Survey on Real-time Activity Detection and Recognition in Video

Ketki Salunkhe, Priyanka Rajaram, Samidha Raut, Samidha Kurle

Department of Computer Engineering, University of Mumbai,Atharva College of Engineering, Malad, Mumbai, India.

Publication Info

Article history:
Received : 12 February 2020
Accepted : 26 May 2020

Keywords:
Object detection, OpenPose, Security monitoring system, Real-time video recognition and detection.

Corresponding author:
Ketki Salunkhe
e-mail: *ketkisalunkhe333@gmail.com*

Abstract

Real-time object recognition and detection is the ability to automatically analyze object to recognize and assess temporal events which do not rely on a single image. It is the mechanism by which a video is stored, data gathered, and data evaluated for the purpose of collecting domain-specific knowledge. Object identification is the method of identifying instances of real-objects. It enables several objects to be identified, focused, and detected within an image, picture, or in real-time. The identification of anomalous events and artifacts through video becomes quite difficult owing to the uncertain existence of the phenomena, the background under which the incident took place, the absence of sufficient amount of anomalous ground truth testing data and other considerations correlated with weather variability, lighting conditions and the operating state of the cameras recorded. This paper aims to research and evaluate different anomalous behavior detection and event tracking strategies based on the film. Various activity and object detection systems were provided the emphasis. The methods are contrasted from both precision-driven activity detection viewpoints, and real-time computation-driven activity detection. This paper further focuses on work problems and obstacles, technology contexts, reviewed databases, and potential operation and object detection directions.

1. INTRODUCTION

The real-time activity recognition and detection through video do the job of analyzing video sequences to detect and recognize the activities and the objects. This paper is about a security monitoring system that uses the identification of activity of the human body and the detection of items to establish surveillance security. Human does not identify unusual circumstances, though, but computers that analyze the photos collected and recognize suspicious behavior or incidents and artifacts. We have used OpenPose Python. The system records videos and monitors a human's actions. The human face and history of human behavior play an important role in the recognition of individuals.

Video surveillance has become a major concern in everyday life nowadays with the increasingly growing needs of citizens protection and personal assets. A consequence of these requests has resulted in cameras being mounted almost everywhere. Most current video surveillance systems have one function; they need a human operator to track continually, displaying the images collected by the cameras. We can use a method called OpenPose that records videos and identifies suspicious activities. Much of this research is devoted to the visual analysis of human actions and the identification of unusual human movements and suspect objects in exam centers.

2. LITERATURE SURVEY

2.1. Convolutional Pose Machines

The work of Shih-En Wei and et al. [3] is based on convolutional pose machines. CPM introduced a sequential architecture which would consist of convolutional networks.

At a particular stage in the CPM, the spatial context of part beliefs gives clear disambiguating signals for a subsequent point. As a result, each point of a CPM generates belief maps with increasingly precise estimates for each part's locations. To capture long range interactions between pieces, the design of the network at each stage of our sequential prediction framework is guided by the goal of achieving a broad receptive field on both the image and the maps of beliefs. Shih-En Wei et al. [3] found, through experiments, that large fields of receptivity on the maps of beliefs are important for studying long-range spatial relationships and result in improved accuracy. CPMs inherit the advantages of pose machine architecture through the implicit learning of long-range dependencies between image and multipart indications, close integration of learning and inference, modular sequential design and combine them with the benefits of convolutional architecture: The ability to learn feature representations to both image and spatial background directly from data; a differentiable architecture that allows for global joint backpropagation training; and

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

the ability to manage massive training datasets efficiently. CPM fixes the problem of the vanishing gradients during training by presenting an objective natural learning mechanism that enforces intermediate supervision. This issue can occur because backpropagated gradients decrease in intensity as they are propagated across the network's many layers.

