Video Analysis of Human Pose Estimation Using Deep Neural And Computer Vision Approaches And Its Networks Applications

Vijay Swaroop A[1], Prekshitha S[2], Zaiba Farheen[3], Supriya P[4]
Atria Institute of Technology

Abstract- *Deep architectures with convolution structure have been found highly effective and commonly used in computer vision. With the introduction of Graphics Processing Unit (GPU) for general purpose issues, there has been an increasing attention towards exploiting GPU processing power for deep learning algorithms. Also, large amount of data online has made possible to train deep neural networks efficiently. The aim of this paper is to perform a systematic mapping study, in order to investigate existing research about implementations of computer vision approaches based on deep learning algorithms and Convolutional Neural Networks (CNN). We selected a total of 119 papers, which were classified according to field of interest, network type, learning paradigm, research and contribution type. Our study demonstrates that this field is a promising area for research. We choose human pose estimation in video frames as a possible computer vision task to explore in our research. After careful studying we propose three different research direction related to: improving existing CNN implementations, using Recurrent Neural Networks (RNNs) for human pose estimation and finally relying on unsupervised learning paradigm to train NNs. We cover key research areas and applications of medical image classification, localization, detection, segmentation, and registration.*

Keywords- convolutional neural network; deep learning algorithm; computer vision; human pose estimation

I. INTRODUCTION

Computers have proven to surpass humans for a variety of tasks, from multiplying large numbers to playing chess. Usually, these certain tasks for which provides a precise rule can be found and even though it seems hard for us humans, it is quite easy for machines, once the respective logic is applied to perform a particular task. Nevertheless, we do perform many tasks to obtain a respective result by performing that particular task, when it comes to tasks like comprehending the world through seeing or listening, what seems trivial for us, is becomes difficult to be implemented by computers.

Based on this premise, we primarily focus our research on computer vision, which in simplicity can be described as there are finding features from various mediums such as, images or videos to help to configure various discriminate objects. From an engineering perspective, it seeks to automate human vision related tasks. One of the major task we are interested in computer vision is human pose estimation. Human pose estimation is related for identifying various human body parts and possibly to track their movements respectively. Real life applications do vary from gaming to augmented and virtual reality to healthcare and last but not least to its gesture recognition. For example, one important aspect we want to try to improve is gesture recognition in sign language videos. In order to be able to translate a particular sign to its respective text, then the first task which we perform to accomplish a respective task is the detection of upper body parts.

Deep learning, a class of machine learning techniques that are used to extract various features from data, and CNN (Convolutional Neural Network), a type of artificial neural network (ANN) that has been extended across space using shared weights, have been found datasets. With the lowered cost of expensive processing hardware, increasing chip processing capabilities and increasing number of data existing online, it was possible to implement deep neural networks in larger data sets and in real-life scenario data sets as well. In particular, Alex Net CNN from Krizhevsky in 2012 has been adopted by the computer vision research community.

II. SYSTEMATIC MAPPING STUDY

Systematic mapping study is one of the method where we use to conduct various mapping study to analyze the respective data, hence, the respective study is been proposed by Peterson. The basic idea is to collect various series of respective publications in the field of their interest, in order to determine the coverage of the respective research field. Systematic mapping study does provides a specific structure for the various type of research reports presented and certain results that have been published successfully by categorizing

them accordingly. There are a number of research questions which are being defined in order to obtain the results in these objectives in a systematic manner. Hence, we choose this study, since its main goals are to respectively present an overview of a certain research area of the respective field and to identify their research gaps accordingly. These are the prior data, what we need to at the beginning of our research.

From our systematic mapping study process, we decided to focus our prior attention on one particular task of computer vision: human pose estimation in video frames. From our systematic mapping study and also from other literature reviews in the same we can say that image analysis has been extensively studied. We are leaning towards video analysis so that we can look at a specter of computer vision that can offer possibility for further study. As for human pose estimation, it represents a task that is present in applications that analyze people. For example: human-computer interaction, gaming (Kinect) or gesture recognition. One case study that we aim is gesture recognition in sign language videos: to understand signs from human upper body movements. As we will see below, in academia and real-life applications, this is an issue which has been tackled, but there is still space for improvement.

