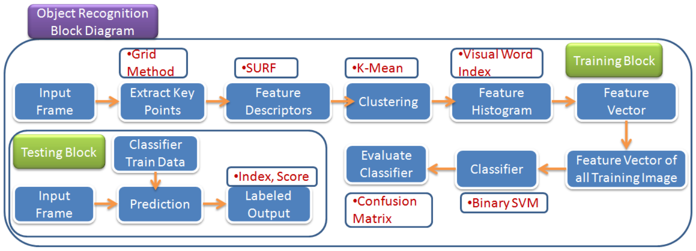

Fig. 4: Block diagram of Object Recognition Method

*1) **Clustering**::* Approximate Nearest Neighbor is used to cluster all the feature. Divide Complete features into K(500) visual vocabulary words. Number of clusters (K): 500. Number of features : M* N; Initialization the cluster centers termination criteria: 100 time loop or cluster distance error ¡ threshold. Feature Histogram: Generate No of word count present in each cluster per each image find how many word are present in 500 cluster. Each cluster represent as visual word index (500 visual index). Generate feature vector corresponding to each image

Training data : Repeat previous two slide for each image in the training set to create the training data Generate 101 feature histogram for car training image. similarly generate feature histogram all training image.

Feature Histogram: Generate No of word count present in each cluster per each image find how many word are present in 500 cluster. each cluster represent as visual word index(500 visual index).

Classifier: Encoded training images from each category are fed into a classifier training process. The function trains a multiclass classifier using the error-correcting output codes (ECOC) framework with the help of binary support vector machine (SVM) classifiers

V. Design and Implement of Algorithm

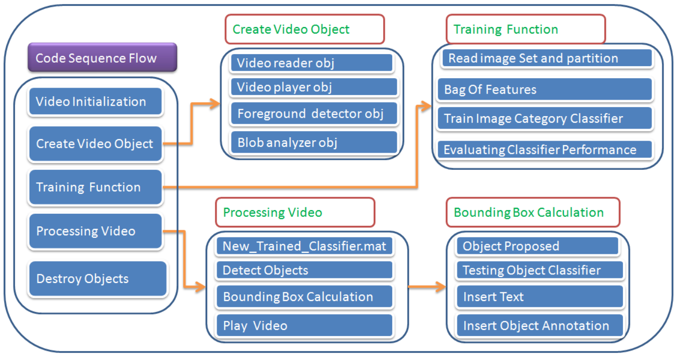

Fig. 5: Block diagram of Object Recognition Method

The steps involved in object detection and recognition system are as follows

- Initialization of video file.
- create video object which contains video read, player, foreground detector object and blob analyzer.
- Training function ,which will read image data and make partition, then apply Bag of word to classify image category.
- Process the video which will make Detect objects and bounding box calculation.
- Bounding box calculation function make object proposal followed by testing object classifier then insert object annotation.

First stage is video initialization .Then create video object by video reader object then video player object then foreground detection object and finally blob analyzer object. The next stage is training function. In these read image set and partition, Bag of features, train image category classifier, evaluating classifier performance blocks present. After training function next stage is processing video. This stage again divided into processing video and bounding box calculation. In processing video contains new trained classifier, detect objects, bounding box calculation and play video blocks present. In bounding box calculation object proposed ,testing object classifier, insert object annotation blocks are present. The final stage is destroy objects.

VI. Result

	carside	leopards	ant	butterfly	motorbikes	ferry	cup	elephant	panda	cub	car	tiger$_cartoon$	rhino	lion
carside	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
leopards	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ant	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
butterfly	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
motorbikes	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ferry	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cup	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0
elephant	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
panda	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0
cub	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0	0.0
car	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0
tiger_cartoon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0	0.0
rhino	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.0
lion	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00

