International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Recognition of Human Emotion Detection and Annotation, using Local Descriptor and Support Vector Machine

Shama P S, Pattan Prakash

Abstract: ***In multimedia data analysis, video tagging is the most challenging and active research area. In which finding or detecting the object with the dynamic environment is most challenging. Object detection and its validation are an essential functional step in video annotation. Considering the above challenges, the proposed system designed to presents the people detection module from a complex background. Detected persons are validated for further annotation process. Using publically available dataset for module design, Viola-Jones object detection algorithm is used for person detection. Support Vector Machine (SVM) authenticate the detected object/person based on it local features using Local Binary Pattern (LBP). The performance of the proposed system presents given architecture is effectively annotating the detected people emotion.***

Keywords: *People Detection, Segmentation, Recognition and Emotion Annotation.*

I. INTRODUCTION

Video annotation is also termed as "Video Tagging". Analysis of video and scenes explanation natural language is easy for humans but for machines it becomes difficult. Increase in computer vision technology makes the machine to identify the object/visual content up to an extent. This extension not yet meets the automatic description of video scenes [01]. The ordinary user easily gets large-scale video data due to the rapid advance in internet technology, network, and data compression techniques. This makes content-based data retrieval is an active research area. This technique is also termed as "Semantic Gap" i.e. discontinuity between both low and high-level features. According to the recent survey, it is revealed that annotation is a promising approach to feel the semantic gap. Surveillance management, education system, entertainment sector, media, medical application area presents a lot of video in day by day. Hence as an increase in video size, it becomes difficult to find out the desired data. In difficulty get solved by video annotation i.e., for example, considers a semantic query is to search a red color book which is passing from person to person in a video.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Shama P S*, Computer Science & Engineering, PDACE, Kalaburagi,Karnataka, India. shm.san1@gmail.com
Dr.Pattan Prakash, Computer Science & Engineering, PDACE, Kalaburagi,Karnataka, India

Detecting an object is easy, where annotating an object with text increase system performance level. Video annotation can be performed on complete video, or individual frame/key frame for particular scenes, a group of peoples.

Advance in computer vision techniques will always help in enhancing the system performance level. But still, annotation result is inaccurate. The challenges involved in video annotation is listed as requirement of large training data to meet the reasonable annotation performance. It is necessary to have sample prototypes with training data to feel the gap between low-level features and high-level features. Training the machine learning algorithm with larger training data with a sample prototype is usually the most challenging. Variety of semantic concepts can be described only when there a large number of low-level features. But working on high dimensional features degrades the system performance.

1. Di stance-based feature extraction is most challenging. Finding various concepts needs different low-level features, but obtaining low-level features with different best distance is most challenging.
2. In video data, temporal features provide most minor and essential features about object/scene etc. But in video annotation, these temporal features are getting neglected. This elimination leads to a loss in some semantic information during video annotation [02].
3. To tackle the video annotation challenges a set of the algorithm has been designed and proposed in the computer. Supervised learning techniques, object segmentation modules, key feature extraction algorithm to overcome the data insufficiency problem. To avoid the dimensionality, curse a multimodal fusion is used.

Video may include a set of different objects, persons. There are different scenarios or applications where it is necessary to annotate the objects sometimes even persons. The proposed module presents effective people detection and their individual emotion annotation approach. A brief summary of the video annotation techniques and its work is summarised in Section II. Considering the annotation challenges from a survey, the designed modules and its functional block explanation are depicted in Section III. Section IV summarise the intermediate results of the proposed module.

II. LITERATURE SURVEY

Vasileios Mygdalis et.al [03] has presents a support vector-based video summarization model. A hierarchically based learning procedure is used for video segmentation and summarization.

This learning process includes two functional operations i.e. unsupervised learning for a video segment and supervised learning techniques for data summarization. The experiment is conducted on three Hollywood movies by using both hierarchical learning and Subclass SVDD algorithm. From the result section, it is concluded that the designed model performs the acceptable classification performance.

Muhammad Usman Ghani Khan et.al [01] has proposed textual description-based video annotation framework. Research work is tested on the self-developed dataset, where the entire video is getting segmented in 10 to 20s lengths. As the model is basically concentrated on human and his/her activity since it is necessary to have clear resolutions. Around 11 human activities are analyzed by using the textual descriptions. Human visual features get extracted by using conventional feature extraction algorithm, whereas nonhuman features get extracted by using Haar feature extraction techniques. The boosted classifier is used for effective annotation techniques. The referred modules perform the task-based video annotation. From the result section, the system designer concluded that system performance is effective but the most challenging task is to get the effective features of human and nonhuman regions.

