Investigating the Impact of Estimated Modalities in Multi-Modal Activity Recognition

Rahul Rajan
rrajan39@gatech.edu

ABSTRACT

RGB-D data obtained from affordable depth-sensors, like the XBox Kinect has allowed for remarkable progress in the field of human activity recognition (HAR). Depth information has been found to significantly increase performance in HAR tasks, especially when it's fused with other modalities like RGB and Optical flow. Unfortunately, the use of depth sensors limits where these models can be used since these sensors are often difficult to use in outdoor settings. Additionally, most videos available today are shot on traditional video cameras, which don't provide depth information needed to run RGB-D based HAR models. Fortunately, deep learning has allowed us to estimate this depth data with high accuracy from just RGB video. This paper investigates the viability of directly using this estimated depth information in RGB-D models for HAR-related tasks.

KEYWORDS

activity recognition, video classification, RGB-D, monocular depth estimation, neural networks

1 INTRODUCTION

The goal of video activity recognition is to identify specific actions happening in a video or a part of a video. These actions can be high-level, such as "playing soccer" or "cooking", to more low-level actions like "kicking" or "standing up". Developing good models for this task has a wide range of applications including automated surveillance in videos, behavior/activity tracking, and improved video summarization. This is not an easy problem to solve since people can perform activities in slightly different ways, and more importantly, videos can have hidden body parts making it difficult for a model to accurately identify an action. The introduction of affordable sensors, like the Xbox Kinect, has sparked further research into activity recognition [17]. These sensors can collect data in more modalities like depth, skeletal data, infrared, in addition to traditional RGB-based video. Activity recognition models can take advantage of this additional context to make more accurate predictions. In this work, we focus on Red Green Blue Depth (RGB-D) video based activity recognition techniques. RGB-D data contains both traditional video and depth information. Depth data provides pixel-level information representing the distance from the camera to objects in the scene. RGB-D video provides models with more dense data that can help them perform well even when parts of the scene are occluded. As a result, depth information can also help in situations where a subject is in a position that is hard to resolve from just RGB information. However, using a depth camera heavily restricts where these models can be used since most videos are recorded on traditional cameras [24] . Requiring depth information for activity recognition places a huge burden on researchers since they have to set up specialized lab experiments with depth cameras. Additionally, these researchers can't take advantage of many large benchmark activity recognition datasets like Kinetics, ActivityNet, and THUMOS-14 [[11], [6], [8], which don't provide depth data. Fortunately, in recent years, the field of monocular depth estimation has made incredible progress. Monocular depth estimation is a method for estimating depth information from a single RGB camera. Several methods have emerged that can recover dense depth maps from just RGB video frames. Many of these models have found success in-depth estimation using end-to-end deep learning techniques with models like convolutional neural networks (CNNs) [5], Generative Adversarial Networks (GANs) [12], and Transformers [4].

Although depth estimation has found a lot of success, there's currently very little research into whether using estimated depth can potentially replace real depth information in RGB-D based activity recognition models. This work serves as an initial look into whether estimated depth information can serve as a viable replacement for depth-sensors. To conduct this study, we use a current state-of-the-art depth estimation model and train a baseline activity recognition model with estimated depth frames, and evaluate on a large benchmark RGB-D activity recognition dataset.

2 LITERATURE REVIEW

2.1 Convolutional Neural Networks

Using convolutional neural networks (CNNs) for vision related tafsks is not a new concept. 2D CNN architectures, such as ImageNet, have been used very successfully in image classification tasks for several years, and more recent research has found ways to modify CNNs to support the temporal dimension provided in video data. The most important advance was the introduction of 3D CNNs in videos by Ji [10]. 3D CNNs are different from traditional CNNs in that their kernels have both spatial and temporal dimensions. Early 3D CNN research found success in activity recognition, but they were not able to obtain state-of-the-art results on benchmark datasets.

