Human Activity Recognition from Multimodal Data

Pengyin Chen, Arial Qin, Junhao Lu

Human Activity Recognition from Multimodal Data

Pengyin Chen
Department of Computing Science
University of Alberta
Edmonton, Canada
pengyin@ualberta.ca

Arial Xiao Qin
Department of Computing Science
University of Alberta
Edmonton, Canada
xqin5@ualberta.ca

Junhao Lu
Department of Computing Science
University of Alberta
Edmonton, Canada
junhao3@ualberta.ca

***Index Terms*—Human Activity, Recognition and Protection**

I. Introduction

In recent years, the interest for 3D human activity analysis, with the amount of applications developed has increased rapidly. People are shifting from 2D image identification to 3D motion activity recognition. [1] Human activity recognition (HAR) is now an important research area in human-to-human interactions. The capacity of humans to detect other people's actions is a major research topic in computer vision and machine learning. As a result of this study, numerous applications such as video surveillance, human-computer interaction, and human behavior modeling require a multiple activity detection system. [2]

The majority of research on HAR is based on the assumption of a figure-centric scenario with an uncluttered background in which the actor is free to execute an activity. The construction of a completely automated system capable of accurately identifying a person's actions is a difficult endeavor due to issues such as background clutter, partial occlusion, changes in scale, perspective, lighting and appearance, and frame resolution. [3] Additionally, annotating behavioral roles takes time and requires prior understanding of the occurrence. What's more, the challenge is made more difficult by intraclass and interclass similarities. That is, acts within the same class may be communicated differently by various persons using distinct bodily motions, whereas actions within classes may be difficult to discern due to the similarity of the information represented. The manner in which humans conduct an activity is determined by their habits, which complicates the task of determining the underlying activity. Lastly, developing a visual model for learning and understanding human motions in real time with insufficient benchmark datasets is a difficult undertaking. [4]

Single person HAR can be implemented through traditional object tracking approaches such as KLT (Kanade–Lucas–Tomasi feature tracker). In earlier years Anup et al.(1999) proposed a method to implement human face tracking with an active camera. It first detects the head region with a frame motion fusion algorithm, then uses deformable template matching and color information to detect facial features, and finally applies a 3D wireframe to track the moving face. Anup et al.(2005) in later years proposed a method to add both Gaussian and LoG weighting functions to increase the tracking performance of KLT with random noise. [5] For machine learning based approaches, single-person HAR mostly rely on linear discriminant functions, which are notoriously difficult to get a decent detection results on a large number of complicated and comparable data, as is the case with deep learning methods. Deep learning based methods can obtain features with richer semantic information, multi-scale and multi-type human joint point feature vectors. And this method has the ability to get rid of the dependence on the structural design of the component model. [6] In this literature review, we are going to summarize the progress in singlet person HAR methods including CNN-based, GAN-based, and component model approaches.

Identify applicable funding agency here. If none, delete this.

II. Literature review

A. HAR Based on Convolutional Neural Networks (CNN)

There are two types of approaches for estimating human posture using CNN: detection-based. [7], [8] and regression-based. [9], [10] A probability distribution map is used as a heatmap to show the joint points' locations in the detection-based technique. The probability value for each pixel may also be estimated using this approach. For this approach, the higher the probability value, the closer the pixel is to the actual joint location, and vice versa. Joint points are expressed as two-dimensional coordinates in the regression-based technique, which is trained by learning how body characteristics are linked to a person's body position and then directly derives the coordinates of each joint point.

When Toshev et al. utilized AlexNet as the fundamental network structure to regress the joint coordinates, they used the cascade network to first compute the coordinates of a joint point and then re-obtained the original picture using this coordinate. Calculate coordinates with a better degree of precision. Although the cascade network enhances the accuracy of the HAR estimation regression network, it is not suited for the scenario when the resolution of the input picture is rather low. It is also time-consuming for performing

numerous convolution processes on a single picture because of the cascade approach.