In Deep architectures with convolution structure which have been found highly effective and commonly used area, in the field computer vision. Hence, by the introduction of Graphics Processing Unit (GPU) for various general purpose issues, there has been a gradually increasing attention towards exploiting GPU processing power for deep learning algorithms. Also, large amount of data online has made possible to train deep neural networks efficiently. We choose human pose estimation in various video frames as a possible computer vision task to explore in our respective research to perform an efficient result. After careful studying the obtained data we propose three different research direction from the field related to improving existing convolutional neural network (CNN), its respective implementations, using Recurrent Neural Networks (RNNs) for its application on human pose estimation and finally relying its respective data on unsupervised learning paradigm to train neural network (NNs). We cover various key research areas and its various applications which can be applicable on medical image classification, localization, detection, segmentation, and registration. We finally collected a total of around 263 articles altogether, but after the screening process, there were only around 119 articles which have remained. Hence, these were finally classified according to its respective field of interest, various network type which can be implemented, learning paradigm respectively, certainly research and its contribution type. Our relevant study demonstrates that this field of interest is a promising area for research.

If we could not derive enough information, than we read the introduction and conclusions for better understanding. In some cases, even the implementation part was read, in order to comprehend better the work of the respective paper. The classification scheme resulted as the one in. We successfully found out that the majority of research papers dealt with image processing, search approaches around (63 %), used convolutional neural network (CNN) as its respective architecture (65 %), successfully used supervised learning paradigm (55 %) and made use of graphical processing unit (GPU)-acceleration (60 %). Also, we noticed that the majority of papers which have been published in the year 2015 and the number of published papers in the year have been continuously growing from year 2012 and onward. We may say that the respective field of interest is relevant and interesting to the researcher community.

III. LITERATURE SURVEY

Using the systematic mapping study process discussed above, we focused on a single task of computer vision: human pose estimation in video frames. The systematic mapping study and other literature reviews in the field say that image analysis is extensively studied. We are learning video analysis so that we can look at the possible computer vision techniques that can offer further studies. For human pose estimation, it represents a task that is present in applications that analyze people. For example: human-computer interaction, gaming (Kinect) or gesture recognition. In a case study, we aim at the gesture recognition in sign language videos: to understand signs from human upper body movements. In real-life applications, this issue is tackled but there is still a space for improvement.

The best method to recognize human body movements and further classify them according to needs correctly and quickly is questioned right from the 1980's until recent years. The earliest methods tend to solve the issue by placing it as a pattern recognition problem, based on geometric properties such as the relationship between parts.

Pictorial structures are another main method, which models the human body parts as a conditional random field (CRF) – a probabilistic method for structured prediction. Pictorial structure representation was introduced by Fischler and Elschlager thirty years ago, where an object is modeled by a collection of parts arranged in a deformable configuration.

While pictorial structures deal with the parts of a decomposed problem, another method: Random Forests (RF), views the issue as a whole. RFs represent a class of methods in machine learning used for classification or regression, that operate as a collection of decision trees during training.

Computer vision is a majorly researched area in current years. With reference to it, DeepPose is suggested as a new method based on Deep Neural Networks (DNN). DNNs are successful in object localization and classification tasks, yet this paper considers solving the problem of localization of articulated objects. The authors formulate the pose estimation as a joint regression problem.

Chen and Yuille combine the flexibility of graphical models with deep CNNs. Neural Networks are used to learn conditional probabilities for the presence of parts and their spatial relationships within image patches.

Other works have also looked into the problem of estimating human pose and also classifying an activity. For example, CNNs are used to address the regression problem of human joint location estimation. We see that CNNs represent a holistic model, by considering the entire image and not being constrained to a local part.

Pfister, Simonyan, Charles and Zisserman look at human pose estimation in gesture videos. They use once again CNN, which regresses the position of head, shoulder, elbows and wrists. Their work studies the upper body positions in human, with the aim to detect gestures in sign language videos. The inputs to the network are RGB video frames and the outputs - the coordinates of the upper-body joints.

IV. RESEARCH QUESTIONS

Based on the articles we have studied, we propose three different research directions for human estimation task. The first is based on taking in consideration existing architectures of CNNs and what we can improve upon. Another idea is to use another type of CNN architecture for the same task and see how it affects the solution. The last one is to user another type of learning paradigm for human pose estimation in video frames.

We pose three research questions:

A. Since it has been shown that CNNs work for human pose estimation, what could be added or changed in their architecture to improve the results?