2.2. Hand Key-point Detection using Multiview Bootstrapping

Tomas Simon and et al. [2] have presented A multi-camera setup method for boosting the performance of a given key-point detector. This approach is based on the following observation: even if there is a substantial occlusion in a specific picture of the side, an un-occluded view also exists. This experience is systematized by multi-view bootstrapping to create a more efficient hand detector that generalizes beyond the capture setup. In particular, it allows a poor detector, trained on a small annotated dataset, to identify key-point subsets in good view, and to filter out incorrect detections using robust 3D triangulation and pictures, where there are extreme occlusions are classified by re-projecting the triangulated 3D hand joints. By adding these newly created annotations into the training collection, the detector is improved iteratively, resulting in more and more accurate detections at each iteration. This strategy produces hand key-point annotations that are geometrically accurate, using multi-view constraints as an external supervisory source. In this way, images that are difficult or impossible to annotate due to occlusion can be labeled. It shows that multi-view bootstrapping produces hand key-point detectors for RGB images, which rival RGB-D hand key-point detectors efficiency. It also proved that applying this single view detector in a multi-camera system enables marker-less 3D hand reconstruction in unparalleled situations, including challenging object manipulation, musical performance, and multiple interacting individuals.

2.3. OpenPose

Zhe Cao et al. [1] suggested an OpenPose method that is one of the most common bottom-up methods for estimating multi-person human poses. OpenPose first identifies parts (key-points) that belong to each person within the picture, then assigns parts to different individuals. The architecture of the OpenPose model is shown Figure 2.

The OpenPose network first removes features from the image using the main few layers (VGG-19 in the flowchart above). The features are then fed onto two parallel divisions of the convolutional layers. A set of 18 confidence maps are projected by the primary division, with each map depicting a particular part of the human skeleton. The second branch estimates a set of 38 Element Fields of Affinity (PAFs) reflecting the degree of interaction between parts.

Successive steps are not meant to optimize the assumptions made by each branch. Bipartite graphs are built between pairs of parts using the part confidence maps. Weaker relations within the bipartite graphs are pruned using the PAF values. Via the above measures, skeletons of human poses are often estimated and allocated within the image to each individual.

2.4. Part Affinity Field (PAF)

The first bottom-up representation of the association scores was given by Zhe Cao et al. [1] through PAF. Each pair of body parts includes a PAF, i.e., neck, nose, elbow, and so forth. A PAF can be a series of flow fields encoding unstructured pairwise connections between parts of the body. Let CC be the number of body parts in pairs. At that point, PAFs are:

The set $L= (L_1, L_2, \ldots, L_C)$ where $L_c \in R^{w \times h \times 2}$, $c \in 1 \ldots C$.

On the off chance that a pixel is on an appendage (body part), the value in Lc at that pixel might be a 2D unit vector from the earliest starting point joint to the top joint.

This approach takes an input image (a) and, at the same time, adds two maps with projections of body parts (b) and PAFs. This then parses body part candidates and runs a special bipartite matching algorithm to connect them (d); this eventually assembles the body parts into complete body poses (e). Figure 1 above displays all the stages from an input image (Figure 1a) to anatomic key-points as an output (Figure 1e). Initially, a feed-forward neural system predicts a lot of body part areas on the picture (Figure 1b) as a confidence map and a lot of PAFs that encode the level of relationship between these body parts (Figure 1c). Accordingly, the calculation gets all data essential for additional coordinating of appendages and individuals. Next, confidence maps and affinity fields are parsed together (Figure 1d) to produce the final location of the limbs for all individuals on the image.

Multi-Person Parsing using Part Affinity Fields (PAFs)

The parsing strategy can be summarized in three phases:

- Step 1: Use the confidence maps to discover all joints areas.
- Step 2: Find which joints go together to make appendages (body parts) utilizing the part affinity fields and joints in stage 1.
- Step 3: Compare limbs belonging to an identical individual and receive the final set of human poses.