Another key development in using CNNs for video tasks was Simonyan and Zisserman's work in two-stream 3D CNN networks [19]. Their model consists of 2 3D CNNs- the first 3D CNN is trained on normal RGB frames. The second model is trained on optical flow frames extracted from the video to help the model better understand motion information across frames. The group found that combining the RGB and Optical flow model's predictions resulted in high performance on video activity recognition benchmarks. The I3D network, proposed in [1], builds on previous work on two-stream CNN's and 3D CNNs [21] [19] [3]. The I3D architecture works by adding a temporal dimension to the kernels of the highly successful image recognition model architecture, ImageNet [2]. The work also proposes employing a two-stream model by having I3D networks for both RGB and Flow frames to help the model extract motion and

spatial information from the video frames. Today, I3D still serves as a key benchmark model for video activity recognition tasks, making them a good candidate to use for our research.

2.2 Depth Estimation

The goal of depth estimation is to calculate a depth map, the distance of every point or pixel in a video from the camera. Obtaining accurate depth maps can help models understand the geometry of the scene in the video. This is very important in activity recognition tasks because knowing the location and orientation of a subject relative to other objects in the scene can provide helpful context when differentiating between similar action categories. Additionally, depth maps can help models process partially occluded features, like objects and body parts. Although depth cameras can provide precise depth maps because they use depth sensors, they are not widespread and more expensive than traditional video cameras. As a result, research has shifted to predicting depth maps directly from a normal video. Traditional video only contains RGB information, making it difficult to recover depth data [15]. Fortunately, deep learning has emerged as a solution and has obtained state-of-the-art results on several depth estimation benchmarks. Two of the primary datasets that were studied for this research were KITTI [14] and NYU-Depth [18].

The KITTI dataset consists of over 93,000 images of outdoor scenes with their corresponding depth maps. The dataset is captured through a car equipped with two color cameras, 2 gray-scale cameras, a laser scanner, and a global positioning system. For monocular estimation, only the data from the cameras and the laser scanner are used. Although performance on this dataset is important in evaluating depth estimation models, it is not the ideal dataset for our use case, since we primarily focus on indoor activities.

The NYU-Depth-v2 dataset contains over 400,000 RGB-Depth Map image pairs for depth estimation tasks. Scenes are all captured indoors with a video camera and a Microsoft Kinect simultaneously collecting data. Each image captured has a size of 640x480 and depth ranges from 0.5 meters to 10 meters. Furthermore, the dataset includes 464 different scene setups, providing plenty of variation for evaluating depth estimation models.

Modern depth estimation models tend to rely on an encoder-decoder convolutional neural network that takes in an input RGB frame and output a new depth image [15]. The encoder traditionally consists of several convolutional and pooling layers that are designed to extract depth-associated features. The decoder, on the other hand, consists of deconvolutional and pooling layers that take in the encoder embedding and generate the depth map. Using a deep encoder-decoder architecture allows the model to still recover depth information even if the input image has small occlusions. One of the current state of the art models, Adabins [4] uses a baseline encoder-decoder architecture, EfficientNet B5 [20], and attempts to process global information using a vision transformer. The transformer is used to predict depth bins, which discretizes the overall depth range of the scene. The centers of these bins are then used as the final depth values for each pixel in the image.

2.3 RGB-D Activity Recognition

There are several methods for performing action recognition from RGB-D sensors. Early work in depth-based methods relied on constructing hand-crafted features that are then processed to identify the activity class. Li et al. extracted bags of 3D points sampled from the silhouette of depth maps and performed clustering on these features for classification [13]. Yang et al. calculated differences between adjacent depth maps and then extracted features using Histogram of Gradients (HOG) [23] . However, these methods rely on constructing specialized features, which can be computationally expensive and not generalizable to multiple datasets/domains. Research has shown that these depth-based methods can perform better than those that only use RGB-based video. As a result, recent work has focused on using deep learning to fuse multiple modalities (i.e. RGB and Depth) to get more accurate predictions in activity recognition tasks. Xu et al. preprocesses RGB-D frames by extracting frames that contain obvious motion and passes these frames into a two-stream I3D network that takes in RGB and Depth frames [22]. Hu et al. trained a CNN with bilinear pooling to fuse temporal information from RGB and Depth data [7]. Imran trained CNNs on motion history information (MHI) from RGB and rotated depth maps (front, side, and top view), and averaged their predictions for activity classification [9].

