Computer Vision Approaches based on Deep Learning and Neural Networks:

Deep Neural Networks for Video Analysis of Human Pose Estimation

Eralda Nishani
Faculty of Contemporary Sciences and Technologies
SEEU, Tetovo, Macedonia
en24236@seeu.edu.mk

Betim Çiço
Department of Computer Engineering
EPOKA University, Tirana, Albania
bcico@epoka.edu.al

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.**

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

I. INTRODUCTION

Computers have proven to surpass humans for a variety tasks, from multiplying large numbers to playing chess. Usually, these are tasks for which a precise rule can be found and even though it seems hard for us humans, it is quite easy for machines, once the logic is applied. Nevertheless, when it comes to tasks like comprehending the world through seeing or listening, what seems trivial for us, is difficult to be implemented by computers.

Based on this premise, we focus our research on computer vision, which in simplicity can be described as finding features from images or videos to help discriminate objects. From an engineering perspective, it seeks to automate human vision related tasks. One task we are interested in is human pose estimation. It is related to identifying human body parts and possibly track their movements. Real life applications vary from gaming to augmented and virtual reality, to healthcare and last but not least to 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 sign into a text, at first detection of upper body parts is needed.

Deep learning, a class of machine learning techniques that are used to extract features from data, and CNN (Convolutional Neural Network), a type of artificial neural network that has been extended across space using shared weights, have been found suitable for computer vision tasks [1], [2]. At the beginning, researchers experimented with small 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, AlexNet CNN from Krizhevsky in 2012 [3] has been adopted by the computer vision research community [4].

In this paper, we look at three directions to improve the performance of human pose estimation through CNNs: improve on existing solutions in this area, implementing deep learning algorithm on a different kind of architecture – Recurrent Neural Networks (RNN) and study the use of unsupervised learning paradigm to train the network.

II. SYSTEMATIC MAPPING STUDY

The method that we use to conduct the mapping study is based on the systematic mapping study proposed by Petersen [5]. The idea is to collect a series of publications in the interested field, in order to determine the coverage of the research field. Systematic mapping study provides a structure of the type of research reports and results that have been published by categorizing them.

A number of research questions are defined in order to obtain these objectives in a systematic manner. We choose this study, since its main goals are to present an overview of a certain research area and to identify research gaps. This is what we need to at the beginning of our research.

We collected a total of 263 articles, but after the screening process only 119 articles remained. To classify the papers and determine the keywords, we used the abstracts as the main source.

978-1-5090-6742-8/17/$31.00 ©2017 IEEE

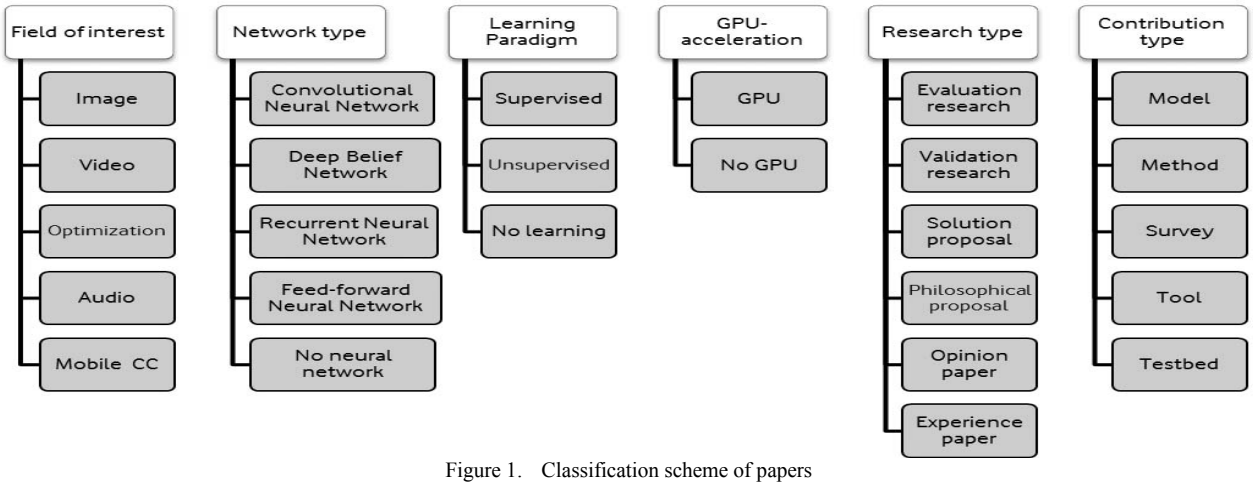

Figure 1. Classification scheme of papers

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 Fig. 1.

We found out that the majority of papers dealt with image search approaches (63 %), used CNN as architecture (65 %), used supervised learning paradigm (55 %) and made use of GPU-acceleration (60 %). Also, we noticed that the majority of papers have been published in 2015 and the number of published papers has been growing from 2012 and onward. We may say that the field is relevant and interesting to the researcher community.

III. LITERATURE REVIEW

From our systematic mapping study process, we decided to focus our 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 field [1], [2] 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.