TABLE I: Confusion Matrix of Training Data

	carside	leopards	ant	butterfly	motorbikes	ferry	cup	elephant	panda	cub	car	tiger$_cartoon$	rhino	lion
carside	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
leopards	0.0	1.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ant	0.05	0.0	0.05	0.41	0.0	0.05	0.0	0.05	0.0	0.0	0.0	0.0	0.0	0.0
butterfly	0.0	0.0	0.09	0.27	0.14	0.0	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0
motorbikes	0.0	0.0	0.0	0.0	0.95	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ferry	0.0	0.0	0.0	0.0	0.0	0.41	0.0	0.0	0.05	0.0	0.0	0.14	0.0	0.0
cup	0.0	0.0	0.0	0.05	0.0	0.0	0.32	0.0	0.0	0.0	0.0	0.14	0.0	0.0
elephant	0.0	0.0	0.0	0.05	0.0	0.0	0.0	0.23	0.0	0.23	0.0	0.0	0.05	0.0
panda	0.0	0.0	0.05	0.0	0.14	0.0	0.0	0.0	0.32	0.0	0.0	0.0	0.0	0.05
cub	0.0	0.0	0.14	0.0	0.05	0.0	0.0	0.0	0.0	0.41	0.0	0.0	0.0	0.05
car	0.0	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.05	0.0	0.68	0.0	0.05	0.0
tiger$_cartoon$	0.0	0.0	0.0	0.0	0.14	0.0	0.05	0.0	0.0	0.0	0.0	0.5	0.0	0.14
rhino	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.14	0.0	0.05	0.32	0.09
lion	0.0	0.05	0.05	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.0	0.09	0.0	0.77

TABLE II: Confusion Matrix of Validation data

VII. Conclusion

our purposed object recognition and localization algorithm is completely implemented using Matlab. we have GMM algorithm for object detection, BOW model for training and classifying object in the video frame. object proposal based on bounding box for location. we have demonstration our algorithm for different video and we got accuracy of object recognition 56%. we have trained caltelt data set and got training results. For our future implementation are planning to implement new object recognition technique and also we will use better object proposal technique purposed by Dallar. et.al. for fast and efficient algorithm.

References

[1] D. G. Lowe, *"Object recognition from local scale-invariant features,"International Conference on Computer Vision, Corfu, Greece, pp. 1150-1157,September,1999.*

[2] D. G. Lowe, *"Distinctive image features from scale-invariant keypoints," International journal of computer vision 60 (2), 91-110,2004.*

[3] M. Brown, D. G. Lowe, *"Recognising panoramas,"International Conference on Computer Vision,1218-1225,2003.*

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, *"Visual categorization with bags of keypoints,"Workshop on statistical learning in computer vision, ECCV 1 (1-22), 1-2,2004*

[5] F. Perronnin, C. Dance, G. Csurka, M. Bressan, *"Adapted vocabularies for generic visual categorization,"Computer VisionECCV-2006 464-475,2006*

[6] D. Nister and H. Stewenius, *"Scalable Recognition with a Vocabulary Tree,"Center for Visualization and Virtual Environments,Department of Computer Science, University of Kentucky 2006*

[7] G. Schindler. and M. Brown., *"A City-scale location recognition","CVPR, 2007*

[8] E. D. Eade. and T. W. Drummond., *"Unified loop closing and recovery for real time monocular SLAM (2008),"BMVC, 2008*

[9] N. Sebe, M. S. Lew, *“Comparing salient point detectors,”Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333CA, Leiden, Netherlands Received 15 August 2001.*

[10] D. Hall, B. Leibe and B. Schiele *“Saliency of Interest Points under Scale Changes”*.

[11] A. Torralba, F. F. Li and R. Fergus.,*“http://people.csail.mit.edu/torralba/cvpr2007/”*

[12] M. H. Yang, *“Object Recognition”,University of California, Merced.*

[13] *“http://www.caa.tuwien.ac.at/cvl/research-areas/object-recognition/”*.

[14] K. Kalirajan and M. Sudha , *“Moving Object Detection for Video Surveillance”,Research Article, Hindawi Publishing Corporation Scientific World Journal,Article ID 907469, 10 pages,Volume 2015.*

[15] Q. Dai, D. Hoiem., *“Learning to Localize Detected Objects”,Department of Computer Science University of Illinois at Urbana-Champaign,cvpr,2012.*

[16] M. B. Blaschko and C. H. Lampert., *“Learning to localize objects with structured output regression”,In ECCV, 2008.*

[17] D. Ramanan., *“Using segmentation to verify object hypotheses”,In CVPR, 2007.*

[18] Y. Li, N.h Snavely, and D. P. Huttenlocher, *“Location Recognition Using Prioritized Feature Matching,”Department of Computer Science, Cornell University, Ithaca, NY 14853,2007.*

[19] A. Farhadi, I. Endres, D. Hoiem, and D.A. Forsyth, *“Describing Objects by their Attributes”, CVPR 2009.*

[20] A. Coates, H. Lee, A. Y. Ng, *“An Analysis of Single Layer Networks in Unsupervised Feature Learning”, AISTATS, 2011.*

[21] Image dataset.,*“http://cs.stanford.edu/ acoates/stl10”*.