HUMAN ACTIVITY RECOGNITION

Deepthi*

Student, Department of MCA, NMAM Institute of Technology, Karkala, India

*Corresponding author: nayakdeepthi20@gmail.com

Abstract *- The computer vision tasks of pose estimation, object recognition, image retrieval, video action recognition, and frame tagging in videos are all intricately related to the task of action detection in still images. The objective of this challenge is to recognise a person's behaviour or conduct from just one frame. In contrast to action detection in movies, a comparably highly established area of research where they are utilised, spatio-temporal features are not available for still images, which makes the problem more challenging. The current work only takes into account actions involving objects. Semantics allows for the breakdown of complex actions into their component parts. The contribution of each of these elements to action recognition is thoroughly examined.*

Keywords *– Faster RCNN, image processing , object detection, tensorflow, deep learning*

---

i. Introduction

It is quite well recognised and well-researched in the field of still-based image recognition. It has gotten a lot of attention because the volume of photos shared on social networks has increased recently. Since motion cannot be easily approximated from a still image and since spatiotemporal information cannot be used to characterise the action, activity detection in still images is still a challenging problem to address. Even if it is easier and more rational to spot activity in pictures, it is still possible and very helpful to spot it in still photos. A single image (without movement or video signal) can effectively depict a variety of action kinds, and these actions can be correctly understood based on human sight. To accurately describe actions for these action kinds, one frame is sufficient. This information lends credence to the development of computer methods for automatic activity analysis and recognition in still images. Loss of spatiotemporal characteristics, backdrop clutter, high intra-class variance and low inter-class variance among specific action classes, change in background lighting, and variation in human stance are the key obstacles to action recognition in still photos. The most crucial feature for describing actions in videos is their spatiotemporal characteristics. Images lose the temporal information, which makes it far more difficult to describe an activity.

To accurately describe actions for these action kinds, one frame is sufficient. This information lends credence to the development of computer methods for automatic activity analysis and recognition in still

images. Loss of spatiotemporal characteristics, backdrop clutter, high intra-class variance and low inter-class variance among specific action classes, change in background lighting, and variation in human stance are the key obstacles to action recognition in still photos. The most crucial feature for describing actions in videos is their spatiotemporal characteristics. Images lose the temporal information, which makes it far more difficult to describe an activity. On the other hand, action recognition uses object recognition as a first step. When performing still image-based action identification, off-the-shelf object detectors are frequently used to obtain the class labels of all the items visible in the image. Their co-occurrence is then modelled and used as a feature. Still picture action detection is closely related to other computer vision tasks like pose estimation and scene interpretation. Pose estimation of the subject and the setting of the action have been used as inputs to action recognition in numerous articles. With some precision, pose and scene information are used to train the action recognition model. A more precise pose estimate and scene understanding model is then created using the trained still picture action recognition model as an input. This creates a cycle where one job is used to enhance the effectiveness of another.

Only activities that involve a person manipulating an object are taken into account in this work. It is intended to deconstruct an action into more manageable semantic components and to comprehend the significance of each of these components in action recognition.

ii. Literature survey

- The human body, body parts, action-related objects, human object interaction, and the complete situation or environment are some of the most frequent high-level indicators used in human action identification in still photographs. These metrics describe a variety of behavioural characteristics [1]
- The overall coarse contour of the human body in the photograph was taken advantage of by Wang et al. A group of edge points acquired using a clever edge detector were used to depict the shape [3]. The shape is utilised as a feature to group and categorise photographs into several tasks.
- Recognizing actions also requires being able to read body language. In order to create deformable models utilising the Conditional Random Field, Ikizler et al. [4] employed the body positions that were derived from photos using edge and region data (CRF).
- In order to find the useful, discriminative patches from the human body region for action recognition, Yao et al. [5] employed a variant of random forests. A saliency map is another visual representation of critical patch information [6].
- A part-based model made up of objects and human poses was employed by Yao et al. [7]. A bike is ridden by a person, for example, and the linked things are either person- or scene-related (for example, the grass in the scene of "horse riding in grassland"). The characteristics are language descriptions of human

behaviour. Objects and human poses make up the parts. Action bases for modelling actions in still photos are attributes and components [8].

- In [9], Le et al. deconstruct input photos into recognisable objects before using a language model to list all conceivable actions when the objects are used in various configurations. While some approaches integrate scene information with the object information for action recognition, others model the co-occurrence of objects to describe actions.Desai et al. [10] exploited contextual information, such as item arrangement discovered by their discriminative models [11], for action recognition.

iii. Methodologies

Since the current investigation is exclusively interested in behaviours involving objects, a unique dataset with five classes is made. Different sources were used to select the images for this dataset. When there was only one person doing an action with an object, photos for a few action classes were directly selected from datasets like the Stanford 40 Action dataset [12] and the Willow dataset [13] Other pictures came from the Google search engine. The unique dataset contains 200 photos for five different motion types, such as standing, sitting, sleeping, climbing stairs, and descending.

