Pose Based Action Recognition using Hierarchical Bidirectional Long Short Term Memory Network

Aditya Agarwal*
Bipasha Sen*
adiagar@microsoft.com
bise@microsoft.com
Microsoft

ABSTRACT

Human body can be represented as an articulation of rigid and hinged joints which can be combined to form the parts of the body. Human actions can be thought of as a collective action of these parts. Hence, learning an effective spatio-temporal representation of the collective motion of these parts is key to action recognition. In this work, we propose an end-to-end pipeline for the task of human action recognition on video sequences using 2D joint trajectories estimated from a pose estimation framework. We use a Hierarchical Bidirectional Long Short Term Memory Network (HBLSTM) to model the spatio-temporal dependencies of the motion by fusing the pose based joint trajectories in a part based hierarchical fashion. Experimental results demonstrate that our method outperforms the existing state of the art on the widely used KTH dataset.

KEYWORDS

Action Recognition, Pose Estimation, Hierarchical BLSTM

1 INTRODUCTION

Human action recognition is a prominent field of research in computer vision with its wide applications in the areas of robotics, intelligent surveillance systems, automated driving etc. Despite the extensive research on action recognition in the vision community, the problem still poses a significant challenge due to large variations and complexities, e.g., occlusion, low frame-rate, camera angle and motion, illumination, cluttered background, and so on.

Traditionally, the spatio-temporal structure has been modeled using handcrafted features and the actions are recognized using well defined discriminative networks [6, 16, 21, 23, 30]. Despite encouraging results for action recognition on several datasets, these approaches suffer from variations of view point and scale, subject and appearance and so on. Recent advances in human pose estimation using deep learning [2–4, 13, 20, 29, 31] and the availability of depth sensors [22, 26, 27, 32] have led to accurate representations of high level features. Studies show that high-level features extracted using current pose estimation algorithms already outperform state of the art low level representations based on hand crafted features implicating their potential in action recognition. In this work, we use a pose estimation framework based on Convolutional Neural Networks (CNN) in tandem with a robust object detection framework to deal with variations in scale and viewpoint to obtain a 2D representation of joint locations. The object detection algorithm filters frames that do not contain the object of interest and are therefore non-discriminative for the task of action recognition.

*Both authors contributed equally to this research.

Human body can be articulated as a system of rigid and hinged joints. These joints can be combined to form the limbs and the trunk. Human actions can be thought of as a collective action of these limbs and the trunk. Human action recognition is considered a time series problem where the characteristics of the body posture and its dynamics are extracted over time to represent the action [12, 14, 34]. [33] proposed a hierarchical approach on a trajectory of 2D skeleton joint coordinates. In this work, we propose an end-to-end part based hierarchical action recognition pipeline on raw video sequences. We use Convolutional Pose Machines (CPM) to estimate a trajectory of 2D joint coordinates. These are combined in a part based hierarchical fashion using Hierarchical Bidirectional Long Short Term Memory (HBLSTM) networks to estimate the spatio-temporal dependencies in the video sequence. The encoded representation of the video sequence is then fed to a discriminative network to classify the action.

The major contributions of this work are two-fold,

(1) Designing an end to end pipeline for pose based action recognition using a part-based hierarchical approach on raw video sequences. We handle the common issues of occlusion, camera angle variations and eliminate non-discriminative frames while learning robust joint coordinates.
(2) We obtain state of the art results on the widely used KTH action recognition dataset [25] that can serve as a baseline for future developments.

2 PIPELINE

The pipeline consists of three modules:

(1) Bounding box detection using YOLOv3 [24]
(2) Pose estimation using CPMs [31]
(3) Action Classification using HBLSTM

The proposed pipeline works in a supervised fashion. Frames of fixed dimension $\in R^{160\times120}$ are extracted from the input video sequence. An object detection algorithm is used to obtain an initial estimate of human's presence in the image. We use a pose estimation framework based on CNNs to generate an estimation of human joints in the extracted bounding boxes. 2D coordinates of 14 joint locations (head, neck, wrist, elbow, shoulder, hip, knee and ankle) are extracted for every frame representing the joint trajectories for the entire video sequence.

The proposed model is based on HBLSTMs that takes joint trajectories as input. HBLSTMs can learn across multiple levels of temporal hierarchy. As shown in Fig. 1., the first recurrent layer encodes the representation of 5 body parts, namely, Right Hand (RH), Left Hand (LH), Right Leg (RL), Left Leg (LL) and Trunk (T).

