ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Object Detection Using Semantic Segmentation With Fisher Vector

Sudhin S[1], Jayalakshmi D.S[2], Veena G.S[3]

P.G Student, Computer-Science Department, M.S.Ramaiah Institute of Technology, Bangalore, India[1]

Associate Professor, Computer Science Department, M.S.Ramaiah Institute of Technology, Bangalore, India[2]

Assistant Professor, Computer Science Department, M.S.Ramaiah Institute of Technology, Bangalore, India[3]

Abstract: Image processing is an aspect of computer-science wherein one of the fundamental problems is that of object detection for a given set of images. There are number of existing works which are used for object detection using different methodology. Here in this work a method of segmentation using semantic rules along with feature extraction procedure is used. A separate training and testing phase is carried out wherein during training phase the SVM is trained for the images from the dataset and similarly in testing phase a test image is tested so as to determine the object of interest for given image. In order to detect the object more accurately semantic rules are applied both during training and testing stages

Keywords: Segmentation, Semantic rules, Feature Extraction, SVM

I. INTRODUCTION

Object detection is a computer vision problem, where in goal is to report both the location of the object in terms of a bounding box, and also the object category in an image. There is significant progress, which has been made earlier. Existing works on object detection using [1] wherein there was a need of feature set that allows the human form to be discriminated cleanly, even in cluttered backgrounds under difficult illumination. So a feature set for human detection known as Histogram of Oriented Gradient (HOG) descriptors were used which provided excellent performance relative to other existing feature sets. Another work using Discriminatively Trained Part Based Models [2] where the problem of detecting and localizing generic objects from categories such as people or cars in static images was considered. An object detection system was described that represented highly variable objects using mixtures of multi scale deformable part models. These models are trained using a discriminative procedure that only requires bounding boxes for the objects in a set of images.

But this work becomes computationally very expensive when rich representations are used. So to overcome this problem, work on real-time object detection was carried out on [3] where a face detection framework is described which is capable of processing images extremely rapidly while achieving high detection rates. This system achieves high frame rates working only with the information present in a single gray scale image. On Experimental evaluation this approach reduces the number of window to be observed and the face detector fails on significantly occluded face Similarly a work on Combining object localization and image classification was carried on [4] where a combined approach was presented for object localization and classification. For image classification the state-of-the-art approaches of the PASCAL VOC 2007 and 2008 challenges was relied upon and the existing sliding window approaches was improved and built for object localization. This approach evaluates a score function for all positions and scales in an image and detects local maxima of score function. The experimental results show that combined object localization and classification methods outperform the state-of-the-art on the PASCAL VOC 2007 and 2008 datasets.

Object detection could also be achieved by branch and bound technique rather than sliding window approach. A work was done based on Efficient Sub window Search [5] where a branch-and-bound scheme that allowed efficient maximization of a large class of classifier functions over all possible sub images was proposed rather than sliding window approach since it increased the computational cost. Here in this method it returned the object locations that an exhaustive sliding window approach would and at the same time it required fewer classifier evaluations than there were candidate regions in the image.

Similarly Measuring the objectness of image windows [6] wherein generic objectness measure is presented quantifying how likely it is for an image window to contain an object of any class and explicitly trained to distinguish objects with a well-defined boundary in space. This measure combines in a Bayesian framework several image cues measuring characteristics of objects, such as appearing different from their surroundings and having a closed boundary. In order to detect object using semantic segmentation a work [7] was carried in order to achieve this. Here a novel deformable part-based model was proposed which exploited region-based segmentation algorithms that compute candidate object regions by bottom-up clustering followed by ranking of those regions. In this approach every detection hypothesis allows to select a segment, and also scores each box in the image using both traditional HOG filters as well as a set of novel segmentation features. The effectiveness of this approach was demonstrated in PASCAL VOC 2010 dataset, and

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

also show that when employing only a root filter this approach outperforms Dalal & Triggs detector by 13% AP, and when employing parts, it outperforms the original deformable part based model by 8%.

Similarly another work wherein an approach to accurately localize detected objects was proposed [8]. The goal is to predict which features pertain to the object and define the object extent with segmentation or bounding box. The detector used is a slight modification of the DPM detector by Felzenszwalb et al. Several color models and edge cues for local predictions, were described and evaluated and also two approaches for localization was proposed, one is learned graph cut segmentation and other is structural bounding box prediction. Here, first the object was detected using a modified version of the deformable parts model (DPM) detector. Then, the pixels, which were part of the object based on color and edge information, were predicted. In order to determine the full extent of the object, the two approaches one is segmentation using graph cut on trained CRF potentials, and other is a structural learning approach to directly predict the bounding box were used. The experiments on the PASCAL VOC2010 dataset showed that this approach leads to accurate pixel assignment and large improvement in bounding box overlap, sometimes leading to large overall improvement in detection accuracy.