There are several benchmark datasets used to evaluate RGB-D based HAR models. The most popular is NTU RGB-D. The dataset is a large-scale indoor activity recognition dataset that contains 56,660 action samples [16]. Each sample consists of an action type/class, an RGB video, depth maps for each video frame, and skeletal pose data. The depth and pose data are recorded using Microsoft Kinect cameras. The dataset consists of 40 subjects performing 60 different actions (i.e. "drink water", "put on jacket", etc.). Each video is also recorded at different angles using 3 Kinect v2 cameras placed at different angles in the scene. As a result, the dataset provides 2 splits that are used for training evaluation. The first split is the cross-subject split where videos of 20 subjects are used for training and the remaining subjects are used for evaluation. The second split is a cross-view split where 2 camera views are used for training and the remaining 1 camera view is used for testing.

3 METHODOLOGY

The goal of our experiments was to first understand the performance of estimated modality data and also the effect of combining multiple modalities has on classification performance. Our work focuses on 2 modalities: RGB and Depth. Since depth can't be extracted directly from raw RGB frames, we rely on off-the-shelf estimation models. We then use the NTU RGB+D dataset that contains ground-truth data for each modality and obtain estimated data from the RGB frames. Next, we pass both ground-truth and estimated modality data into a classification model and compare performance. After conducting per-modality analysis, we conducted experiments on combining modality data to understand which modality combinations are the best for activity recognition tasks.

3.1 Modality Estimation

Since RGB frames can be extracted without an estimation model, we focus on obtaining estimated depth data. For obtaining estimated depth sequences, we use the current state-of-the-art depth estimation model: Adabins [4]. Adabins outputs high-density depth maps from single RGB frames. To obtain depth sequences, we run the model on every frame in a given video. For our experiments, we used an Adabins model pre-trained on the NYU-Depth-v2 dataset. NYU-Depth contains over 1400 depth annotated images focusing on indoor scenarios. We believe using this dataset is sufficient since we run our experiments on the action recognition dataset, NTU RGB-D, which also contains similar indoor scenes. Figure 1 shows a comparison between the real and estimated depth information from NTU RGB-D. It is important to note that even though the edges in the estimated depth frame are not as defined, Adabins was still able to recover similar depth features compared to the real depth frame.

Figure 1: Comparison of Real Depth Image (left) vs. the Predicted Depth Image (right)

3.2 Activity Classification

To evaluate modality estimation data, we need ground-truth RGB and Depth Map data. To obtain this, we use the NTU RGB+D dataset [16], which provides activity labels for videos with per-frame depth annotations. The dataset contains over 56,000 annotated videos across 60 activity classes. To test each modality, we use a baseline I3D architecture for RGB and Depth since these modalities are represented as images. We train our model on the subject split of the NTU RGB+D dataset for each modality. We then evaluate the model on both the ground truth and our estimated modality data for accuracy and F1-score.

3.3 Implementation Details

Individual RGB and depth frames from the NTU Dataset are resized to 300x300 and normalized between -1 and 1 before being passed into the I3D network. Estimated depth is obtained at a frame level from the pre-trained Adabins model and preprocessed in a similar fashion. We then pass in the first 32 frames from each video and apply random horizontal flipping during the training process. Each I3D model is trained for 64,000 iterations with a batch size of 24 across 2 NVIDIA Titan Xp's.

