Action Recognition using Deep Convolutional Neural Networks and Compressed Spatio-Temporal Pose Encodings

William McNally — University of Waterloo, ON, Canada
Alexander Wong — University of Waterloo, ON, Canada
John McPhee — University of Waterloo, ON, Canada

Abstract

Convolutional neural networks have recently shown proficiency at recognizing actions in RGB video. Existing models are generally very deep, requiring large amounts of data to train effectively. Moreover, they rely mainly on global appearance and could potentially underperform in single-environment applications, such as a sports event. To overcome these limitations, we propose to shortcut spatial learning by leveraging the activations within a human pose estimation network. The proposed framework integrates a human pose estimation network with a convolutional classifier via compressed encodings of pose activations. When evaluated on UTD-MHAD, a 27-class multimodal dataset, the pose-based RGB action recognition model achieves a classification accuracy of 98.4% in a subject-specific experiment and outperforms a baseline method that fuses depth and inertial sensor data.

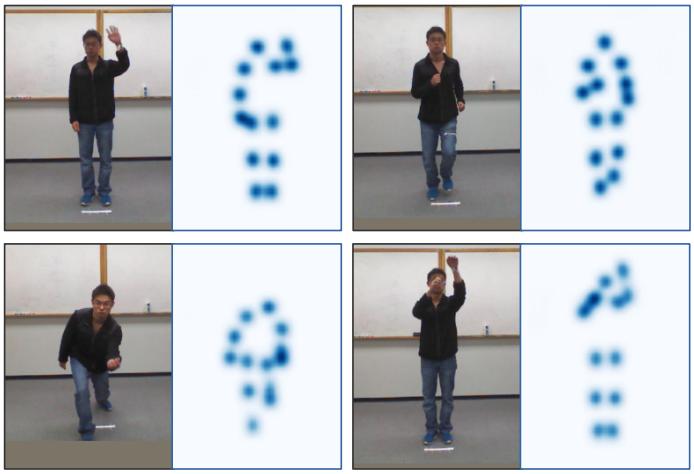

Fig. 1: Spatial slices of compressed spatio-temporal pose encodings. The spatial fusion compresses the pose information, enabling it to be processed by a standard CNN classifier. Sample images taken from the *wave, jog, bowling, and basketball shot* classes of UTD-MHAD [10].

1 Introduction

Human action recognition has been an active area of research in computer vision for several decades due to its wide range of applications in intelligent video surveillance, sports analytics, rehabilitation, and human-computer interaction. With recent advancements in sensor technology, action recognition has benefited from the use of various data modalities, such as 3D skeletal coordinates obtained from depth cameras and wearable inertial sensors [1]. However, depth cameras are severely limited by their working range, often fail in outdoor scenes due to sunlight interference [2], and are not as widely available or economically viable as RGB cameras. Additionally, wearable inertial sensors can be impractical outside of research, in cases like professional sporting events or "in the wild" applications like intelligent surveillance. Consequently, performing action recognition strictly using RGB image data is highly desirable and remains a considerable challenge in the field of computer vision.

Recently, convolutional neural networks (CNNs) have shown a great aptitude for visual recognition tasks [3]. For this reason, they have become the state-of-the-art for RGB-based action recognition. Incorporating temporal information into CNNs has been accomplished by performing 3D convolutions over RGB image sequences directly [4], using recurrent networks [5], or by fusing spatial and temporal features (*i.e.*, RGB and optical flow) using dual-stream networks [6]. Yet, these classifiers rely mainly on global appearance and thus could potentially underperform in situations where multiple unique actions exist within a single environment (*e.g.*, in a sports match).

In a parallel line of computer vision research, CNNs have also been used extensively to infer 2D human pose from RGB images [7, 8, 9]. Although these two streams of research share many similarities (*i.e.*, pose estimation and action recognition), using the *spatial activations* contained within human pose estimation networks to recognize human actions remains relatively unexplored.

In this work, we integrate a state-of-the-art human pose estimation model with a CNN classifier to perform end-to-end human action recognition using RGB image sequences as input. This is accomplished using compressed encodings of spatio-temporal activations produced by the pose estimation network (see Fig. 1). Leveraging the pose information allows us to shortcut spatial learning. As a result, this action recognition classifier can be trained quickly and is tractable enough to achieve good performance on small-scale datasets. When evaluated on a 27-class multimodal action recognition dataset [10], the proposed RGB-based method offers comparable performance to a baseline method fusing two richer, yet impractical data modalities, namely depth maps and inertial sensor data.

