IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

A Brief Study on Human Action Recognition

Mr. Nagesh UB[1], Abhishek V Doddagoudra[2], Adarsh KM[3], Mayoori K Bhat[4], Shreya L[5]
Assistant Professor, Department of Information Science and Engineering[1]
Students Department of Information Science and Engineering[2,3,4,5]
Alva's Institute of Engineering and Technology, Mijar, Mangalore, Karnataka, India

Abstract: *Human action recognition is the process of labelling image sequences with action labels. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. The task is challenging due to variations in motion performance, recording settings and inter-personal differences. In this survey, we explicitly address these challenges. We provide a detailed overview of current advances in the field. Image representations and the subsequent classification process are discussed separately to focus on the novelties of recent research. Moreover, we discuss limitations of the state of the art and outline promising directions of research.*

Keywords: Human action recognition.

I. INTRODUCTION

We consider the task of labelling videos containing human motion with action classes. The interest in the topic is motivated by the promise of many applications, both offline and online. Automatic annotation of video enables more efficient searching, for example finding tackles in soccer matches, handshakes in news footage or typical dance moves in music videos. Online processing allows for automatic surveillance, for example in shopping malls, but also in smart homes for the elderly to support aging in place. Interactive applications, for example in human–computer interaction or games, also benefit from the advances in automatic human action recognition. In this section, we first discuss related surveys and describe the most common datasets and algorithms. Also, we outline the main characteristics and challenges of the field as these motivate the various approaches that are reported in literature. In its simplest form, vision-based human action recognition can be regarded as a combination of feature extraction, and subsequent classification of these image representation.

II. SURVEYS

Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review [1]

Di Wu, Nabin Sharma, Michael Blumenstein [1] Video-based human action recognition has become one of the most popular research areas in the field of computer vision and pattern recognition in recent years. It has a wide variety of applications such as surveillance, robotics, health care, video searching and human-computer interaction. There are many challenges involved in human action recognition in videos, such as cluttered backgrounds, occlusions, view point variation, execution rate, and camera motion. A large number of techniques have been proposed to address the challenges over the decades. Three different types of datasets namely, single viewpoint, multiple view point and RGB-depth videos, are used for research. This paper presents a review of various state-of-the art deep learning-based techniques proposed for human action recognition on the three types of datasets. In light of the growing popularity and the recent developments in video-based human action recognition, this review imparts details of current trends and potential directions for future work to assist researchers.

Action Recognition for Surveillance Applications Using Optic Flow and SVM [2]

Somayeh Danafar, Niloofar Gheissari [2] Low quality images taken by surveillance cameras pose a great challenge to human action recognition algorithms. This is because they are usually noisy, of low resolution and of low frame rate. In this paper we propose an action recognition algorithm to overcome the above challenges. Author used optic flow to construct motion descriptors and apply a SVM to classify them. Having powerful discriminative features, significantly reduce the size of the feature set required. This algorithm can be applied to videos with low frame rate without scarifying efficiency or accuracy, and is robust to scale and view point changes. To evaluate our method, author used a database consisting of walking, running, jogging, hand clapping, hand waving and boxing actions. This grayscale database has

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

images of low resolution and poor quality. This image database resembles images taken by surveillance cameras. The proposed method outperforms competing algorithms evaluated on the same database.

Violent Flows: Real -Time Detection of Violent Crowd Behavior [3]

Tal Hassner, Yossi Itcher, Orit Kliper-Gross [3] Although surveillance video cameras are now widely used, their effectiveness is questionable. Here, author focus on the challenging task of monitoring crowded events for out breaks of violence. Such scenes require a human surveyor to monitor multiple video screens, presenting crowds of people in a constantly changing sea of activity, and to identify signs of breaking violence early enough to alert help. With this in mind, author proposed the following:

Novel approach is described to real-time detection of breaking violence in crowded scenes. Our method considers statistics of how flow vector magnitudes change over time. These statistics, collected for short frame sequences, are represented using the Violent Flows (ViF) descriptor. ViF descriptors are then classified as either violent or non-violent using linear SVM.