While DeepPose does not account for local appearance in pose estimation, and it is challenging to estimate complicated human postures, this paper proposes a novel approach to solving these challenges. Fan and colleagues suggested a dual-source deep convolutional neural network technique for human position estimation (DS-CNN). [11] R-CNN is extended to a DS-CNN model, which offers researchers with global view pose estimation, a new feature of the approach When compared to sliding windows or entire pictures used as input in many earlier techniques of estimating human position, DS-CNN is trained using a series of class-independent image patches recognised from the input image. To better represent the semantic significance of local bodily components. The author uses DS-CNN unified learning to accomplish joint detection, decide if the image block contains human joints and joint locations, and locate the image block by taking the local (body) part image and the overall view of each local part as independent inputs, where the midway joint actually is. More accurate human posture estimate is achieved by combining the joint detection/localization findings obtained from all picture patches together. Researchers have developed a novel approach for joint point localization utilizing optical flow in deep convolutional networks, which uses a heatmap to represent pose estimation as a detection issue. [12] It combines information from many frames using optical flow to provide context.

The location estimation of occluded body components in the HAR issue is a challenging one to tackle. Nathaniel and his team created a new pipeline for pose estimation with multisensor, which includes a pre-trained classifier to measure which kind of position could lead to higher estimation error. [13] Bulat et al. developed a detection-follow-regression cascade network. Joint locations in the human body may be detected using the network's cascade network, which has two sub-networks. [14] In order to determine the location of each joint on the body, the part detection network is employed. [15] A heatmap of occluded joints, on the other hand, does not have a high degree of certainty. Because the original images are stacked with the heatmap output by the part detection network and fed into a deep regression network, which predicts the location of occluded joint points by regressing a set of confidence maps near the true location of the occluded joint points, further improving the detector's robustness in detecting occlusion of specific body parts.

While analyzing live captured data, efficient data compression is also a key performance factor to quickly receive and process large amounts of data. In earlier years, Anup et. al. has applied KL (Karhunen–Loève) transformation to compress MPEG video to address the issue of large amount of uncompressed video data. [16] Anup et. al. later proposed a novel approach for lossy compression of motion capture data, that is both faster and able to get down to a compression ratio of 25:1 with small impact on the actual quality of the data. [17] The quality of the video captured by video cameras is another key factor to be considered. The general camera produced videos are usually 30 fps, Magnetic resonance imaging (MRI) can only capture 6 - 7 fps. Anup et. al. proposed a method to use event dynamics and ED matching to reconstruct the video into higher resolution. [18]

Chu et al. introduced a multi-context attention model for the first time, which split the picture into two components, the human body and the human body's local joints. [19] Attention maps with various resolution characteristics (e.g., hourglass networks) may be generated, and the various resolution features correlate to various meanings. CNNs are coupled with a multi-context attention mechanism in order to recognise human action from beginning to conclusion in real time. This system (UniPose) was proposed by Artacho, et al. using the "Waterfall" Shrinking Spatial Pooling Architecture (WASP). [20] Unified Pose does not require distinct branches for the detection of bounding boxes and joints, combining context segmentation and joint localization methods in order to compute the locations of joint points and bounding boxes for human body detection simultaneously. Without the use of statistical post-processing procedures, a stage may accurately assess a human's stance. When comparing the waterfall atrous spatial pooling (WASP) module with the atrous spatial pyramid pooling (ASPP) component of the framework, it is shown that the cascade approach of atrous convolution yields a more parallel structure. [21] Another approach is predicting joint locations based on contextual information. It includes information about the full frame and does not require post-analysis using statistical or geometric approaches. The UniPose-LSTM approach has also been applied to multi-frame processing in related literature, with outstanding results in video pose estimation.

B. HAR Based on Generative Adversarial Networks (GAN)

Chou et al. offered a method employing generative adversarial networks to address the lack of effective feature representation for human activity recognition. [22] It creates two layered hourglass network frameworks, generator and discriminator, with the same architecture. The generator network is a fully convolutional network that includes residual blocks. Following the passage of the input picture through the generator, a collection of heat maps indicating the confidence score of each joint point at each location is produced. The heatmap generated by the generator network and the real heatmap are fed into the discriminator network, and the two sets of heatmaps are reconstructed to calculate the loss between the discriminator output and the real heatmap and the heat map generated by the generator, using the notation $Lfake$ and $Lreal$, respectively.