2.5. Multi-hypothesis Tracking

Wei Niu et al. [4] suggested a stable, real-time algorithm that is well adapted for examining outdoor, far-field events. Such mechanisms enhance overall robustness and precision by retaining detection and recovery from both non-catastrophic

 2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

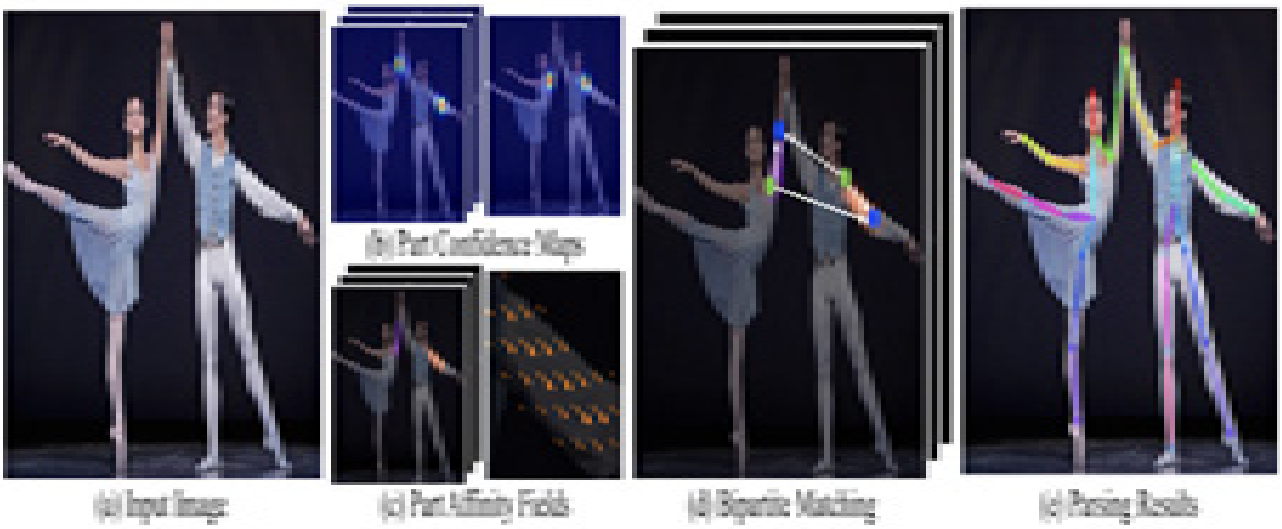

Figure 1: Pipeline for PAF

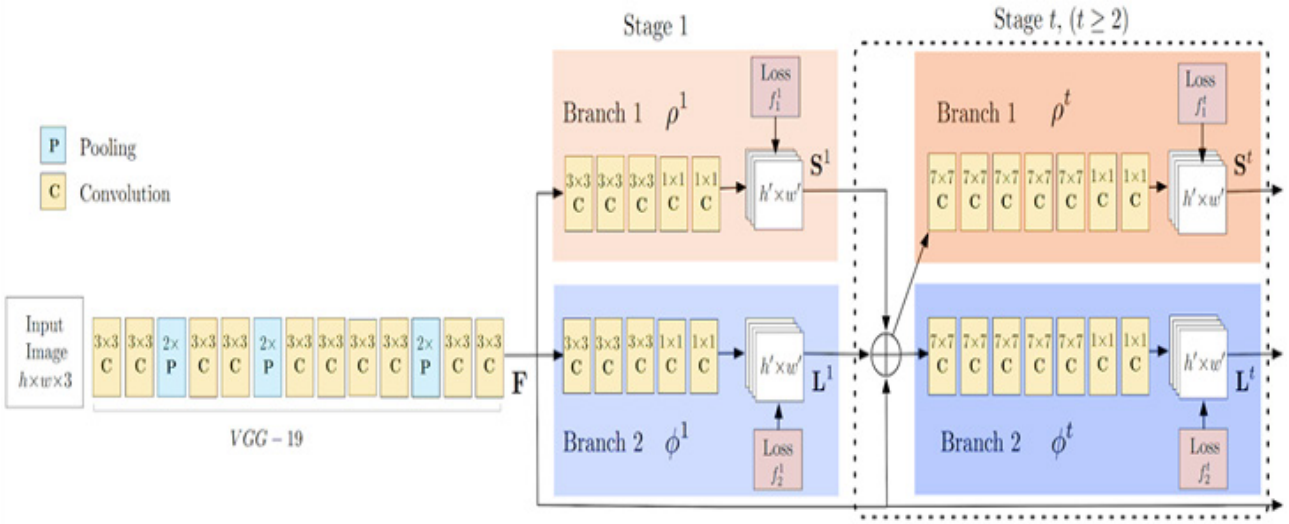

Figure 2: Flowchart of the OpenPose architecture

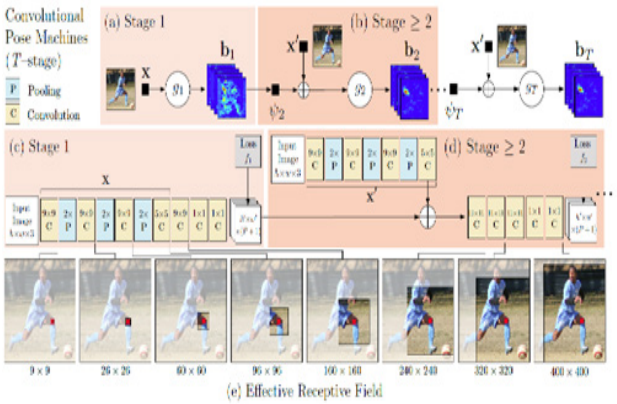

Figure 3: Stages of CPM

errors (such as rare, brief periods of occlusion and shadow merging) and catastrophic errors (such as lengthy stretches of lack of operation from the field of view of the camera). Action identification is based on tracked pathways. A scheme has been developed that distinguishes between various forms of contact within a group of people by defining distinct signatures in the relative location and velocity of the trajectories of the participants.

2.6. Feature Extraction

Gowsikhaa D, Manjunath, et al. [7] suggested a methodology that would incorporate real-life video into the framework. It is divided into frames, and pre-processing happens. Pre-processing consists of subtraction of background and

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

reduction of noise. Those pre-processed frames are then incorporated into the system. The two main methods, head motion detection, and contact detection, detect suspicious behaviors. The detection of head motion is rendered by the identification of the human face. This is achieved through artificial neural networks. The hands are identified for contact detection using an and