ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 9, September 2020

An Efficient Data Augmentation CNN -Network for Skeleton-based Human Action Recognition

L.Arthi[1], Mrs.Nisha Priya[2]
[1]Student, [2]Assistant Professor, Head of the Department, Department of Computer Science and Engineering,
CSI College of Engineering, Ketti, Nilgiris, Tamilnadu, India.

Abstract: Data augmentation is a widely used technique for enhancing the generalization ability of deep neural networks for skeleton-based human action recognition (HAR) tasks. Most existing data augmentation methods generate new samples by means of handcrafted transforms. These methods often cannot be trained and then are discarded during testing because of the lack of learn able parameters. To solve those problems, a novel type of data augmentation network called a sample fusion network (SFN) is proposed. Instead of using handcrafted transforms, an SFN generates new samples via a long short term memory (LSTM) auto encoder (AE) network. SFN and an HAR network can be cascaded together to form a combined network that can be trained in an end-to-end manner. Adaptive weighting strategy is employed to improve the complementation between a sample and the new sample generated from it by an SFN, thus allowing the SFN to more efficient improve the performance of the HAR network during testing. Experimental results on various data sets verify that the proposed method outperforms state-of-the-art data augmentation methods. More importantly, the proposed SFN architecture is a general framework that can be integrated with various types of networks for HAR. For example, when a baseline HAR model with 3 LSTM layers and 1 fully connected (FC) layer was used, the clarification accuracy was increased from 79.53% to 90.75% on the NTU RGB+D data set using a cross-view protocol, thus outperforming most other methods.

Keywords: Authentication, Criminals, Mobile Application, Reporting.

I. INTRODUCTION

Human brain gathers more information visually and processes these visual information faster than the textual information. There is a popular axiom "A Picture is worth a thousand words", which refers to the meaning that a picture conveys information more effectively than words. According to Forrester Research's Dr. James McQuivey, 'one minute of a video is worth 1.8 million words". To understand an occurrence, single image is measy so we need a video (i.e. moving visual image). Now-a-days no one could imagine the world without videos. Videos are a meaningful and powerful way of conveying the information. Cameras are used almost everywhere to record videos. Video surveillance cameras are introduced in many public places such as schools, hospitals, banks, shops, airports etc. This is widely accepted by the society also. There are many accessories for recording, storing and sharing videos. The videos are evolving rapidly. So the need for understanding video is also increasing. The real world scene is a continuous 3D signal (temporal, horizontal and vertical). The human can understand the videos naturally by visual means, but in order to make computers understand the videos automatically, intelligent systems are developed. These intelligent systems could analyze and recognize the activity or the happenings in videos. Such systems are already available and they are performing the video understanding and action recognition much faster than humans. The skeleton is a high-level representation of human action that is robust to variations in location and appearance. Moreover, rapid advances in imaging technology and the development of a recently, deep learning methods using skeletons have been undergoing rapid development because they can automatically extract spatial-temporal relationships among joints. Applications of these works have achieved outstanding performance in skeleton-based HAR. However, since skeleton data are far less abundant than RGB data, over fitting has become a very serious problem for deep learning methods, even in shallow networks. This problem limits the generalization ability of deep learning methods.To overcome such limitations, many regularization methods have been proposed. These methods can be broadly categorized into three groups, namely, loss function regularization, network structure regularization and data augmentation. In contrast to the two other types of regularization methods, data augmentation focuses on the data level and does not require the design of a new loss function or modification of the network structure. Because of these merits, data augmentation is widely used during the training of deep neural networks to improve their

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 9, September 2020

generalization ability. Data augmentation methods generate new samples by means of handcrafted transforms, the parameters of which cannot be learned. Therefore, these methods cannot be trained along with the training of an HAR network.

II. Literature Survey

J.K. Aggarwal [1] has developed Understanding of Human Motion, Actions and Interactions. The efforts to develop computer systems able to detect humans and recognize their activities form an important area of research in computer vision today. Motion is an important cue for the human visual system and for understanding human actions. The research included the study of interactions at the gross (blob) level and at the most detailed (head, torso, arms and legs) level. For blob level analysis, a modified Hough transform called the Temporal Spatio-Velocity transform to isolate pixels with similar velocity profiles is used. For the detailed-level analysis, a multi-target, the multi - assignment strategy to track blobs in consecutive frames is used.