Human joint structure prediction must be addressed when estimating body parts that are severely occluded by adjacent body parts or that appear obstructed by backdrops comparable to body parts. The solution to this challenge lies in learning the real body joint point distribution from a huge amount of training data. Directly learning such structures, on the other hand, might occasionally result in approximated postures

that are biologically inexplicable. Chen et al. incorporated information of human body structure, suggested a structure-aware convolutional network to estimate occluded body components, and trained pose generation networks using a new conditional adversarial network (adversarial PoseNet) with two discriminators. [23] When the predicted part is occluded, the pose generation network G generates heatmaps with low confidence, and the confidence discrimination network C marks these heatmaps as fakes, forcing the pose generation network G to generate heatmaps with higher confidence, improving the accuracy of positioning the occluded parts.

Another difficulty with human activity identification is scale instability, which occurs when the input bounding box of the human detector is slightly perturbed, causing the activity detection result to vary. For the unstable scale of human activity recognition, the basic hourglass structure approach will overfit the body joints at a given scale, resulting in scale "domination" on a single scale. The approach is to repeat training at several scales, do posture estimation, and output the most accurate results. However this technique lacks a consistent scale representation. Yang et al. also observed changes in the shape and perspective of the human body, as well as inconsistency in the proportions of body parts, which made it difficult for the body part detector to correctly locate the joints of body parts, and proposed a pyramid residual model (PRM), which is a stacked hourglass network that serves as the basic network structure for constructing a multi-scale feature pyramid with learning function with the goal of enhancing the scale invariance of a feature pyramid. [24] Ke et al. suggested an approach based on a multi-scale structure-aware neural network that merged a multi-scale supervised network (MSS-Net) and a multi-scale regression network (MSR-Net) to match multi-scale characteristics and enhance joint point localization's resilience. Both multi-scale supervised networks and multi-scale regression networks employ structure-aware losses to learn human skeletal structural elements from multi-scale data. [25] These priors can be quite useful for recovering occluded body parts in complicated scenarios. This approach, which is a close combination of multi-scale supervision and regression network, locates joint points well and does global HAR via the structural link between numerous joint points. Moreover, Ashraf and his team proposed a technique for detecting moving objects from a moving camera using the background constraint. Motion is detected by computing a mapping that corresponds to pixels in successive images whenever possible. [26] This idea is also inspiring and could be utilized in reducing the influence of scale instability. Another method to eliminate distortions is also introduced by Sergio and his team, which has a good performance dealing with fish lenses. [27]

C. HAR Based on Component Model

Component models have been used in a variety of human activity identification systems during the last decade [28]–[30]. The human body composition model entails the representation of the entire human body as a hierarchical structure of parts and sub-parts, with certain joint constraints met. Combination models include a set of discrete variables to represent the compatibility of components, including information about the parts' orientation and size, as well as information about semantic classifications such as straight arm and curved arm. Due to the vast number of possible combinations of various component types and their subcomponent types, the resultant state space of higher-level components can increase exponentially, which is computationally and memory intensive. In earlier years, Xiaobo and Anup proposed a variable resolution approach for character thinning for preprocessing the data set. The goal is to greatly downsample the data and remain the skeleton of the human in the data source, that aims to solve the possible performance issue. [31] Tang et al. proposed a deep learning component model (DLCM) for HAR based on learning the complex structural relationships of the human body, as well as a spatial localization model based on the composition of human bones, in order to address this issue to the greatest extent possible. [32] A spatial local information summary (SLIS) representation model comprises of 16, 12, and 6-component combination models at three semantic levels. It correctly records the scale, direction, and shape of each component to avoid improper component combinations. In comparison to prior HAR networks, DLCM is hierarchical across several semantic levels, and human HAR estimation follows a bottom-up/top-down network stage similar to multi-person HAR estimation. Additionally, none of the prior CNN-based structural models for human activity identification deconstructed things into meaningful and reusable hierarchies or inferred between different semantic levels.