Taking in consideration the recent work in dealing with human pose estimation through CNNs, we will look at ways to improve the existing models already successful in this task. Our base model will be the work done by Pfister [15]. Before going in details of what we will improve upon, we will look at three models Pfister proposes. The idea behind his work is to estimate human pose estimation so that the network can detect human gestures in sign language videos.

1) CoordinateNet: In this network, the task of estimating human pose is treated as a regression problem, where the input are RGB video frames and the output are the (x,y) coordinates of the joints.

2) HeatmapNet: The next flavor of implementation of CNNs is a heatmap network. In this case a heatmap of the joints positions are the target of the regression problem. In the beginning of the training process, for a given joint multiple locations may fire, but as the learning continues, the correct predictions prevail. The HeatmapNet performs better than CoordinateNet.

HeatmapNet using optical flow: The idea is to exploit temporal information in videos, by using optical flow to warp pose predictions from neighboring frames. The procedure is to predict joint positions for all neighboring frames and then align them to the specific frame by warping them backwards and forwards using dense optical flow. This method performs better than the previous ones. One problem to focus on is predicting joint coordinates and heatmaps jointly. Experiments have shown that CoordinateNet performs better in the case of shoulders and elbows (more stable position), while HeatmapNet performed better in the case of wrists (highly variable position). The idea is to study the possibility of using both loss targets jointly. For example, losses could be calculated separately for each case and then the weighted average loss could be back propagated through the network. Another issue are the cases where there are multiple modes in the heatmap and the wrong ones are selected. One solution would be to use a spatial model on top of heatmap net.

It could be learned from another convolutional network where the heat maps are used as an input. The higher-level spatial model will be used to remove strong outliers from the output of convolutional networks, which represent the false positives that derive from predictions. A simple method has been used in [16] and we will try to see if this model will further improve HeatmapNet.

B. What would the result be if we used Recurrent Neural Networks to deal with the problem of human pose estimation?

If we refer to the articles [17], [18] we can see examples of previous work that use RNNs respectively for scene labeling and object recognition. For scene labeling, RNN gives the possibility to consider a large input, while limiting the capacity of the model. In the case of object recognition, RNN mimics the visual system recurrent connections. This means that even though the input is static, the activities of the RNN units evolve over time, where each unit depends on its neighbor.

Our idea is to combine ConvNet with RNNs. For each input frame in time (t+1), information from frame in time (t) is passed on. After the heatmap is produced, which can be considered as a spatial model of the probability of the joint locations, we use recurrent layer which memorizes and passes the information on to the next layer.

C. How can we use unsupervised learning to make the most of the amount of unlabeled data that exist online?

Compared to image data domain, there is relatively little work on applying CNNs to video classification. Video is more complex than images since it has another dimension - temporal.

V. APPLICATIONS

To the researcher, CNNs have been put to task for classi - cation, localization, detection, segmentation and registra-tion in image analysis.

A. Classification

Classi cation is sometimes also known as Computer-Aided Diagnosis (CADx). Lo *et al.* described a CNN to detect lung nodules on chest X-rays as far back as 1995 [45]. They used 55 chest x-rays and a CNN with 2 hidden layers to output whether or not a region had a lung nodule. The relative availability of chest x-ray images has likely accelerated deep learning progress in this modality.

B. Localization

Localization of normal anatomy is less likely to interest the practicing clinician although applications may arise in anatomy education. Alternatively, localization may nd use in fully automated end-to-end applications, whereby the radiological image is autonomously analyzed and reported without any human intervention. Yan *et al.* [61] looked at transverse CT image slices and constructed a two stage CNN where the rst stage identi ed local patches, and the second stage discriminated the local patches by various body organs, achieving better results than a standard CNN.

C. Detection

Detection, sometimes known as Computer-Aided Detection (CADe) is a keen area of study as missing a lesion on a scan can have drastic consequences for both the patient and the clinician. The task for the Kaggle Data Science Bowl of 2017 [64] involved the detection of cancerous lung nod-ules on CT lung scans. Approximately 2000 CT scans were released for the competition and the winner Fangzhou [65] achieved a logarithmic loss score of 0.399. Their solution used a 3-D CNN inspired by U-Net architecture [19] to isolate local patches rst for nodule detection.