2 Method

To summarize, a person detector is used to extract images from an action video sample. The detections are resized, cropped and padded in accordance with the input size of the top-down pose estimation network. For each frame, the pose estimation model generates spatial activations for several *keypoints* (*i.e.*, joints) on the body. The temporal sequence of spatial activations are compressed into spatio-temporal encodings and are used as the input to a relatively shallow CNN classifier.

2.1 Compressed Encodings of Spatial Activations

The 2D human pose estimation model used in this action recognition framework placed first in the 2017 COCO keypoints challenge. The aptly named Cascaded Pyramid Network (CPN) [9] achieved an average precision (AP) of 0.721, a remarkable improvement over the previous year's winning AP of 0.605. We chose the CPN more for its efficient network architecture than its high AP.

Recently, hourglass networks have shown prevalence for the task of human pose estimation [7]. The principle of the hourglass architecture lies in repeated top-down, bottom-up processing to consolidate features across multiple scales and encode the local and global context required for the spatial relationships of the human body. These hourglass modules are then stacked with intermediate supervision to improve performance. However, there are computational inefficiencies associated with hourglass stacking as performance gains drop after two stages, leading to wasteful computations in subsequent stages [9]. The CPN[1] was designed to mitigate these inefficiencies using a feature pyramid network [11] with a ResNet-50 [3] backbone. Consequently, the CPN outperforms an 8-stage hourglass network at less than a third of the computational cost [9].

As with all top-down pose estimation models, a person detector is generally required. In this work, a simple HOG person detector [12] was implemented with non-maxima suppression. Using the bounding box centers returned by the person detector, the images are cropped, resized, and padded in accordance with the 256×192 input size of the pose estimation network.

The action recognition model, illustrated in Fig. 2, takes as input the batch of detections from an action video sample. The pose

[1]The CPN TensorFlow model is available at `https://github.com/chenyilun95/tf-cpn`

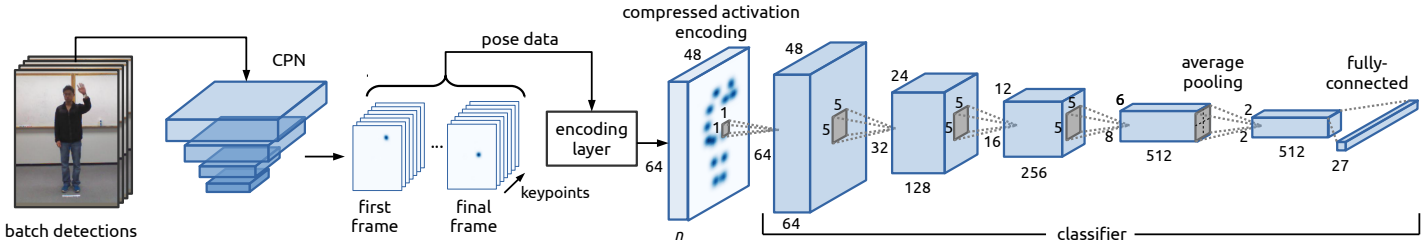

Fig. 2: Proposed action recognition model. The classifier takes as input the compressed spatio-temporal pose encodings and includes four convolutional layers, an average pooling layer, and a final fully-connected layer.

estimation network outputs four-dimensional spatio-temporal pose activations having the shape (frame, height, width, keypoint). As a result of the downsampling present in the pose estimation network, the spatial resolution of the output feature map is one-fourth the size of the input (*i.e.*, 64x48). The pose data is fed into an encoding layer that sums the activations about the keypoint axis (see Fig. 1), concatenates the activations temporally along the channel dimension, and then resizes to a temporal depth of n frames using bilinear interpolation. We considered 13 keypoints in total, including the nose, shoulders, elbows, wrists, hips, knees and ankles. Although the eyes and ears were available, they were not included as their positions are fixed relative to the nose and thus do not provide any additional spatial information. To facilitate batch training the classifier, the encodings were generated prior to training. The effect of the parameter n on performance is discussed in Section 3.2.