Mohak Sukhawani et.al [04] has designed a content-based video annotation module for Tennis video. A brief text description of the respective video segment is used for frame-level annotation. The model is worked on London Olympic 2012 tennis video. Fine gained techniques are used for video annotation. The annotation function is segmented into two sections i.e. dictionary-based system learning and K-SVD algorithm. Integration of both techniques presents the proper textual labeling object in each frame. From the result analysis, it is concluded that designed fine-grained application is useful for video retrieval as well as video summarization.

Shangan Sah et.al [05] proposed text-based video annotation techniques on user-generated video. Based on user preference and cinematographic long video get segmented in a smaller length. The most impact full frame is considered as keyframe for annotation. Using neural network and descript text data input is got annotated. Based on the system performance it is concluded that designed automated system is good for long video.

Yu Liu et.al [06] has designed advanced multimodal textual based video annotations. Initially, both text and visual features are extracted and embedded in matching components. A deep learning-based classification is combined both the feature level for better performance. The preferred architecture is trained in multilevel, which decreases the classification losses. Based on the result it is concluded that the designed approach presents effective output for multimodal feature matching and data classification (i.e. in video annotation).

III. METHODOLOGY

There are so many kinds of object annotation or video tagging in video processing. Annotation can be done based on the object name, its characteristics or based on its appearance throughout the video. It can also be done by using other multi module technique like text. Considering these different video annotation methods and video annotation challenges (i.e. object segmentation, recognition) a simple model is designed which is shown in below Figure 1. Training and testing are the two main functional section of the implemented design. A set of sample videos are trained and knowledge base is created using supervised machine learning algorithms.

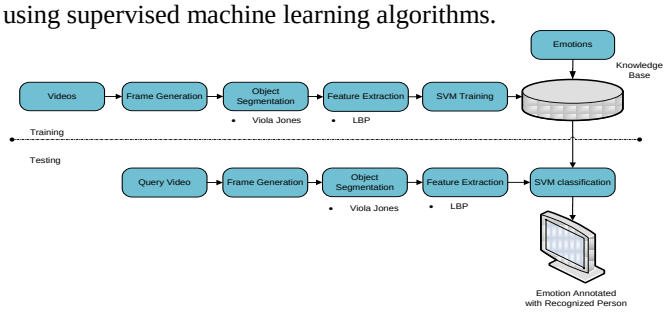


Figure 1: Proposed Functional Block Diagram

In testing, input query video is converted into a number of frame sequences. Every frame proceeds further for object segmentation, in which person is segmented or detected by using viola jones object detection algorithm. Further detected object region is passed to a descriptor module to extract the local pattern/features of a segmented frame region. These segmented object features are stored in a separate file format which can be loaded to SVM classifier during object recognition. The mathematical working and individual response of each function unit are briefly depicted in below section.

3A. Input Video/Frame Generation

Input video is converted into a number of frames with 15 frames/sec frame rate. Each converted frame is in RGB plane. These color frames passed to next object detection module to segment the object region.

3B.Object Detection using Viola-Jones Algorithm

To detect the object/person each frame is processed sequentially using viola jones algorithm. The Viola-Jones face detector contains three main ideas that make it possible to build a successful face detector that can run in real time: the image integral, classifier learning with AdaBoost, and the attentional cascade structure.

i) Image integral and feature extraction

The first step of the Viola-Jones face detection algorithm is to turn the input image into an integral image. Integral image, also known as a summed area table, is an algorithm for quickly and efficiently computing the sum of values in a rectangle subset of a pixel grid. The integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive:

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x',y') \qquad (1)$$

where $i(x,y)$ is the pixel value of the original image and

1	1	1
1	1	1
1	1	1

1	2	3
2	4	6
3	6	9

(a) Input Image (b) Integral Image

Figure 2: Integral Image Example

$i(x', y')$ is the corresponding image integral value. Using the integral image to compute the sum of any rectangular area is extremely efficient, as shown in Figure 2. The sum of pixels in rectangle ABCD can be calculated with only four values from the integral image:

$$\sum_{(x,y)\in ABCD} i(x,y) = ii(D) + ii(A) - ii(B) - ii(c) \quad (2)$$

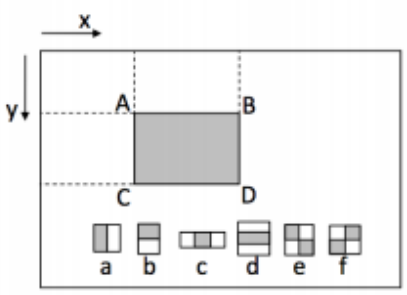


Figure 3: Integrated Image with 6 Different Haar Features

ii)AdaBoost Learning

Given a feature set and a training set of positive and negative images, any number of machine learning approaches could be used to learn a classification function. The ViolaJones uses a variant of AdaBoost to both select a small set of features and trains the classifier. A single AdaBoost classifier consists of a weighted sum of many weak classifiers, where each weak classifier is a threshold on a single Haar-like rectangular feature. The weight associated with a given sample is adjusted based on whether or not the weak classifier correctly classifies the sample. A single weak classifier is defined as:

$$h(x, f, p, \theta) = \begin{cases} 1 & pf(x) < p\theta \\ 0 & otherwise \end{cases} \quad (3)$$

Where f denotes the feature value, θ is the threshold and p is the polarity indicating the direction of the inequality

iii)Cascade Classifier

The cascaded classifier is made out of stages each containing a solid classifier from AdaBoost. The activity of each stage is to decide if a given sub-window is unquestionably not a face or perhaps a face. At the point when a sub-window is grouped to be a non-face by a given stage it is quickly disposed of. On the other hand, a sub-window delegated a possibly face is given to the following stage in the course. It pursues that the more stages a given sub-window pass, the higher the possibility the sub-window contains a face [07][08][09]..

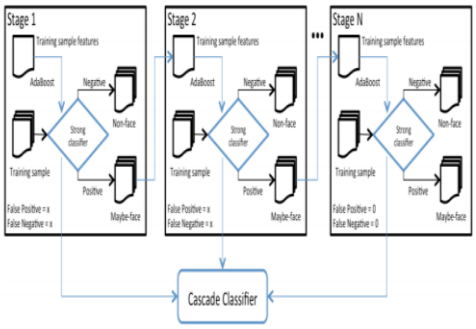


Figure 4: Work Flow of Cascade Classifier

3C. Local Feature Extraction using LBP

The segmented object region is acted as input to the LBP algorithm. The application of LBP finds out the local pattern of segmented regions. Before extracting the features, the segmented color region of the frame is translated into a grayscale region. In which pixel intensities are varying from 0 to 255. In grayscale considering 3x3 gray scale regions or 8 neighbors of the center pixel to extract the LBP features. The working of LBP is explained in Algorithm 1with the help of Figure 2[10][11].

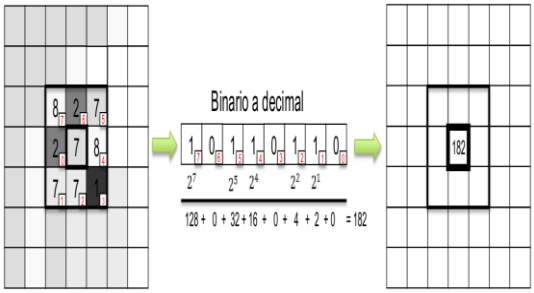


Figure 5: Working of Local Binary Pattern Algorithm

Algorithm 1: Local Binary Pattern

Input : Segmented RGB Image

Output: LBP Feature Vector

Step.1: Convert the RGB Image region into grayscale using Eq. (1)

$$Gray\ Scale = 0.30R + 0.59G + 0.11G$$

Step.2: Find out the size of the grayscale image

Step.3: Define the block size

Step.4: Extract the coordinates the blocks

Step.5: Initialize the result matrix with zeros

Step.6: Point the center pixel of the block

Step.7: for i = 1:no.of block size

```
a = neighbourPixel(i);
c = centre pixel;
if a>c
  a = 1;
else
  a = 0;
end
```

Step.8: Generate the 8-bit binary bit stream

Step.9: Compute the decimal equivalent values of 8-bit binary stream

Step.10: Replace the center pixel with a computed decimal value

Step.11: Repeat the procedure for the entire grayscale image.

Step.12: Form the feature vector

End Algorithm

The computed features from LBP are loaded in TestFeatures.dat file. This file is passed to the classifier for object authentication. A supervised learning based machine learning algorithm is used for object recognition.