A. Background

Starting from 1980's until recent years, the question has been which is the best method to correctly and quickly recognize human body movements and further classify them according to need. 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. For example, in [6] researchers build what they call a 'body plan', adapted to segmentation and recognition in complex environments. They suggest that the body plan can be fixed or can be learned using statistical learning techniques.

Another main method are pictorial structures, which model the body parts of a human as a conditional random field (CRF) – a probabilistic method for structured prediction as in [7]. Pictorial structure representation were introduced by Fischler and Elschlager thirty years ago [8], where an object is modeled by a collection of parts arranged in a deformable configuration.

While pictorial structures tend to deal with the parts of a decomposed problem, another method: Random Forests (RF), views the issue as a whole.

RFs [9] represent a class of methods in machine learning used for classification or regression, that operate as a collection of decision trees during training.

B. State-of-the-art

There is a lot of work done about computer vision related tasks in current years. In reference [10], DeepPose is suggested as a new method based on Deep Neural Networks (DNN). Even though DNNs have been successful in object localization and classification tasks, the paper tries to solve the problem of localization of articulated objects. The authors propose a cascade of DNN-based pose predictors. They formulate the pose estimation as a joint regression problem.

Another work in the field [11] chooses CNNs to recognize a person in an image, to estimate the pose and classify the action. In this case we see a broad category of problems involved in the CNN implementation, which are trained jointly for multiple tasks.

Chen and Yuille [12] combine the flexibility of graphical models with deep CNNs. NNs 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, two Stanford students [13] use CNNs to address the regression problem of human joint location estimation. Once again, we see that CNNs represent a holistic model, by taking in consideration the entire image and not being constrained to a local part.

Pfister, Simonyan, Charles and Zisserman [14] 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 input 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.

3) 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

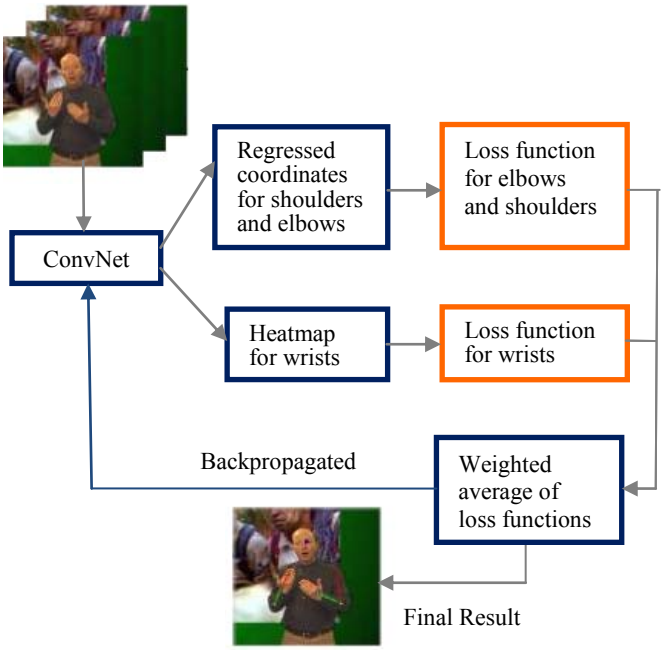

Figure 2. Our proposed model for improving on existent Convolutional Neural Network architectures

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 backpropagated through the network, as can be seen on Fig. 2. 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. A model

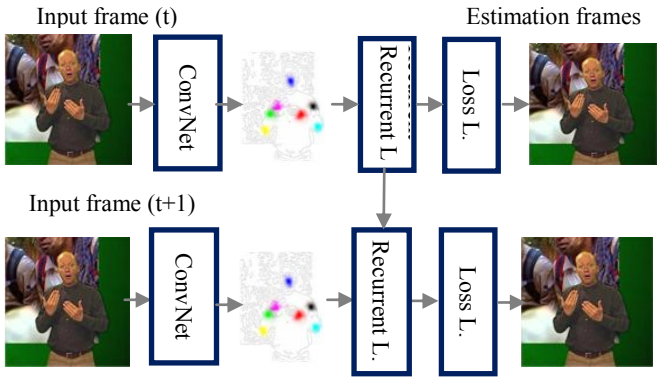

Figure 3. Using Recurrent Neural Networks to deal with human pose estimation

that we propose can be seen in Fig. 3. 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 [19]. Video is more complex than images since it has another dimension - temporal.

Unsupervised learning exploits temporal dependencies between frames and has proven successful for video analysis. Some extensions of CNNs into the video domain have been explored. One way is to fuse the features of different CNNs, responsible for spatial and temporal stream [20].

One problem encountered in video analysis, is the fact that videos contain high dimensionality and more labeled data are needed for credit assignment. Finding proper labeled data in the training phase is not always easy and is time-consuming. This gives us another reason to think that unsupervised learning may be fitting to video analysis and we can exploit it for our purpose. We propose a solution that will be adapted from [21], where an encoder-decoder RNN architecture is used to learn representations of video sequences for the task of action recognition.

V. 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.