Another work was carried out for providing a unified framework for object detection, segmentation, and classification using regions [9]. The Region features are appealing since they encode shape and scale information of objects naturally and they are only mildly affected by background clutter. Here a robust bag of overlaid regions for each image using Arbel ´aez et al., CVPR 2009 was produced. The idea was, first each image were represented by a bag of regions derived from a region tree. These regions were described by a rich set of cues inside them. For these region weights were learned using a discriminative max-margin framework and then a generalized Hough voting scheme was applied to cast hypotheses of object locations, scales, and support, followed by a refinement stage on these hypotheses which deals with detection and segmentation separately. This approach significantly outperformed on the ETHZ shape database and achieved competitive performance on the Caltech 101 database. In our work semantic rules along with feature extraction procedure have been used in order to detect the object in an image more accurately.

II. PROPOSED SYSTEM

In the proposed system there are two stages

- Training
- Testing

In training stage the classifier is trained for a set of images. Initially the image from the database is taken and a pre-processing is done for this image, which involves converting the image to gray scale and resizing the image. The processed image is then fed into feature extraction step which involves SIFT and color feature block in order to detect the key points involved in the image and to extract the color histogram values from the image. The results from these two blocks are combined to form a fisher vector representation of the image. This representation is then added into SVM train along with the semantic rules wherein the results obtained are stored in knowledge base for further retrieval. In testing phase a test image is taken and it involves a similar process that is done in training stage. Here the image is processed and sift and color descriptors are used which results in fisher vector representation. This representation is added to the SVM classifier along with the information that is stored in the knowledge base is retrieved along with semantic rules to form a segmented object of the image.

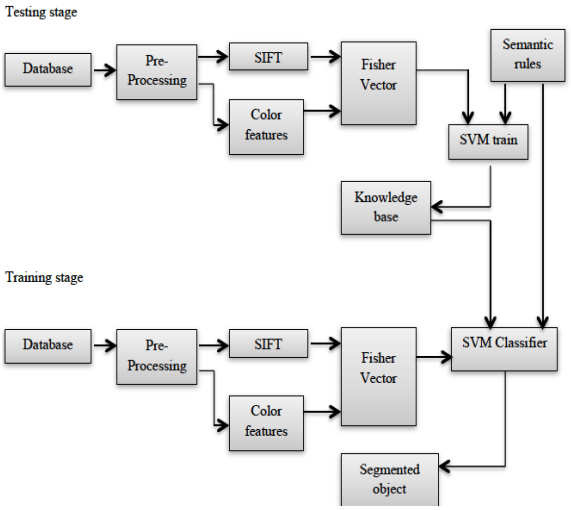

Fig 1:Architectural Diagram

A. *Pre-Processing*

Pre-processing of images involves removing low frequency background noise, normalizing the intensity of individual particles images, reflections and masking portions of images. This can increase the reliability of an optical inspection. Several filter operations, which intensify or reduce certain image details enables an easier or faster evaluation. So one such operation, which is used is Histogram Of Oriented Gradients

B. *Scale Invariant Feature Transform*

Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in images. For any object in an image, interesting points on the object can be extracted to provide a feature description of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. Such points usually lie on high contrast regions of the image, such as object edges and another important characteristic of these features is that the relative positions between them in the original scene shouldn't change from one image to another.

SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes. SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches

C. Fisher Vector

Fisher Vector is an image representation obtained by pooling local image features. It is frequently used as a global image descriptor in visual classification. The Fisher Vector (FV) representation of images can be seen as an extension of the popular bag-of-visual word (BOV). Both of them are based on an intermediate representation, the visual vocabulary built in the low level feature space. The FV representation has many advantages with respect to the BOV. First, it provides a more general way to define a kernel from a generative process of the data, since the BOV is a particular case of the FV where the gradient computation is restricted to the mixture weight parameters of the Gaussian Mixture Model (GMM), it is shown experimentally that the additional gradients incorporated in the FV bring large improvements in terms of accuracy. The second advantage of the FV is that it can be computed from much smaller vocabularies and therefore at a lower computational cost and the third advantage of the FV is that it performs well even with simple linear classifiers.