3D.Object Authentication and Video Annotation

As in mentioned in training section SVM classifier is trained object features and annotation parameters of each object/person. These features create a knowledge base. The respective segmented object local features are loaded to SVM classifier for object recognition. The working of SVM in linear classification is depicted in below section.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional spaces, this hyperplane is a line dividing a plane into two parts wherein each class lay in either side. The pictorial representation of the class separation is shown in below Figure 6.

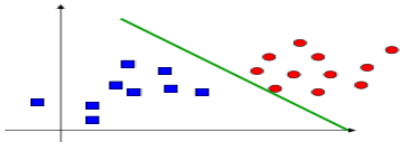


Figure 6: Linear Data Separation by using SVM classifier.

The green line in above image is termed as hyperplane and its mathematical representation is given in Eq. (5)

$$w^T x + w_0 = 0 \quad (5)$$

If testing features region is greater than hyperplane than SVM label it as class 1 object else class 2. The mathematical terms involved in data classification is given in below Eq. (6) and (7)[12].

$$Class\ 1\ if\ w^T x + w_0 > 0 \quad (6)$$

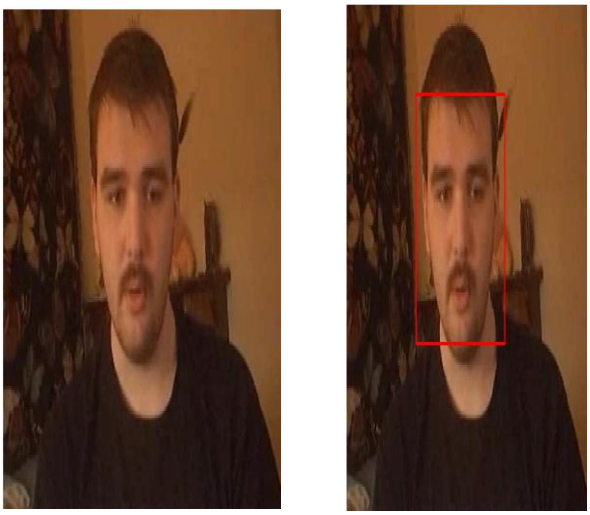
$$Class - 1\ if\ w^T x + w_0 < 0 \quad (7)$$

The application of SVM can effectively recognize the segmented object and annotate the according to trained sequences. Experimental result briefly summarises the intermediate result of the proposed design and overall function flow of the implemented architecture is shown depicted in Figure 7.

Figure 7: Function Flow Chart of Proposed Design

IV. EXPERIMENTAL RESULT

Due to the unavailabity of the standard dataset for video annotation, a simple module is generated in which person are detected by using advanced segmented method. This segmented person can be annotate in two different way i.e. either annotating the person identity or by his/her facial expression. In designed module, the segmented person is annotated with his/her name with their emotions. There are two types emotion considered i.e. emotion 1 indicated the normal behaviour of the person whereas emotion 2 indicates abnormal/stressed behaviour of the segmented person.



(a) (b)

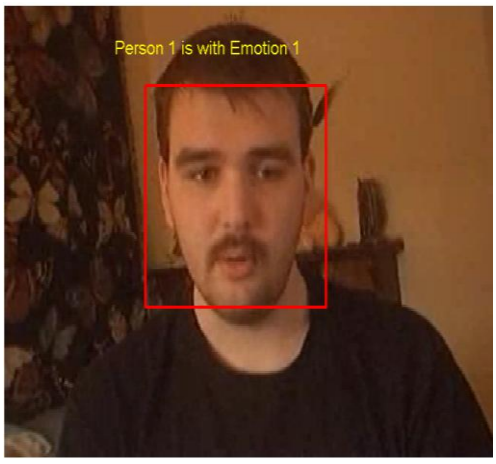
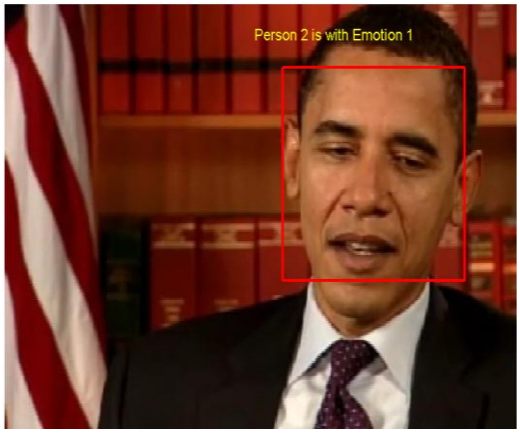