Fig1: custom data-set collection

The data sets are pre-processed and normalised in fig 2 before being labelled with the "labelImg" tool in accordance with their categorization. Following feature extraction, the images go through training where they are classified based on their features. Testing is conducted after training is finished, and the output is then shown using the anaconda prompt.

Google Colab and Anaconda IDE were employed in the development of this system. One of the best IDEs for creating systems like these is Google Colab's Anaconda. To write and run code in a single window, use this tool. Python has been used to create this project. Python has been utilised because it is a fairly adaptable language with a wide range of library support, which makes programming easier.

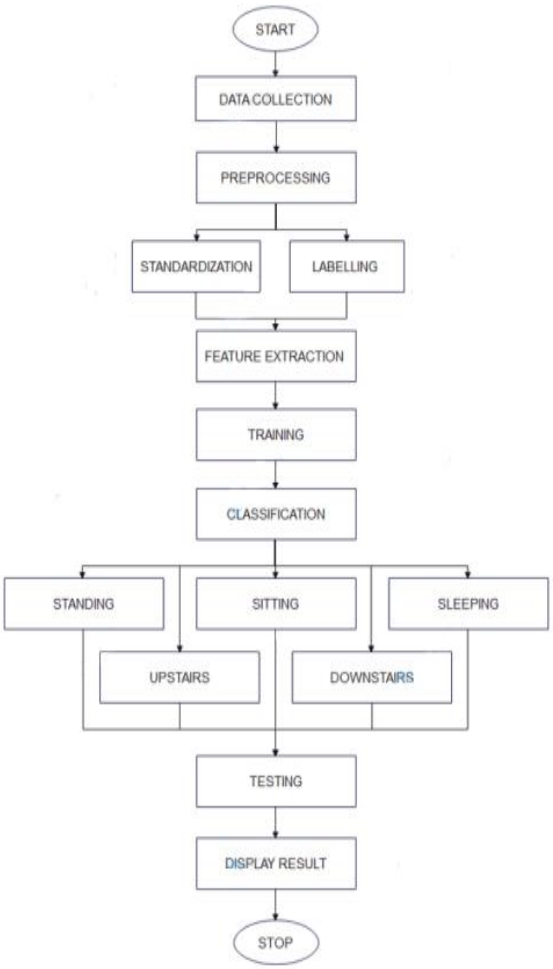

Fig 2: flow diagram

iv. Proposed solution

Working of RCNN:

Step 1: After running a region proposal method similar to a selective search on the input image, we receive 2,000 candidate region proposals in the image that need to be evaluated.

Step 2: We will warp each potential region into a set size, say, because region suggestions can have varying sizes and aspect ratios (224x224)

Step3: We will individually process each warped image region via a Convolutional Neural Network (CNN), and the CNN will produce a classification score for each of these regions.

We contrast these bounding boxes using the Intersection over Union metric (IOU).

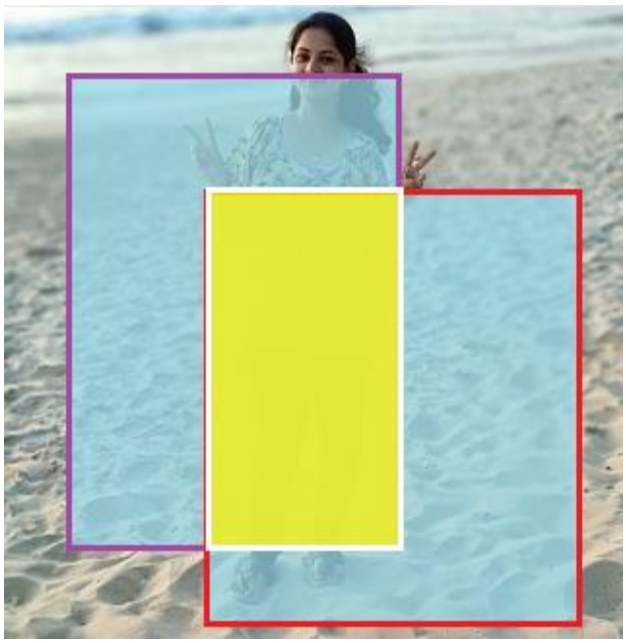

Fig3: Intersection over union

Intersection over union(IOU)
= size of ▭ / size of ▭

If IOU >= 0.5, it is "correct."

Area of Intersection = IOU (Area of Union)

The measure of overlap between these two bounding boxes in Fig. 3 is called IOU.

IOU < 0.5 is considered "BAD"

IOU > 0.5, we deem it deplorable.

IOU>0.7 is considered "GOOD"

IOU > 0.9 is considered "ALMOST PERFECT."

v. Output

The evaluation of each combination of component attributes for classifying actions is shown in the table below. These classifications are reported to be accurate.