operation on motion detection, skin color detection and edge detection. Finally, the detected activity is classified into suspicious and non-suspicious. The alert reported on detecting suspicious activity is the output of the system. If there are no suspicious activities detected, there is no output from the system.

3. DATASETS USED

3.1. MPII Dataset

Zhe Cao et al. [1], Tomas Simon et al. [2], and Shih-En Wei et al. [3] have used the MPII dataset. Human Pose data set is a state of the art benchmark for calculating articulated human pose estimation. The dataset contains approximately 25 K images of over 40 K individuals with annotated body joints. The photos were periodically collected using an existing taxonomy of day-to-day human activities. The total dataset includes 410 human activities, and a laboratory for each image is given. Each image was taken from a YouTube video, and the corresponding frames were provided to follow.

3.2. LSP Dataset

This dataset includes 2,000 pose annotated photos of mostly sportspeople using the tags obtained from Flickr, which was used by Shih-En Wei et al. [3]. The images have been scaled so that the most prominent person is approximately 150 pixels in length.14 joint positions were annotated for each image. From a person-centered perspective, the left and right joints are consistently marked.

3.3. FLIC Dataset

The FLIC-full dataset is a complete set of frames collected from films and sent to Mechanical Turk for the hand-annotation of joints. FLIC is the dataset which is used by Shih-En Wei et al. [3] for detecting elbow and wrist joints. It consists of 3987 training images and 1016 test images. This gives more accuracy in results.