Present a unique data set of real-world surveillance videos, along with standard benchmarks designed to test both violent/nonviolent classification, as well as real-time detection accuracy. Finally, Provide empirical tests, comparing our method to state-of-the art techniques, and demonstrating its effectiveness.

Human Activity Recognition: A Review [4]

Ong Chin Ann, Lau Bee Theng [4] Human Activity Recognition is one of the active research areas in computer vision for various contexts like security surveillance, healthcare and human computer interaction. In this paper, a total of thirty-two recent research papers on sensing technologies used in HAR are reviewed. The review covers three area of sensing technologies namely RGB cameras, depth sensors and wearable devices. It also discusses on the pros and cons of the mentioned sensing technologies. The findings showed that RGB cameras have lower popularity when compared to depth sensors and wearable devices in HAR research.

Human Action Monitoring for Healthcare Based on Deep Learning [5]

Yongbin Gao, Xuehao Xiang, Naixue Xiong, Bo Huang, Hyo Jong Lee, Rad Alrifai, Xiaoyan Jiang, and Zhijun Fang [5] Human action recognition enables efficient and accurate monitoring of human behaviors, which can exhibit multifaceted complexity attributed to disparities in viewpoints, personality, resolution and motion speed of individuals, etc. The spatial-temporal information plays an important role in the human action recognition. In this paper, Author proposed a novel deep learning architecture named as recurrent 3D convolutional neural network (R3D) to extract effective and discriminative spatial temporal features to be used for action recognition, which enables the capturing of long-range temporal information by aggregating the 3D convolutional network entries to serve as an input to the LSTM (Long Short-Term Memory) architecture. The 3D convolutional network and LSTM are two effective methods for extracting the temporal information. The proposed R3D network integrated these two methods by sharing a shared 3D convolutional network in sliding windows on video streaming to capturing short-term spatial-temporal features into the LSTM. The output features of LSTM encapsulate the long range spatial-temporal information representing high-level abstraction of the human actions.

Human Action Recognition using Deep Learning Methods [6]

Zeqi Yu, Wei Qi Yan [6] Human action recognition seeks to recognize and comprehend the behaviors of individuals in films and export pertinent tags. Actions in a video also possess the qualities in the time domain in addition to the spatial correlation seen in 2D images. The recognition will be impacted by the complexity of human behavior, such as shifting views, background noises, and other factors. In this research, three methods are created and put into practice to address these challenging issues. TwoStream CNN, CNN+LSTM, and 3D CNN, which are based on convolutional neural networks (CNN), are used to recognize human actions in videos. Each method is explained in detail and examined. To test these algorithms and obtain the best results, the HMDB-51 dataset is used.

Human Action Recognition Using Deep Neural Networks [7]

Rashmi R. Koli Tanveer I. Bagban[7] In deep neural networks, human behaviors like body language pose the greatest technical and practical challenges. Human Action recognition is nothing more than the recognition of human gestures. A gesture is a movement of bodily components that indicates a deeper meaning. The best and most natural way for humans to connect with systems (computers) is through gestures, which creates a bridge between humans and technology.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

Recognition of human action offers a platform for communication with the dumb and deaf. In this study, authors present the creation of a platform for hand movement detection. Using CNN, it can distinguish human gestures in a picture by recognizing hand movement (gestures).

Human Action Recognition using Machine Learning in Uncontrolled Environment [8]

Inzamam Mashood Nasir, Mudassar Raza, Jamal Hussain Shah, Muhammad Attique Khan, Amjad Rehman [8] The most crucial stage in action recognition is human detection in videos, which is an active area of machine learning research called video-based human action recognition (HAR). Recently, a number of methods and algorithms have been put out to boost the HAR process' accuracy, yet room for improvement still exists. The unpredictable variations in human look, attire, lighting, and background make it difficult to identify and categorize human actions. This article describes an effective method for categorizing human actions using steps like trimming out unnecessary frames from videos, extracting segments of interest (SoIs), mining feature descriptors using geodesic distance (GD), 3D Cartesian-plane features (3D-CF), Joints MOCAP (JMOCAP), and n-way point trajectory generation (nPTG).