2.2 Classifier Architecture

The classifier takes as input refined spatio-temporal information. Thus, its architecture does not need to be as deep as typical action recognition models and does not require any pretraining. The architecture details are included in Fig. 2. The classifier contains just four convolutional layers, an average pooling layer, and a final fully-connected layer upon which softmax regression is applied to predict the class probabilities. The first convolutional layer is a 1D convolution with 64 output channels. The following 3 convolutional layers use a kernel size of 5 and stride of 2, and halve the spatial resolution of the feature-map. Each time the spatial resolution is halved, the number of output channels is doubled. The average pooling layer uses a window size *and* stride of 4×3. In other words, the average pooling layer averages the four quadrants of each 8×6 channel. This "quandrant average pooling" layer was responsible for roughly 5% improvement in performance over the commonly used global average pooling layer. The resulting $2 \times 2 \times 512$ feature-map is then flattened and used as input to a final fully-connected layer.

2.3 Implementation Details

Each convolutional layer is followed by the ReLU activation function. Batch normalization was not used. All network weights were initialized with Xavier initialization [13]. Dropout [14] was applied at a rate of 50% after flattening the $2 \times 2 \times 512$ feature-map and proved to be critical to training (see Fig. 3). Various forms of data augmentation were implemented. Small translations were applied to account for HOG detector noise, small rotations for posture and movement variation, and horizontal and vertical scaling between 0.75 and 1.25 for different body types. All data augmentation was applied randomly with a probability of 50%. To optimize the network, we use the Adam [15] algorithm with a fixed learning rate of 0.001 and batch size of 64. All training was performed on a single Titan Xp GPU.

3 Experimental Results

3.1 Dataset

UTD-MHAD [10] was used for evaluation. UTD-MHAD is a 27-class multimodal dataset containing RGB images, depth images, 3D skeletal coordinates, and inertial sensor data. The dataset was chosen because it comprises high-level body movements with minimal scene interaction, making it suitable for pose-based action recognition. Furthermore, previous works have used this dataset to evaluate action recognition models using other data modalities. Assessing the practicality of these approaches through comparison with an RGB approach is of interest and has not yet been done, likely because the dataset is too small to effectively train any existing CNN action recognition models end-to-end.

The 27 actions are as follows: (1) *swipe left*, (2) *swipe right*, (3) *wave*, (4) *clap*, (5) *throw*, (6) *cross arms*, (7) *basketball shot*, (8) *draw X*, (9) *draw circle clockwise*, (10) *draw circle counter-clockwise*, (11) *draw triangle*, (12) *bowling*, (13) *boxing*, (14) *baseball swing*, (15) *tennis forehand*, (16) *arm curl*, (17) *tennis serve*, (18) *push*, (19) *knock on door*, (20) *catch*, (21) *pick up and throw*, (22) *jog*, (23) *walk*, (24) *sit to stand*, (25) *stand to sit*, (26) *lunge*, (27) *squat*.

The actions were performed by 8 subjects, four male and four female, with each subject performing an action four times. After removing three corrupted samples, the dataset includes a total of 861 samples. Three experimental protocols have been used previously. The first is "cross-subject" (CS), where the data from subjects 1, 3, 5, 7 are used for training and the data from subjects 2, 4, 6, 8 are used for testing. The second is "subject-generic" (SG), where an 8-fold leave-one-out cross-validation is performed, *i.e.*, each subject is used as a test subject while the remaining seven subjects are used for training. The third protocol is "subject-specific" (SS), where the model is evaluated on each subject using 2 trials for training and 2 trials of testing. For the SG and SS experiments, the classification accuracy is averaged over the 8 trials.

3.2 Impact of Temporal Resizing and Dropout

To investigate the impact of the temporal resizing parameter n and dropout, the model was trained using values of n ranging from 2 to 64, with and without dropout. For this hyperparameter search, subject 1 was held out for testing and no data augmentation was used. Due to the stochastic nature of training on a GPU, the model was trained 3 times for each value of n. The mean classification accuracies are shown in Fig. 3. The results show that dropout regularization improves training significantly. Furthermore, classification performance saturated around $n = 32$, which was used to generate the results in Section 3.3.

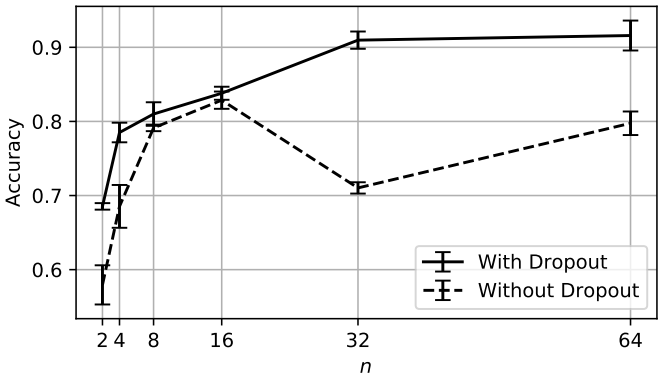

Fig. 3: The effect of temporal resizing parameter n and dropout on classifier performance. Subject 1 was held out for testing, and no data augmentation was used. The error bars represent two standard deviations over the 3 trials.