3.4. COCO Dataset

Zhe Cao et al.[1] have referred to the COCO dataset. COCO is a huge image dataset designed for object detection, segmentation, individual key-point detection, object segmentation, and caption generation. This package provides the Matlab, Python, and Lua APIs to help load, read, and visualize of COCO annotations The Common Objects in Context (COCO) dataset has 200,000 images in 80 categories, and over 500,000 object annotations. It is the largest publicly available database for object detection.

4. APPLICATIONS

- The proposed system can be used for bank robbery detection.
- It can also be used for developing a patient monitoring system.
- Detecting and reporting suspicious activities at the railway station is also one of the important applications of the proposed system.
- It can also be used for developing security-related applications such as defense, military, checking at airports.

5. FUTURE SCOPE

Most researchers have struggled to manage multiple entities in a single end-to-end system, and are an important avenue for future research. The multi-view bootstrapping method proposed by Tomas Simon et al. [2] can be made robust enough to function with fewer cameras and allow the development of even richer data sets that more closely resemble real-world capturing conditions in less regulated environments. The PAF approach [1] appears to combine annotations from different people in extremely crowded images where people overlap while ignoring others due to the overlapping PAFs that make the greedy multi-person parsing fail. It is also a very challenging task to detect multiple human joints through a single camera. Current literature also lags behind when it comes to real-time analysis of the images. The potential approach could be to adopt online learning methods to understand the anomalies in real-time and integrate deep models of online learning.

6. ACKNOWLEDGMENT

We would like to express our sincere thanks to Prof. Samidha Kurle for her co-operation and guidance. We would also like to thank our Computer Department HOD, Prof. Suvarna Pansambal, our Project Coordinators Dr. Mamta Meena, Prof. Shweta Sharma, and all the Computer Department staff who have directed us to grow this project concept throughout.

7. REFERENCES

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, IEEE Transactions on Pattern Analysis and Machine Intelligence- 2019.

[2] Tomas Simon, Hanbyul Joo, Iain Matthews, Yaser Sheikh, “Hand Keypoint Detection in Single Images using Multiview Bootstrapping”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2017.

[3] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser

Sheikh, "Convolutional Pose Machines," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2016.

[4] Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang, "Human Activity Detection and Recognition for Video Surveillance," IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), 2004.

[5] Rajat Singh, Sarvesh Vishwakarma, Anupam Agrawal, M.D Tiwari, "Unusual activity detection for video surveillance ", Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia - IITM '10, 2010.

[6] Seyed Yahya Nikoueia, Yu Chena, Alexander Avedb, Erik Blaschb, Timothy R. Faughnanc, "I-SAFE: Instant Suspicious Activity identiFication at the Edge using Fuzzy Decision Making," presented at the Fourth ACM/IEEE Symposium on Edge Computing, Washington DC,2019.

[7] Gowsikhaa D, Manjunath, Abirami S, "Suspicious Human Activity Detection from Surveillance Videos" (IJIDCS) International Journal on Internet and Distributed Computing Systems. Vol: 2 No: 2, 2012

[8] Karishma Pawar, Vahida Attar, "Deep learning approaches for video-based anomalous activity detection."

[9] Sumalatha Ramachandran, Lakshmi Harika Palivela, and C. Giridharan, "An intelligent system to detect human suspicious activity using deep neural networks."

[10] T. Senthil Kumar and G. Narmatha, "Video Analysis for Malpractice Detection in Classroom Examination," Conference on Soft Computing Systems, Advances in Intelligent Systems and Computing 397, 2016.

AUTHORS

Ketki Salunkhe
Student
Atharva College of Engineering

Priyanka Rajaram
Student
Atharva College of Engineering

Samidha Raut
Student
Atharva College of Engineering

Prof. Samidha Kurle
Teacher
Atharva College of Engineering