Venet Osmani et al. [2] have developed "Human activity recognition in pervasive healthcare: Supporting efficient remote collaboration". The activity recognition system that is described in conjunction with their efficiency mechanism has the potential to cut down healthcare costs by making the working environments more efficient. The activity recognition process that has the ability to infer user activities based on the self-organization of surrounding objects that the user may manipulate is developing. The results show an accurate activity recognition process for individual users with respect to their behavior. At the same time supported remote virtual collaboration through task allocation process Between doctors and nurses with results shows maximum efficiency within the resource constraints.

Dhruv Batra et al. [9] have developed Gabor Filter based Fingerprint Classification Using Support Vector Machines. Fingerprint classification is important for various practical applications. Gabor filter based Feature extraction scheme is used to generate a 384 dimensional feature vector for each fingerprint image. The classification of these patterns is done through two stage classifier in which K Nearest Neighbor (K") acts as the first step and finds out the two most frequently represented classes amongst the K nearest pattern, followed by the pertinent SVM classifier choosing the most apt class of the two.

Yu Su, Shiguang et al. [12] have developed Patch-Based Gabor Fisher Classifier for Face Recognition. Face representations based on Gabor features have achieved great success in face recognition, such as Elastic Graph Matching, Gabor Fisher Classifier (GFC), and AdaBoosted Gabor Fisher Classifier (AGFC). A patch-based GFC (PGFC) method is presented, in which Gabor features are spatially partitioned into a number of patches, and on each patch one GFC is constructed as a component classifier to form the final ensemble classifier using sum rule.

III. Problem Identification

- Identify the region within a web document where the relevant data is most likely to reside and the text of the input documents requires them to be well-formed
- Searches for mismatches and then tries to find out if they must be generalised to a capturing group.
- A repetition or an optional expression, which is a complex procedure that requires backtracking.

IV. Proposed System

CNN-based method

A novel action recognition method based on space, time interest points and Euclidean similarity measure is presented. The input video is split into frames and these frames are enhanced using preprocessing technique. The proposed method requires some prior knowledge about actions, namely foreground, background estimation and motion estimation. Salient features are extracted from the descriptors like the space, time, motion value. The feature vectors are normalized and then concatenated. The concatenated features are given as input to the classifier. In this approach an action is considered as a particular class of image sequences and the unknown image sequence (i.e. An input) is recognized as an action by categorizing it into its class.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 9, September 2020

Skeleton sequences are converted into images, thus converting the task of skeleton-based HAR into an image classification task. Therefore, the key question is how to effectively represent temporal information in the form of image properties, including color and texture. Skeleton sequence as a matrix by concatenating the joint coordinates at each instant and arranging the vector representations in chronological order. proposed a method called joint trajectory maps (JTM), in which the trajectories are mapped into the hue, saturation, and value (HSV) space, to encode spatio-temporal information into multiple texture images. Used joint distance maps (JDM) to encode the pairwise distances between the skeleton joints of single or multiple subjects into image textures. Drew skeleton joints with a specific pen onto three orthogonal canvases and then encoded the dynamic information in the skeleton sequences in color. Encoded skeletons into a series of color images and then applied visual/motion enhancement methods to the color images to enhance their local patterns. proposed a generic graph-based model called a spatial-temporal graph convectional network (ST-GCN) to automatically learn both spatial and temporal patterns from data shown in the figure.