Analysis of Human Activity Recognition using Deep Learning [9]

Lamiyah Khattar, Chinmay Kapoor, Garima Aggarwal [9] With new technologies emerging every day, the amount of data is exploding today. These recent developments have also contributed to an increase in growth in industries like robotics and the internet of things (IoT). The use and precision of various Human Activity Recognition models can be compared thanks to this study. Specifically, 2-D Convolutional Neural Network and Long-Short Term Memory models will be discussed. Both models are trained using the same dataset, which contains data obtained from a public website and contains wearable sensor data, in order to preserve the consistency and validity of the survey.

Sequential Deep Learning for Human Action Recognition [10]

Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt [10] In this research, authors offer a completely automated deep model that can categorize human actions without any prior knowledge. Automatic learning of spatiotemporal characteristics is the initial stage of our approach, which is based on the 3D extension of convolutional neural networks. The next stage is to train a recurrent neural network to categorize each sequence taking into account the temporal evolution of the previously learnt features for each timestep. Experimental findings on the KTH dataset demonstrate that the suggested strategy outperforms current deep models and produces results that are comparable to those of the best related research.

III. CHALLENGES AND CHARACTERISTICS OF THE DOMAIN

In human action recognition, the common approach is to extract image features from the video and to issue a corresponding action class label. The classification algorithm is usually learned from training data. In this section, we discuss the challenges that influence the choice of image representation and classification algorithm.

3.1 Intra- and Inter-Class Variations

For many actions, there are large variations in performance. For example, walking movements can differ in speed and stride length. Also, there are anthropometric differences between individuals. Similar observations can be made for other actions, especially for non-cyclic actions or actions that are adapted to the environment (e.g., avoiding obstacles while walking, or pointing towards a certain location). A good human action recognition approach should be able to generalize over variations within one class and distinguish between actions of different classes. For increasing numbers of action classes, this will be more challenging as the overlap between classes will be higher. In some domains, a distribution over class labels might be a suitable alternative.

3.2 Environment and Recording Settings

The environment in which the action performance takes place is an important source of variation in the recording. Person localization might prove harder in cluttered or dynamic environments. Moreover, parts of the person might be occluded in the recording. Lighting conditions can further influence the appearance of the person. The same action, observed from different viewpoints, can lead to very different image observations. Assuming a known camera viewpoint restricts the use to static cameras. When multiple cameras are used, viewpoint problems and issues with occlusion can be alleviated, especially when observations from multiple views can be combined into a consistent representation. Dynamic

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

backgrounds increase the complexity of localizing the person in the image and robustly observing the motion. When using a moving camera, these challenges become even harder. In vision-based human action recognition, all these issues should be addressed explicitly.

3.3 Temporal Variations

Often, actions are assumed to be readily segmented in time. Such an assumption moves the burden of the segmentation from the recognition task, but requires a separate segmentation process to have been employed previously. This might not always be realistic. Recent work on action detection (see Section 3.3) addresses this issue. Also, there can be substantial variation in the rate of performance of an action. The rate at which the action is recorded has an important effect on the temporal extent of an action, especially when motion features are used. A robust human action recognition algorithm should be invariant to different rates of execution.

3.4 Obtaining and Labelling Training Data

Many works described in this survey use publicly available datasets hat are specifically recorded for training and evaluation. This provides a sound mechanism for comparison but the sets often lack some of the earlier mentioned variations. Recently, more realistic datasets have been introduced (see also Section 1.4). These contain labelled sequences gathered from movies or web videos. While these sets address common variations, they are still limited in the number of training and test sequences. Also, labelling these sequences is challenging.