Figure 8: Performance Analysis of Input Video 1; (a) Input Frame ;(b) Object/Person Detected Frame ;(c)Annotated Video Frame;

Frame 1

(a)

(b)



(c)

Frame 2

(a)

(b)

(c)

Figure 9: Performance Analysis of Input Video 2; (a) Input Frame ;(b) Object/Person Detected Frame ;(c)Annotated Video Frame;

The performance is tested with publically available standard videos. Figure 8 represents the intermediate results of Input Video 1. Figure (a) denotes the input video frame, whereas a person detected in an input frame is shown in Figure (b). An authenticated person with annotated information is depicted in Figure (c). The similar testing procedure is repeated for other videos, and its intermediate outputs are shown in Figure 9. From the above performance results, it is cleared that, application of MATLAB, computer vision algorithms, and machine learning modules can effectively segment the people even in complex backgrounds. In the future, the work is extended by using the video with a number of background objects, where we can effectively segment the object and annotate them based on a given input text.

V. CONCLUSION

The proposed model is designed with an optimized objective to segment the object and annotate the object its characteristics. The applied segmentation module effectively segmented the person. Further used SVM module effectively authenticated the person and annotate its emotions based on the created knowledge base. From the result section it is concluded that proposed architecture meet the acceptable accuracy in object segmentation and its facial emotional annotation. Further, in futureproposed work is extended by considering the video with multiple background objects and its textual description. Out of the set of objects, a specified object is detected and recognized by using supervised machine learning.

REFERENCES

1. Muhammad Usman Ghani Khan, Nouf Al Harbi and Yoshihiko Gotoh, “A Framework for Creating Natural Language Descriptions of Video Streams”, Information Sciences, Elsevier, Vol. 303, pp.61-82, 2015.
2. Khushboo Khurana, and M. B. Chandak, “Study of Various Video Annotation Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp.909 - 914, 2013.

3. Vasileios Mygdalis, Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas, "Video Summarization Based on Subclass Support Vector Data Description", In Computational Intelligence for Engineering Solutions (CIES), 2014 IEEE Symposium on, IEEE, pp. 183-187.2014.
4. Mohak Sukhwani and C.V Jawahar, "Frame Level Annotation for Tennis Videos", 23rd International Conference on Pattern Recognition (ICPR), IEEE, pp. 4-8, 2016.
5. Shagan Sah, Sourabh Kulhae, Allison Gary, Subhashini Venugopalan, Emily Prud hommeaux, and Raymond Ptucha, "Semantic Text Summarization of Long Videos", IEEE Winter Conference on Application of Computer Vision, IEEE, pp. 989-997, 2017.
6. Yu Liu, Li Liu, Yanming Guo, Michael S Lew, "Learning Visual and Textual Representations for Multimodal Matching and Classification", Pattern Recognition, Elsevier, 2018.
7. Ole Helvig Jensen, "Implementing the Viola Jones Face Detection Algorithm", Kongens Lyngby 2008.
8. Yi-Qing Wang, "An analysis of the Viola-Jones face detection algorithm", Image Processing On Line, pp. 128-148, 2014.
9. Laxmi Narayan Soni, Ashutosh Datar and Shilpa Datar, "Implementation of Viola Jones Algorithm Based Approach for Human Face Detection", International Journal of Current Engineering and Technology, pp. 2347-5161, Vol. 07, No. 5, 2015.
10. Esa Prakasa, "Texture Feature Extraction by using Local Binary Pattern", INKOM, Vol. 9, No. 2, pp. 45-48, 2015.
11. Di Huang, Caifeng Shan, Mohsen Ardevilian, Yunhong Wang and Liming Chen, "Local Binary Patterns and its Application to Facial Image Analysis: a Survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 41, Issue 6, pp.765-781, 2011.
12. Shuai Zheng, and Chris Ding, "Minimal Support Vector Machine", arXiv preprint arXiv, pp.1804-02370, 2018.

AUTHORS PROFILE

Shama P S did her B.E (ISE), M. Tech (CSE) from PDACE, Kalaburagi. Currently she's pursuing Ph. D from VTU, Belagavi. She has published various papers in national and international journals. Her area of interest is Digital Image Processing.

Dr. Pattan Prakash, presently working as Professor in Dept. of Computer Science and Engineering, PDACE**,** Kalaburagi. Currently 08 Ph. D Scholars are working under his Supervision. He published 11 papers in international Journals and 06 papers presented in International Conferences.