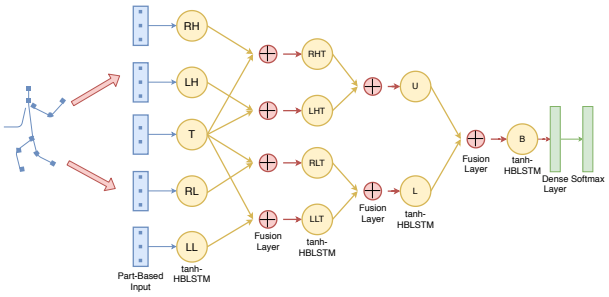

Figure 1: Proposed Part Based Hierarchical Architecture.

The next set of layers encodes the part representations into Upper Right (RHT), Upper Left (LHT), Lower Right (RLT) and Lower Left (LLT) vectors by fusing the encoded representation of T with RH, LH, RL and LL respectively. The subsequent layers generate the encoded representation of Upper (U) and Lower (L) bodies followed by the encoded representation of the entire body. Finally a dense layer followed by a softmax layer are added to classify the action.

2.1 Bounding Box Detection

We use an object detection algorithm to obtain an initial estimate of human's presence in the image. YOLOv3 [24] is an efficient object detection algorithm pretrained on the ImageNet [7] and MSCOCO [18] datasets. It uses 53 successive 3×3 and 1×1 convolutional layers. The input to the bounding box algorithm are the grey scaled image frames of fixed dimension $\in R^{160 \times 120}$ extracted from the video sequences in the KTH action recognition dataset.

2.2 Pose Estimation

Convolutional Pose Machines (CPMs) [31] were introduced for the task of articulated pose estimation. CPMs consist of a sequence of convolutional neural networks that repeatedly produce 2D belief maps for the location of each part. At each stage in a CPM, image features and belief maps produced in the previous stage are used as inputs producing increasingly refined locations of each part (Eq. 1).

$$g_t(x'_z, \psi_t(z, b_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p \in \{0...P+1\}} \tag{1}$$

We operate CPMs directly on the bounding boxes generated from the previous stage to produce joint coordinates.

2.3 Hierarchical Bidirectional Long Short Term Memory

We denote a video sequence as $X = [x_1, ..., x_T]$, with each frame x_t denoting the $2D$ coordinates of 14 joints. We work in a supervised classification setting with a training set,

$$\chi = \{(X_i, y_i)\}_{i=1}^{N} \in R^{T \times 28 \times 1} \times \{1, ..., C\} \tag{2}$$

where X_i is a training video sequence and y_i is its class label (from one of the C possible classes).

Human body can be decomposed into five parts - two arms, two legs and a trunk and the global action can be modeled as the collective motion of these five parts. Benefitting from the LSTM's ability to model contextual dependencies from temporal sequences, we propose a Hierarchical Bidirectional LSTM (HBLSTM) for the task of pose based action recognition (Fig. 1). For the first layer, given the inputs $I_{i,j}^t$ as trajectories of part j at i^{th} layer for time t, the corresponding encoded part representation is expressed as

$$h_{i,j}^t = \overrightarrow{h_{i,j}^t} \oplus \overleftarrow{h_{i,j}^t} \tag{3}$$

where $\overrightarrow{h_{i,j}^t}$ and $\overleftarrow{h_{i,j}^t}$ are the forward and backward layers passes respectively with tanh activations.

For the fusion layer at time t, the newly fused pth representation as the input for the $(i+1)^{th}$ encoding layer is:

$$I_{i+1,p}^t = h_{i,j}^t + h_{i,k}^t \tag{4}$$

where $h_{i,j}^t$ are the concatenated hidden representations of the j^{th} part in i^{th} encoding layer and $h_{i,k}^t$ is for the k^{th} part in i^{th} layer.

The encoded representation of the entire body is given by the T^{th} unit of the last encoding layer, $h_{4,body}^T$, which is given as input to the dense layer. The output of the dense layer is expressed as:

$$O = v_{h_{4,body}^t} \cdot h_{4,body}^T + b_{h_{4,body}^t} \tag{5}$$

O is given as input to the softmax layer.

2.4 Evaluation Results

Similar to many evaluation approaches on the KTH dataset [25], we carried out our experiments using a leave-one-out cross validation strategy [10] (i.e. all subjects except one were used for training and the learned model was evaluated on the remaining one). Average accuracy on the comparative architectures and the proposed model are reported along with the existing approaches in Table 1.

Table 1: Recognition rates against different approaches on the KTH dataset.

Methods	Accuracy
Schuldt *et al.* [25]	71.72%
Dollar *et al.* [8]	81.17%
Fathi *et al.* [9]	90.5%
Laptev *et al.* [17]	91.8%
Baccouche *et al.* [1]	94.39%
Chen and Hauptman [5]	95.83%
Gilbert *et al.* [11]	96.7%
Mona *et al.* [19]	97.89%
Proposed Approach	**99.3%**

We achieve almost full separation between different actions and a slight misclassification occurs between similar actions i.e. between "hand-clapping" and "hand-waving" and between "running" and "jogging". The classification accuracy is averaged over all selections of test data to achieve a recognition rate of **99.3%** using the proposed approach. Although the results are not directly comparable due to different evaluation strategies being adopted by different researchers, we show that our proposed approach outperforms the state of the art on the KTH dataset for the task of action recognition.