Let X = $\{x_t$, t = 1...T$\}$ be the set of T local descriptors extracted from an image. Assume that the generation process of X can be modeled by a probability density function$u_\lambda$with parameters λ. X can be described by the gradient vector:

$$G_\lambda^x = \frac{1}{T}\nabla_\lambda \log_{u_\lambda}(X) \quad (1)$$

The gradient of the log-likelihood describes the contribution of the parameters to the generation process. The dimensionality of this vector depends only on the number of parameters in λ, not on the number of patches T.A natural kernel on these gradients is:

$$K(X,Y) = G_\lambda^{X'} F_\lambda^{-1} G_\lambda^Y \quad (2)$$

WhereF_λ is the Fisher information matrixu_λ:

$$F_\lambda = E_{x\sim u_\lambda}[\nabla_\lambda \log u_\lambda\,(x)\nabla_\lambda \log \mathrm{u}_\lambda(\mathrm{x})'] \quad (3)$$

As $F_\lambda$is symmetric and positive definite, it has a Cholesky decomposition $F_\lambda = L_\lambda' L_\lambda$ and (X, Y) can be rewritten as a dot-product between normalized vectors G_λ with:

$$G_\lambda^x = L_\lambda G_\lambda^x \quad (4)$$

HereG_λ^x is referred as the Fisher vector of X

D. SVM Classifier

Support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class since in general the larger the margin the lower the generalization error of the classifier. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function K(x,y) selected to suit the problem.

E. Semantic Rules

For given images from the datasets, the images are divided into patches and for each of these patches ground truth clustering is carried out so that for each of these patches there is clear distinction between the colors used which helps in differentiating between the object of interest and other objects in the image. Once there is a clear distinction between the colors used, for each of the colors sift and color features are extracted which results in the segmented object of interest

III. RESULTS

The below figures are some of the images which have been tested for this system. The result obtained suggests that the system works accurately.

Fig 2:Original image

Candidate window

Detecting regions in the image

Fig3:Original image

Candidate window

Detecting regions in the image

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014


Fig 4: Original image Candidate window

Detecting regions in the image

Fig5: Original image Candidate window

Detecting regions in the image

In this work carried out there are two stages one is training and other is testing. In training stages the images are divided into patches and each of these patches are passed through pre-processing and feature extraction procedure, which involves sift and color descriptors. The result of these stages is combined to form a fisher vector representation. Semantic rules are applied to SVM, which are used to differentiate the object region with other regions along with fisher vector representation, and the result obtained is stored in knowledge base for future access. Similarly in testing stage a test image is taken and it undergoes a series of pre-processing and feature extraction process..Then the resultant fisher vector representation is passed on to the SVM classifier along with semantic rules and information stored in the knowledge base. The result obtained is the object of interest along with the labels associated with the regions in the image. In the below Table 1 the values is determined for the images mentioned above. Here the values are interpreted by taking true positive, true negative, false positive and false negative values corresponding to human and non-human regions. True positive values correspond to the human regions and false negative values correspond to non-human regions in these human region. Similarly true negative values correspond to non-human region and false positive values correspond to human regions in these non-human region.

List Of Figures	True Positive	True Negative	False Positive	False Negative
Fig 2	13	19	0	4
Fig 3	25	9	1	1
Fig 4	12	24	0	0
Fig 5	12	19	2	3

Table 1:Result set for the images taken above

IV. CONCLUSION

The framework was presented for object detection, segmentation, and classification using regions by applying histogram of gradient orientations features together with SIFT and color descriptors on a image gives good results for human detection along with fisher vector representation and semantic rules which helps in labelling the human region. This method makes use of INRIA person dataset along with some selective search strategy in order to train and test the Support Vector Machine detector and derive very efficient features, which can capture the essential information encoded in the image segments. The results obtained using this method on INRIA dataset were accurate in detecting the region of interest along with labeling the regions as human and non-human in the image. In the future work there is a need to apply the semantic rules on non-human regions of the image so that there is complete information regarding the objects in the image

REFERENCES

[1] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection. In CVPR, 2005
[2] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan, Object Detection with Discriminatively Trained Part Based Models
[3] P. Viola and M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, May2004.
[4] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In ICCV, 2009.
[5] C. Lampert, M. Blaschko, and T. Hofmann. Efficient sub window search: a branch and bound framework for object localization. PAMI, 31(12):2129–2142, 2009.
[6] B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image windows. PAMI, 34(11):2189–2202, 2012.
[7] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In CVPR, 2013.
[8] Q. Dai and D. Hoiem. Learning to localize detected objects. In CVPR, 2012
[9] C. Gu, J. Lim, P. Arbel´aez, and J. Malik. Recognition using regions. In CVPR, 2009