IV. COMMON DATASETS

The use of publicly available datasets allows for the comparison of different approaches and gives insight into the (in)abilities of respective methods. We discuss the most widely used sets.

4.1 KTH Human Motion Dataset

The KTH human motion dataset contains six actions (walking, jogging, running, boxing, hand waving and hand clapping), performed by 25 different actors. Four different scenarios are used: outdoors, outdoors with zooming, outdoors with different clothing and indoors. There is considerable variation in the performance and duration, and somewhat in the viewpoint. The backgrounds are relatively static. Apart from the zooming scenario, there is only slight camera movement.

4.2 Weizmann Human Action Dataset

The human action dataset recorded at the Weizmann institute contains 10 actions (walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip), each performed by 10 persons. The backgrounds are static and foreground silhouettes are included in the dataset. The viewpoint is static. In addition to this dataset, two separate sets of sequences were recorded for robustness evaluation. One set shows walking movement viewed from different angles. The second set shows fronto-parallel walking actions with slight variations (carrying objects, different clothing, different styles).

4.3 INRIA XMAS Multi-View Dataset

Weinland et al. introduced the IXMAS dataset that contains actions captured from five viewpoints. A total of 11 persons performs 14 actions (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw overhead and throw from bottom up). The actions are performed in an arbitrary direction with regard to the camera setup. The camera views are fixed, with a static background and illumination settings. Silhouettes and volumetric voxel representations are part of the dataset.

4.4 UCF Sports Action Dataset

The UCF sports action dataset contains 150 sequences of sport motions (diving, golf swinging, kicking, weightlifting, horseback riding, running, skating, swinging a baseball bat and walking). Bounding boxes of the human figure are

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

provided with the dataset. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and background.

4.5 Hollywood Human Action Dataset

The Hollywood human action dataset contains eight actions (answer phone, get out of car, handshake, hug, kiss, sit down, sit up and stand up), extracted from movies and performed by a variety of actors. A second version of the dataset includes four additional actions (drive car, eat, fight, run) and an increased number of samples for each class. One training set is automatically annotated using scripts of the movies, another is manually labelled. There is a huge variety of performance of the actions, both spatially and temporally. Occlusions, camera movements and dynamic backgrounds make this dataset challenging. Most of the samples are at the scale of the upper-body but some show the entire body or a close-up of the face.

V. ALGORITHMS

5.1 LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is particularly well-suited for modelling long-term dependencies in sequential data. It is often used in action recognition tasks because it is able to learn the dependencies between poses or movements in a sequence over a long period of time, allowing it to accurately classify actions even if they span a large number of frames.

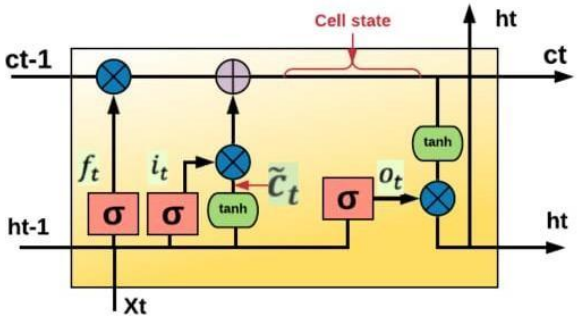

Figure 1: Single LSTM cell

LSTMs work by using a series of "gates" to control the flow of information through the network. These gates allow the network to selectively choose which information to remember and which to forget, enabling it to effectively model long-term dependencies.

In the context of action recognition, an LSTM model might be trained to recognize an action by inputting a sequence of poses or movements into the network and training it to classify the action based on this input. The LSTM would then learn to recognize the dependencies between different poses or movements in the sequence, allowing it to accurately classify the action being performed even if it spans a large number of frames.