D. Segmentation

CT and MRI mage segmentation research covers a variety of organs such as liver, prostate and knee cartilage, but a large amount of work has focused on brain segmentation, including tumor segmentation. The latter is especially important in surgical planning to determine the exact boundaries of the tumor in order to direct surgical resection. Sacri cing too much of eloquent brain areas during surgery would cause neurological de cits such as limb weakness, numbness and cognitive impairment. Traditionally, medical anatomical seg-mentation was done by hand, with a clinician drawing out-lines slice by slice through an entire MRI or CT volume stack, therefore it is ideal to implement a solution that automates this laborious task. An excellent review of brain MRI seg-mentation was written by Akkus *et al.* [75], who reviewed various CNN architectures and metrics used in segmentation. Additionally, he also detailed the numerous competitions and their datasets, such as Brain Tumor Segmentation (BRATS), Mild traumatic brain injury outcome prediction (MTOP) and Ischemic Stroke Lesion Segmentation (ISLES).

E. Registration

Although the registration of medical images has many potential applications, which were reviewed by El-Gamal *et al.* [82], their actual clinical use is encountered in niche areas. Image registration is employed in neurosurgery or spinal surgery, to localize a tumor or spinal bony landmark, in order to facilitate surgical tumor removal or spinal screw implant placement. A reference image is aligned to a second image, called a sense image and various similarity measures and reference points are calculated to align the images, which can be 2 or 3-dimensional. The reference image may be a pre-operative MRI brain scan and the sense image may be an intraoperative MRI brain scan done after a rst-pass resection, to determine if there is remnant tumor and if further resection is required. Using MRI brain scans from the OASIS dataset,

Yang *et al.* [83] stacked convolution layers in an encoder-decoder fashion, to predict how an input pixel would morph into its nal con guration

VI. CONCLUSIONS

This paper proposes three research questions related to deep neural networks for video analysis of human pose estimation. The merit of the research is threefold. It gives an overview of state-of-the-art research in the field. It paves the way for further study of video analysis, an area not tackled as much as images by computer vision community. Also, it proposes two models that bring a new development to human pose estimation problem. In the future, we intend to work on the three proposed ideas and implement them for gesture recognition in sign languages videos.

REFERENCES

[1] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA Transactions on Signal and Information Processing, vol. 3, 2014.

[2] J. Schmidhuber, "Deep learning in neural networks: An Overview," Elsevier, 2014.

[3] A. Krizhevsky, S. Ilya and G. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, 2012.

[4] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," Frontiers in Robotics and AI, 2016.

[5] K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson, "Systematic mapping studies in software engineering," in Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, 2008.

[6] D. Forsyth and M. Fleck, "Body plans," in IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[7] P. Buehler, M. Everingham, D. P. Huttenlocher and A. Zisserman, "Upper body detection and tracking in extended signing," International Journal Computer Vision, vol. 95, no. 2, pp. 180-197, 2011.

[8] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," IEEE Transactions on Computers, vol. 22, no. 1, pp. 67-92, 1973.

[9] J. Charles, T. Pfister, M. Everingham and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," International Journal Computer Vision, vol. 110, no. 1, pp. 70-79, 2013.

[10] A. Toshev and C. Szegedy, "DeepPose: human pose estimation via deep neural networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[11] G. Gkioxari, B. Hariharan, R. Girshick and J. Malik, "R-CNNs for pose estimation and action detection," in Computer Vision and Pattern Recognition (cs.CV), 2014.

[12] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in Advances in Neural Information Processing Systems, 2014.

[13] A. Bearman and C. Dong, "Human pose estimation and activity classification using convolutional neural networks," 2015.

[14] T. Pfister, K. Simonyan, J. Charles and A. Zisserman, "Deep convolutional neural networks for efficient pose estimation in gesture," in Asian Conference on Computer Vision (ACCV), 2014.

[15] T. Pfister, Advancing Human Pose and Gesture Recognition, University of Oxford, DPhil Thesis, 2015.

[16] A. Jain, J. Tompson, M. Andriluka, G. Taylor and C. Bregler, "Learning human pose estimation features with convolutional networks," in Computer Vision and Pattern Recognition, 2013.

[17] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, 2015.

[18] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in International Conference on Machine Learning, 2014.

[19] G. W. Taylor, R. Fergus, Y. LeCun and C. Bregler, "Convolutional learning of spatio-temporal features," in EECV'10 Proceedings of the 11th European conference on Computer Vision , 2010.

[20] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Computer Vision and Pattern Recognition, 2016.

[21] N. Srivastava, E. Mansimov and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in Computer Vision - ECCV 2016, Amsterdam, 2016.