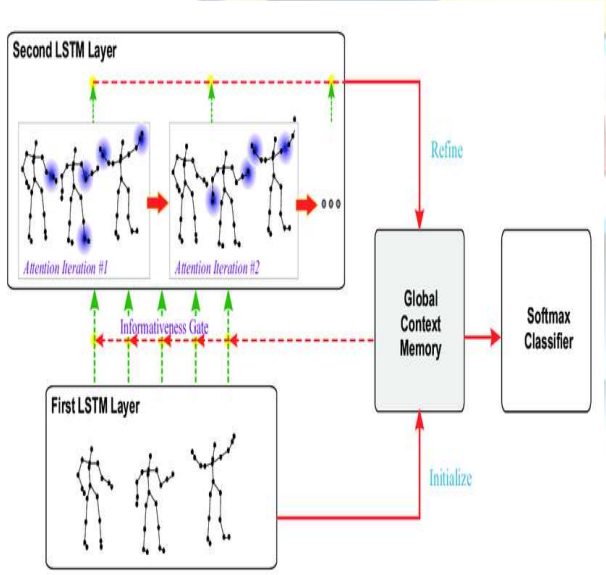

Figure1: LSTM Layer on convectional network

Background subtraction is one of the task in machine vision applications. There are two phases in the background subtraction technique, namely the background modeling and the foreground detection. Background Modeling It is important to model the background properly in order to prevent foreground misclassification such as shadows and reflections. Background modeling is achieved by averaging the number of frames. The background model contains only the static background without any foreground object.

Foreground Detection It is assumed that any kind of change in the image in one frame and the other is due to the moving object. Hence, Foreground object (i.e. the human) is detected by segmenting the current image from the background model image

Depth and Skelton Features With the application of depth evaluation sensors (RGBD sensors), the human action detection approaches have gained improved recognition results due to the depth analysis of data. These approaches are further classified as depth sequence-based approaches and Skelton based approaches .These approaches use the basic global and local features s well to extract a composite feature vector from every action sequence. In the depth-based feature extraction approaches, initially, the motion changes are analyzed through the depth map of the human body. Under this class, a video captured through an RGBD sensor is seen as a space-time structure having the depth information. For a given action sequence, the feature is extracted as a Spatio-temporal feature with a motion or an appearance having changes in the depth. a new action recognition technique based on the additional depth information, i.e., body shape and motion information. In this approach, the depth maps are projected into three orthogonal planes and then accumulate global activities to generate "Depth Motion Maps (DMM)". Further HoGs are computed multi-fusion based action recognition technique based on the DMM and "Local Binary Patterns (LBPs)". This approach employs DMMs for three projection views (top, side, and front) to capture the motion cues and then applies LBPs to extract a composite feature for every action. This approach accomplished two fusion phases; one is feature level fusion and the other is decision level fusion. The compressed depth maps for action recognition. In this approach, every depth map is encoded with a scalable encoder which has multi-scale breakpoints and ad Adaptive "Discrete Wavelet

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 9, September 2020

Transform (DWT)". Here the sharp edges are obtained through breakpoints and the smooth variations are obtained through DWT and are extracted from the bit-stream and are used to construct a set of features that are fed to classifier for recognizing the action.

V. EXPERIMENTAL RESULTS

Weizmann Action dataset is used for analyzing the results. The performance of the system is analyzed by testing different human actions. A common quantitative analysis is performed to assess the overall performance of recognition process. To analyze the performance precision, recall and Fmeasures are used. Precision, also called as the positive predictive value is the fraction of retrieved action instances that are relevant.

$$\boldsymbol{F-score} = (2 * P * R)/(P + R)$$

$$\boldsymbol{Precision} = No.\ of\ instances\ of\ correct\ positive\ recognition\ /\ Total\ No.\ of\ positive\ recognition$$

$$\boldsymbol{Recall} = No.\ of\ instances\ of\ positive\ recognition\ found/\ Total\ No.\ of\ relevant\ input\ instance$$

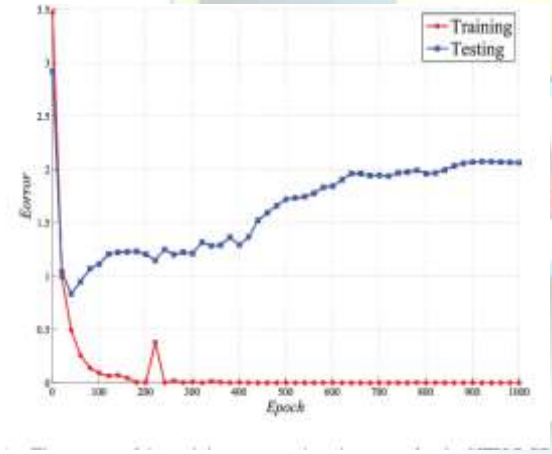

Figure2: Epoch Vs Error

The graph shows error rate on training and testing dataset while compare to training dataset error rate increase in testing dataset