There are several other algorithms that can be used for human action recognition, including:

1. **Hidden Markov Models (HMM)**: These models are based on the Markov assumption that the future state of a system can be predicted based on its current state and the transitions between states. In the context of action recognition, HMMs can be used to model the sequence of poses or movements that make up an action.
2. **Dynamic Time Warping (DTW)**: DTW is an algorithm that can be used to find the optimal alignment between two sequences of data. It is often used in action recognition to align the sequence of poses or movements in a video with a reference sequence of poses or movements.
3. **Conditional Random Fields (CRF)**: CRFs are a type of probabilistic graphical model that can be used to model the dependencies between different variables in a sequence. In the context of action recognition, CRFs can be used to model the dependencies between different poses or movements in a sequence, allowing the model to make more informed predictions about the action being performed.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

4. **Recurrent Neural Networks (RNNs)**: RNNs are a type of neural network that can process sequential data. They are often used in action recognition tasks because they are able to learn the dependencies between poses or movements in a sequence.
5. **Convolutional Neural Networks (CNNs)**: CNNs are a type of neural network that is particularly well-suited for image classification tasks. They can be used in action recognition by analysing the spatial and temporal features of video frames to classify the actions being performed.
6. **Spatiotemporal Graph Convolutional Networks (ST-GCNs)**: ST-GCNs are a type of neural network that can process both spatial and temporal data and are specifically designed for action recognition tasks. They operate by convolving over a graph representation of the data, which allows them to effectively capture the dependencies between different poses or movements in a sequence.

VI. CONCLUSION

Human action recognition is a field of study that involves developing algorithms and systems to automatically recognize and classify human actions in video or other types of data. It has a wide range of applications, including video surveillance, sports analysis, and human-computer interaction. There are many different algorithms and approaches that have been developed for action recognition, including Hidden Markov Models, Dynamic Time Warping, Conditional Random Fields, Recurrent Neural Networks, Convolutional Neural Networks, and Spatiotemporal Graph Convolutional Networks. These approaches have been applied to a variety of different tasks, including recognizing actions in real-time video streams, analysing sports footage to identify specific plays or techniques, and detecting and classifying actions in large datasets of annotated video. Overall, action recognition is a challenging but important area of study that has the potential to have significant impact in a variety of fields.

REFERENCES

[1]. D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review", in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2017, pp. 2865–2872.
[2]. S. Danafar and N. Gheissari, "Action recognition for surveillance applications using optic flow and SVM", in Proc. Asian Conf. Comput. Vis. Berlin, Germany: Springer, 2007, pp. 457.
[3]. T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior" in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2012, pp. 1–6.
[4]. Ong Chin Ann,Lau Bee Theng "Human Activity Recognition: A Review" in Proc. 2014 IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014, Penang, Malaysia.
[5]. Yongbin Gao, Xuehao Xiang, Naixue Xiong, Bo Huang, Hyo Jong Lee, Rad Alrifai, Xiaoyan Jiang, and Zhijun Fang "Human Action Monitoring for Healthcare Based on Deep Learning" in Proc. 2018 IEEE International Conference.
[6]. Zeqi Yu and Wei Qi Yan "Human Action Recognition Using Deep Learning Methods" in Proc. IEEE International Conference.
[7]. Rashmi R. Koli Tanveer I.Bagban "Human Action Recognition using Deep Neural Network" in Proc. IEEE International Conference.
[8]. Inzamam Mashood Nasir, Mudassar Raza, Jamal Hussain Shah, Muhammad Attique Khan, Amjad Rehman "Human Action Recognition using Machine Learning in Uncontrolled Environment" in Proc. IEEE International Conference.
[9]. Lamiyah Khattar, Chinmay Kapoor, Garima Aggarwal "Analysis of Human Activity Recognition using Deep Learning" in Proc. IEEE International Conference.
[10]. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt "Sequential Deep Learning for Human Action Recognition" in Proc. IEEE International Conference.