	RGB	Pred. Depth	RGB+Pred.Depth
Accuracy	0.53	0.41	0.54
F-1 Score	0.55	0.43	0.56

Table 1: Final Accuracy and Average Weighted F-1 Scores

3.4 Multimodal Learning

After obtaining trained models on each individual modality, we perform late fusion by averaging the final logits from the I3D network for different combinations of modalities. In our experiments, we tested fusing RGB and Depth.

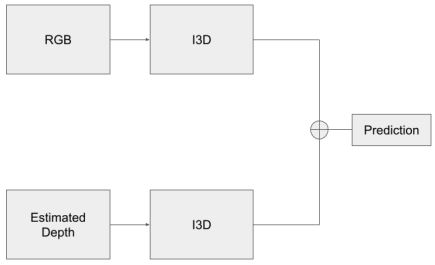

Figure 2: Fusion Architecture for Multimodal Learning

4 RESULTS

Figure 3: RGB vs Depth data for Clapping Activity

Due to technical issues, we were unable to obtain results using ground truth depth data in time. However, we present results from the RGB and Predicted depth models.

Table 1 shows the performance results after 64,000 training iterations. Although we were unable to get results with the real depth, the metrics for predicted depth shows that the model is able to learn activity classes from just estimated depth frames.

4.1 Classification Performance

Although the predicted depth model performed worse than the RGB model, it performed better for certain classes where body

parts or objects are occluded ("Clapping", "play with phone/tablet", and "Wipe face"). The most notable example was in the "Clapping" activity class. The predicted depth model obtained an F-1 score of 0.26, while the RGB model obtained an F-1 score of 0.21. Figure 7 shows a comparison between RGB and depth frames from a sample clapping video. When the subject's hands are together during a clap, one arm is mostly occluded. However, we found that in the depth frames, the subject's arms have different depth values. We also found a similar pattern with the "Wipe Face" class where some of the predicted depth frames showed clear depth differences between the subjects' arm/hand and their face. This indicates that predicted depth information provides features that can improve classification performance for certain activities.

4.2 Fusion Performance

In addition to single modality results, we also reported performance from late fusion with RGB and predicted depth. We found a slightly improved F-1 score and accuracy when averaging the RGB and depth model predictions, with an increase of 0.1 for both. Although this is not a huge improvement, it provides further indication that estimated depth can improve classification.