VI. CONCLUSION

In this paper, a detailed and comprehensive survey is carried out over various human action recognition techniques. Initially, an introduction to HAR is explored and in that the need for human action recognition is illustrated. Further detailed information is provided about various datasets. Further, a detailed survey is described and the complete description is done under two phases, one is the survey over various feature extraction techniques and the other is on various action classification techniques. Under the feature extraction techniques, this paper reviewed the local features, depth, and skeleton features. Next, under the action classification techniques, this paper reviewed traditional classifications techniques and also the deep learning strategies. Further, a fine-grained analysis is accomplished over the deep learning approaches with respect to the convolution dimension. Based on the review explored, this paper makes the following conclusions. To perform effective human action recognition, first, the feature extraction must be more effective. Since a system with more detailed information can only recognizes the action even under occlusions, noisy and complex backgrounds. For this purpose, the feature fusion will get priority and need an effective combination. 2. Though the feature fusion gives more prominent results in action recognition, there will be an excessive computational complexity at the classifier. Definitely, the complexity is more for an HAR system which analyzes the data in multiple views than the HAR system which analyzes the data in only one point of view. This can be compensated by an effective classifier design which is also more important in the HAR system.

REFERENCES

[1]. J. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," Pattern Recognition Letters, vol. 48, pp. 70–80, Oct 2014.

[2]. R. Vemulapalli and R. Chellappa, "Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun 2016, pp. 4471–4479.

[3]. L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," Pattern Recognition, vol. 53, pp. 130–147, 2016.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 9, September 2020

[4]. S. Escalera, V. Athitsos, and I. Guyon, "Challenges in Multi-modal Gesture Recognition," in Gesture Recognition. Springer, 2017, pp.1–60

[5]. F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," Computer Vision and Image Understanding, vol. 158, pp. 85–105, 2017

[6]. J. Weng, C. Weng, and J. Yuan, "Spatio-Temporal Naive-Bayes Nearest_x0002_Neighbor (ST-NBNN) for Skeleton-Based Action Recognition," in 2017 IEEE on offer once on Computer Vision and Pattern Recognition (CVPR), vol. 017-Jnua. IEEE, July 2017, pp. 445–454.

[7]. P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," Computer Vision and Image Understanding, may 2018.

[8]. Ramezani M. and Yaghmaee F, 2016, "A review on human action analysis in videos for retrieval applications", pp. 485–514.

[9]. Xia L. and Aggarwal J, 2013,"Spatio-temporal depth cuboid similarity features for activity recognition using depth camera".

[10]. Koppula H. S. and Saxena A, 2016, "Anticipating human activities using object affordances for reactive robotic response",pp. 14–29.

[11]. Li K. and Fu Y, 2014, "Prediction of human activity by discovering temporal sequence patterns", pp.1644–1657.

[12]. C. Sch¨uldt, I. Laptev, and B. Caputo Recognizing human actions: Alocal SVM approach, in IEEE ICPR, 2004.

[13]. Blank M., Gorelick L., Shechtman E, 2005, "Actions as space-time shapes",in Proc. ICCV, 2005.

[14]. Weinland D, RonfardR, andBoyer E, 2006, "Free viewpoint action recognitionusing motion history volumes",page no. 2-3.

[15]. RyooM. S. and. K J. Aggarwal, 2010, "UT-Interaction Dataset, ICPRcontest on Semantic Description of Human Activities (SDHA)",http://cvrc.ece.utexas.edu/SDHA2010/HumanInteraction.

[16]. Monfort M., Zhou B, S. A. Bargal, T. Yan, "Momentsin time dataset: one million videos for event understanding".

[17]. Khurram Soomro A. R. Z. and Shah M, 2012, "Ucf101: A dataset of 101human action classes from videos in the wild".

[18]. Kuehne H., Zhuang H., and Serre T, 2011,"HMDB: Alarge video database for human motion recognition".

[19]. Laptev I., MarszalekM, and RozenfeldB, 2008, "Learning realistic human actions from movies".

[20]. W. Li, Z. Zhang, and Z. Liu, Action recognition based on a bag of 3d points, in CVPR workshop, 2010.