Fig4 . Downstairs detection

Fig5 . Sitting detection

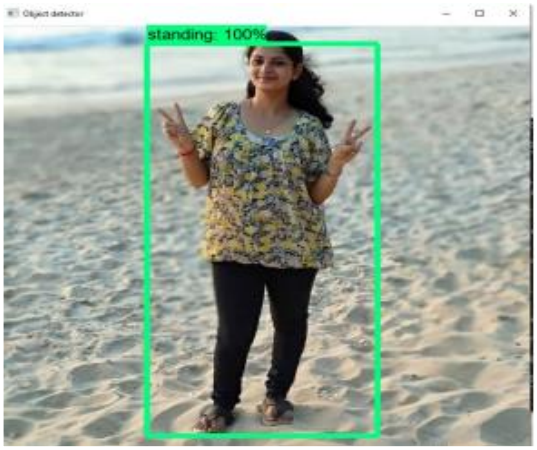

Fig6 . Standing detection

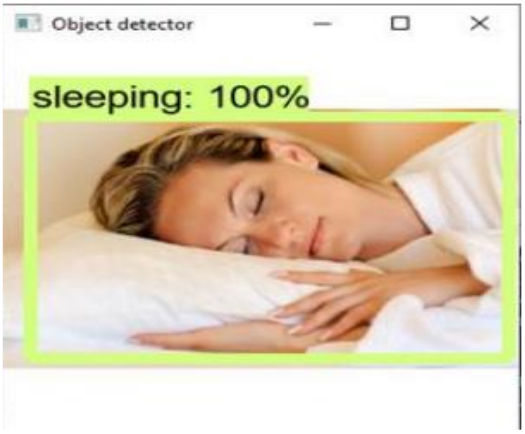

Fig7 . Sleeping detection

Fig8 . Upstairs detection

vi. Conclusion and Future scope

Recognition of human activity has numerous uses in human survey systems and medical research. We created a recognition system for five human activities in this project: standing, sitting, sleeping, going upstairs, and going downstairs. Mobile devices and other social media platforms are used to gather the information or photographs. They are then trained to produce reliable findings after being pre-processed, labelled, and classified according to their category. Here, a faster version of RCNN has been employed as the algorithm.

Future implementation may include the following:

- Putting in place a real-time system and employing mobile devices

References

1. Anca Ralescu and Koji Miyajima. The representation and identification of basic spatial relations in 2D segmented pictures. 225–236 Fuzzy Sets and Systems, 65(2-3):1994,
2. 2. Greg Mori, Yang Wang, Hao Jiang, Mark S. Drew, and Ze-Nian Li. Action class discovery without supervision. Volume 2, pages 1654–1661 of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006.
3. John Canny, an edge detection method using computing. IEEE Transactions on pattern analysis and machine intelligence, vol. 6, no. 6, 1986, pp. 679–698.
4. Pinar Duygulu, Selen Pehlivan, R. Gokberk Cinbis, and Nazli Ikizler. identifying actions in still pictures. Pages 1-4 of the 2008 19th International Conference on Pattern Recognition. IEEE, 2008.
5. Li Fei-Fei, Aditya Khosla, and Bangpeng Yao. utilising a combination of randomness and discrimination to classify images finely. Pages 1577–1584 of CVPR 2011. IEEE, 2011.Gaurav Sharma, Fred´ eric Jurie, and Cordelia Schmid. Dis- ´ criminative spatial saliency for image classification. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3506–3513. IEEE, 2012.
6. Leonidas Guibas, Andy Lai Lin, Bangpeng Yao, Xiaoye Jiang, and Li Fei-Fei. Recognizing human action through studying the components and features of action. Pages 1331–1338 in: 2011 International Conference on Computer Vision. IEEE, 2011.
7. Alice Lai and Guodong Guo. a study on recognising human action from still images. 2014;47(10):3343–3361. Pattern Recognition
8. Jasper Uijlings, Raffaella Bernardi, and Dieu Thu Le. using language models to detect hidden activities 2013; pages 231-238 in Proceedings of the Third ACM Conference on International Conference on Multimedia Retrieval.
9. Charless Fowlkes, Deva Ramanan, and Chaitanya Desai. Discriminative models for interactions between static humans and objects. Pages 9–16 in Computer Vision and Pattern Recognition-Workshops, 2010 IEEE Computer Society Conference. IEEE, 2010.

10. Charless C. Fowkes, Deva Ramanan, and Chaitanya Desai. Models that discriminate for multi-class object arrangement. 95(1):1–12 in International Journal of Computer Vision, 2011.

11. Joint pose estimation and action identification in picture graphs. Kumar Raja, Ivan Laptev, Patrick Perez, and Lionel Oisel. Pages 25–28 of 2011's 18th IEEE International Conference on Image Processing. IEEE, 2011.

12. J. Sivic, I. Laptev, and V. Delaitre. Identifying human actions in still images: a comparison of part-based versus bag-of-features representations. Available at http://www.di.ens.fr/willow/research/stillactions/, this version was modified in 2010.