DLCM learns the compositional relationships between body parts through the network. It has a hierarchical compositional architecture and bottom-up/top-down inference stages. In the bottom-up stage, the heatmaps of the target joints are directly regressed from the image, and the heatmaps of higher-level parts are recursively derived through the body joint sub-nodes. In the top-down stage, the heatmaps of the lower-level parts are described by recursive refinement using their associated parent joint score maps and the self-score maps of the bottom-up stage; using the mean squared error(MSE) [33] as loss to distinguish the predicted heatmap from the real heatmap to guide the network to learn the correct connection relationship between body parts. On the FLIC dataset, the accuracy of DLCM in wrist recognition is improved by 1.5% compared with the literature [25]. On the MPII dataset, the accuracy increases by 2.6%, 2.0%, 1.7%, 1.6% and 1.4% on the ankle, knee, hip, wrist and elbow, respectively. In both datasets, the accuracy of this method surpasses that of the previous state-of-the-art methods, while the 3-level DLCM has significantly less amount of parameters and lower computational complexity. Moreover, Mark and his team also proposed a new method to solve the feature detection in non-SVP (single viewpoint) situations. [34] They represent a new mathematical model to detect features in panoramic non-SVP images using a modified Hough transform and significantly improve the performance in identifying line features with only estimated calibration. A

strong feature extraction could greatly improve the accuracy of localizing joints and parts. Similarly, Meghna and his team also introduce a method using Radon transform. [35] They create a way to present the human skeleton as a binary vector and utilize parametric Radon transform to extract pose features, which generates maximum corresponding to specific orientations of the skeletal representation. Anup et al. proposed a novel approach for ground air traffic human hand gestures recognition. [36] It also uses Radon transform to obtain parametric representation of the input images. Restrictions are applied to focus on the upper body of the movement only, along with background separation to further increase the accuracy of motion detection. The result indicates an 88% accuracy and has an advantage over other traditional template based approaches in terms of performance and the necessity to normalize the image. Anup et al. proposed a method that utilizes color information to improve the accuracy of movement track. [37] The paper discusses human eye movements tracking with Hough transform and template matching, combined with color information. The results showed that it is able to even track the movement of iris and lids. The color information makes the model more robust. Anup et al. applied similar techniques for nose shape estimation, tracking and facial expression prediction. [38] Each part of the human face (such as eye, nose) has a distinguishable darkness, compared to the human face skin, which is smooth. This further improves the human facial expression tracking accuracy on top of the existing models.

III. Conclusion

The single-person HAR approach based on deep learning is a more fundamental technique for recognising human activities. The regression-based technique learns the mapping between body features and their positions by extracting picture characteristics at several resolutions. Then, it directly derives the coordinates of each joint point using a multi-stage network. However, this approach of directly regressing the coordinates of joint points is insufficient for learning the structural information associated with the human body's joint points. The output heat map approach learns the structural information of joint points by the construction of a probabilistic graphical model or the use of multi-scale receptive fields, and then derives correct joint point coordinates via the planned cascade network. Due to the fact that this technique completely understands the failure process of joint point localization, the network outputs the confidence in several spatial places rather than just one, which makes learning more three-dimensional. As a result, the majority of commonly used HAR algorithms are based on heat map detection.

References

[1] S. Berretti, M. Daoudi, P. Turaga, and A. Basu. Representation, analysis, and recognition of 3d humans: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 14(1s), 2018.

[2] Michalis Vrigkas, Christophoros Nikou, and Ioannis A. Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2, 2015.

[3] Zhu Jiadong, San-Segundo Rubén, and Pardo José M. Feature extraction for robust physical activity recognition. *Human-Centric Computing and Information Sciences*, 7(1):1 – 16, 2017.

[4] Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, and Arousha Haghighian Roudsari. Human pose, hand and mesh estimation using deep learning: a survey. *The Journal of Supercomputing: An International Journal of High-Performance Computer Design, Analysis, and Use*, pages 1 – 39, 2022.

[5] L. Yin and A. Basu. Integrating active face tracking with model based coding. *Pattern Recognition Letters*, 20(6):651–657, 1999.

[6] Neil Robertson and Ian Reid. A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104(2):232–248, 2006. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[7] Xianjie Chen and Alan Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. 2014.

[8] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. 2019.

[9] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. 2013.

[10] T. (1) Pfister, K. (1) Simonyan, A. (1) Zisserman, and J. (2) Charles. *Deep convolutional neural networks for efficient pose estimation in gesture videos.*, volume 9003 of *Lecture Notes in Computer Science*. Springer Verlag, (1)Visual Geometry Group, Department of Engineering Science, University of Oxford, 2015.