3.3 Comparison with Other Modalities

To our knowledge, no existing RGB-based action recognition models have been evaluated on UTD-MHAD. This is likely because ex-

Method	CS [10]	SG [1]	SS [1]
Kinect	66.1	74.7	85.1
Inertial	67.2	76.4	88.3
Kinect + Inertial	**79.1**	**91.5**	97.2
Proposed (RGB)	76.1	89.1	**98.4**

Table 1: Classification accuracies on UTD-MHAD for the three experimental protocols. A comparison of the proposed RGB method to baselines that use Kinect and inertial sensor data. To date, no existing RGB methods have been evaluated on UTD-MHAD.

isting RGB CNN models are not tractable enough to perform well on small datasets. One of the key advantages of our system is that it can be trained effectively using limited data, as demonstrated by the results in Table 1.

We chose to evaluate our model against the dataset baselines that use two richer data streams, including depth motion maps generated from the Microsoft Kinect, and inertial sensor data. Specifically, three baselines for each experiment are considered, including those using depth and inertial data individually, and a fused approach where both modalities are combined at the decision level using a collaborative representative classifier [1, 10]. The results in Table 1 indicate that our RGB-based method outperforms the data-fusion baseline in the SS experiment. For the SG and CS experiments, our approach falls within 2.4% and 3.0% of the data-fusion baselines, respectively. As expected, the best performance is seen in the SS experiment due to the minimal amount of variance within the intra-subject video samples.

Fig. 4 depicts the confusion matrix for the evaluation of the proposed model on the test dataset in the CS experiment. Frequent misclassifications were observed between the *bowling* (12) and *lunge* (26) classes, and the *throw* (5) and *knock on door* (19) classes. These classes consist of very similar body movements.

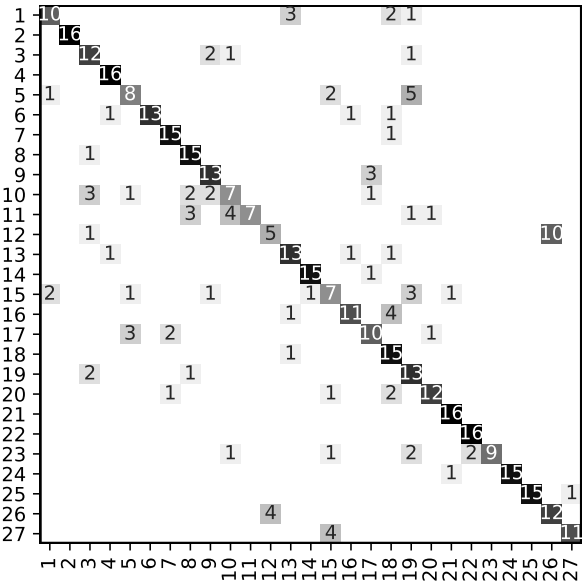

Fig. 4: Confusion matrix for the CS experiment. *Bowling* (12) was frequently misclassified as a *lunge* (26).

4 Discussion

The proposed RGB action recognition model was demonstrated to perform comparably to a method that fused two richer, yet impractical data streams from depth cameras and inertial sensors. These results put into question the practicality of action recognition models utilizing these modalities, given their limitations in comparison to RGB images. Furthermore, it was demonstrated that tractability can be maintained using RGB by shortcutting spatial learning with the integration of a human pose estimation network. Currently, human pose estimation is a highly active area of research. Future pose-based RGB action recognition models may reap the benefits of pose estimation advancements such as 3D pose estimation, and body segment identification.

Acknowledgments

We acknowledge financial support from the Canada Research Chairs program and the Natural Sciences and Engineering Research Council of Canada, as well as a hardware grant from NVIDIA.

References

[1] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[2] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 44, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.

[7] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, pp. 483–499, 2016.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[9] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, 2018.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proceedings of the IEEE Conference on Image Processing*, pp. 168–172, 2015.

[11] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients (hog) for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, IEEE, 2005.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.