3 CONCLUSION AND FUTURE WORK

In this work, we present an end-to-end pipeline for the task of human action recognition in videos. Using an object detection approach, we first estimate the presence of human and discard remaining frames as they are non-discriminative for the task of action recognition. We use a pose estimation framework to generate the trajectory of joint coordinates and combine them in a part-based hierarchical fashion to obtain a global representation of the entire video sequence. We showed that adding the part based hierarchical fusion helps us achieve better results over other comparative architectures. Experimental evaluation on the KTH action recognition dataset shows that our proposed approach outperforms the existing state of the art approaches on this dataset.

As our proposed approach shows great promise on the KTH action recognition dataset, we are currently extending this approach on larger datasets like [28] and HMDB51 [15].

REFERENCES

[1] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. 2011. Sequential Deep Learning for Human Action Recognition. In *Human Behavior Understanding*, Albert Ali Salah and Bruno Lepri (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 29–39.

[2] E. Brau and H. Jiang. 2016. 3D Human Pose Estimation via Deep Learning from 2D Annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*. 582–591.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs.CV]

[4] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2015. Human Pose Estimation with Iterative Error Feedback. arXiv:1507.06550 [cs.CV]

[5] Mingyu Chen and Alexander G. Hauptmann. 2009. MoSIFT : Recognizing Human Actions in Surveillance Videos CMU-CS-09-161.

[6] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human Detection Using Oriented Histograms of Flow and Appearance. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 428–441.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. 65–72.

[9] A. Fathi and G. Mori. 2008. Action recognition by learning mid-level motion features. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[10] Zan Gao, Ming-yu Chen, Alexander G. Hauptmann, and Anni Cai. 2010. Comparing Evaluation Protocols on the KTH Dataset. In *Human Behavior Understanding*, Albert Ali Salah, Theo Gevers, Nicu Sebe, and Alessandro Vinciarelli (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 88–100.

[11] Andrew Gilbert, John Illingworth, and Richard Bowden. 2009. Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features. 925 – 931. https://doi.org/10.1109/ICCV.2009.5459335

[12] D. Gong, G. Medioni, and X. Zhao. 2014. Structured Time Series Analysis for Human Action Segmentation and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1414–1427.

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. arXiv:1802.00434 [cs.CV]

[14] Hyesuk Kim and Incheol Kim. 2015. Human Activity Recognition as Time-Series Analysis. *Mathematical Problems in Engineering* 2015 (10 2015), 1–9. https://doi.org/10.1155/2015/676090

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*. 2556–2563.

[16] Laptev and Lindeberg. 2003. Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision*. 432–439 vol.1.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]

[19] Mona M. Moussa, Elsayed Hamayed, Magda B. Fayek, and Heba A. El Nemr. 2015. An enhanced method for human action recognition. *Journal of Advanced Research* 6, 2 (2015), 163 – 169. https://doi.org/10.1016/j.jare.2013.11.007

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. arXiv:1603.06937 [cs.CV]

[21] Janez Perš, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec, and Stanislav Kovačič. 2010. Histograms of optical flow for efficient representation of body motion. *Pattern Recognition Letters* 31, 11 (2010), 1369 – 1376. https://doi.org/10.1016/j.patrec.2010.03.024

[22] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. 2010. Real-time identification and localization of body parts from depth images. In *2010 IEEE International Conference on Robotics and Automation*. 3108–3113.

[23] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28, 6 (2010), 976 – 990. https://doi.org/10.1016/j.imavis.2009.11.014

[24] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs.CV]

[25] C. Schuldt, I. Laptev, and B. Caputo. 2004. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. 32–36 Vol.3.

[26] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. 1297–1304.

[27] M. Siddiqui and G. Medioni. 2010. Human pose estimation from a single view point, real-time range sensor. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 1–8.

[28] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs.CV]

[29] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Jun 2014). https://doi.org/10.1109/cvpr.2014.214

[30] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2015). https://doi.org/10.1109/cvpr.2015.7299059

[31] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. arXiv:1602.00134 [cs.CV]

[32] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. 2013. *A Survey on Human Motion Analysis from Depth Data*. 149–187. https://doi.org/10.1007/978-3-642-44964-2_8

[33] Yong Du, W. Wang, and L. Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1110–1118.

[34] Yong Zhang, Yu Zhang, Zhao Zhang, Jie Bao, and Yunpeng Song. 2018. Human activity recognition based on time series analysis using U-Net. arXiv:1809.08113 [cs.LG]