[11] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. 2015.

[12] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. 2015.

[13] N. Rossol, I. Cheng, and A. Basu. A multisensor technique for gesture recognition through intelligent skeletal pose analysis. *IEEE Transactions on Human-Machine Systems*, 46(3):350–359, 2016.

[14] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. 2016.

[15] Ning Zhang, Evan Shelhamer, Yang Gao, and Trevor Darrell. Fine-grained pose prediction, normalization, and recognition. 2015.

[16] L. (1) Yin and A. (2) Basu. Generating realistic facial expressions with wrinkles for model-based coding. *Computer Vision and Image Understanding*, 84(2):201–240, 2001.

[17] A. (1) Firouzmanesh, I. (2) Cheng, and A. (2) Basu. Perceptually guided fast compression of 3-d motion capture data. *IEEE Transactions on Multimedia*, 13(4):829–834, 2011.

[18] 3) Basu, A. (1, 2) Mandal, M. (1, and M. (2) Singh. Event dynamics based temporal registration. *IEEE Transactions on Multimedia*, 9(5):1004–1015, 2007.

[19] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. 2017.

[20] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. 2020.

[21] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 2016.

[22] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen. Self adversarial training for human pose estimation. 2017.

[23] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. 2017.

[24] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. 2017.

[25] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. 2018.

[26] A. Elnagar and A. (2) Basu. Motion detection using background constraints. *Pattern Recognition*, 28(10):1537–1554, 1995.

[27] Anup Basu and Segio Licardie. Modeling fish-eye lenses. In *1993 International Conference on Intelligent Robots and Systems*, pages 1822–1828, Univ of Alberta, 1993.

[28] Y. (1) Tian, S.G. (1) Narasimhan, and C.L. (2) Zitnick. *Exploring the spatial hierarchy of mixture models for human pose estimation.*, volume 7576 LNCS of *Lecture Notes in Computer Science*. (1)Carnegie Mellon University, 2012.

[29] B. (1) Rothrock, S. (1) Park, and 2) Zhu, S.-C. (1. Integrating grammar and segmentation for human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number Proceedings - 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pages 3214–3221, (1)Department of Computer Science, University of California, 2013.

[30] S. Park and S.-C. Zhu. Attributed grammars for joint estimation of human attributes, part and pose. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 International Conference on Computer Vision, ICCV 2015, pages 2372–2380, Center for Vision, Cognition, Learning and Autonomy, Department of Computer Science and Statistics, UCLA, 2015.

[31] X. Li and A. Basu. Variable-resolution character thinning. *Pattern Recognition Letters*, 12(4):241–248, 1991.

[32] W. Tang, P. Yu, and Y. Wu. *Deeply Learned Compositional Models for Human Pose Estimation.*, volume 11207 LNCS of *Lecture Notes in Computer Science*. Springer Verlag, Northwestern University, 2018.

[33] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. 2014.

[34] M. Fiala and A. Basu. Hough transform for feature detection in panoramic images. *Pattern Recognition Letters*, 23(14):1863–1874, 2002.

[35] M. (1) Singh, M. (1) Mandai, and A. (2) Basu. Pose recognition using the radon transform. In *Midwest Symposium on Circuits and Systems*, volume 2005, pages 1091–1094, (1)Department of Electrical and Computer Engineering, University of Alberta, 2005.

[36] M. (1) Singh, M. (1) Mandal, and A. (2) Basu. Visual gesture recognition for ground air traffic control using the radon transform. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, number 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pages 2586–2591, (1)Department of Electrical and Computer Engineering, University of Alberta, 2005.

[37] S. Bernogger, A. Basu, L. Yin, and A. Pinz. Eye tracking and animation for mpeg-4 coding. In *Pattern Recognition, International Conference on*, volume 2, page 1281, Los Alamitos, CA, USA, aug 1998. IEEE Computer Society.

[38] L. Yin and A. Basu. Nose shape estimation and tracking for model-based coding. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 3, pages 1477–1480, Department of Computing Science, University of Alberta, 2001.