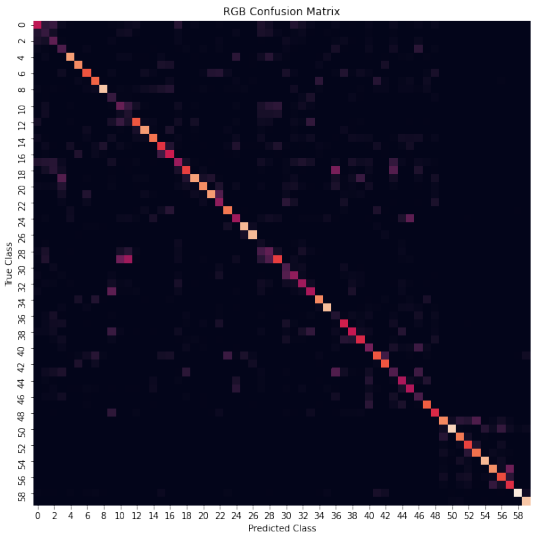

Figure 4: Confusion Matrix for RGB

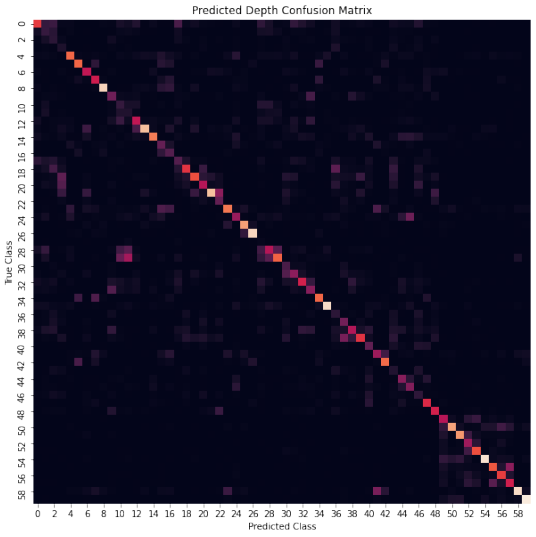

Figure 5: Confusion Matrix for Predicted Depth

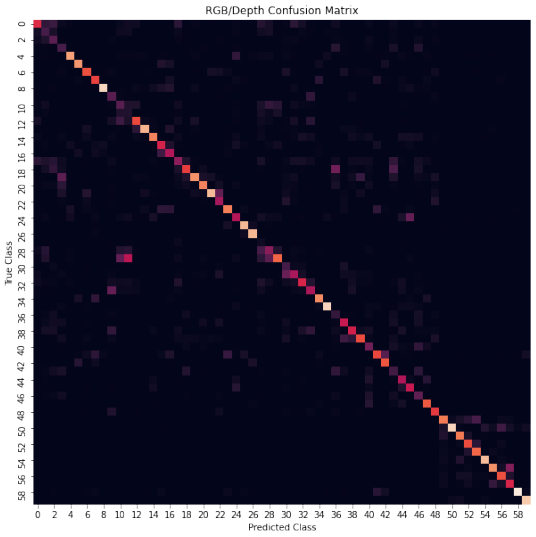

Figure 6: Confusion Matrix for RGB+Predicted Depth

4.3 Misclassifications

Figures 4 and 5 plot the confusion matrices for RGB and predicted depth, respectively. Lighter colors signify a higher frequency, while the darker colors signify a lower frequency of samples. When comparing our baseline RGB model's confusion matrix with the predicted depth model's, there are a couple of differences. The most significant difference is the brighter point between classes 58 (two people walking towards each other) and 41 (staggered walking). The predicted depth model had a higher rate of classifying staggered walking when the actual video is two people walking towards each other. This is likely due to depth data lacking information required to distinguish two subjects. Since walking and staggered walking are similar activities, it's possible the predicted depth model was

unable to classify them accurately. That being said, this is not a problem with using estimated depth data since real depth information would lack the same features. Additionally, the confusion matrix for the fused RGB+Predicted Depth results (Figure 7) shows that combining the modalities resulted in the misclassification going away.

Similarly, there are also cases where the estimated depth model had a lower misclassification rate between certain classes compared to the RGB model. One example is between the true class 36 (wiping face) and predicted class 43 (having a headache) in Figure 4. Since we've found that estimated depth has performed better when there are occluded objects/limbs, it's possible that the depth model was able to decipher between the two actions more accurately. Additionally, Figure 6 shows that fusing the two modalities lowered the number of misclassifications for this pair of activities.

5 DISCUSSION/FUTURE WORK

Our results serve as an initial look into the feasibility of using estimated depth data in activity recognition models. We found that our estimated depth model was to take achieve adequate performance on the NTU dataset. We believe we achieved good benchmark results since we used a baseline activity recognition model without any special preprocessing of depth data or video frames. Furthermore, there is currently very limited results on using a vanilla I3D model with depth data, making it difficult to compare results.That being said, we found evidence indicating that the estimated depth model was using depth related information to achieve better classification performance in key activity categories. Our fusion results also show that the RGB and estimated depth modality complement each other well by reducing misclassificions in certain activity classes. Thus, we see promising results indicating that activity classification models can enjoy benefits of 3D information from monocular depth estimation.

Due to the time and resources needed to train these models, we're unable to report results on using real depth information. However, future research should continue this work by comparing classification performance with real depth data and identifying activity classes that are not easily learned when using estimated information. Additionally, further study is required to identify the best way to fuse RGB and estimated depth information together.

Research should also look into other estimated modalities, such as skeletal data, that can be extracted from just RGB video and whether they can replace ground truth information in activity recognition tasks. Additionally, modality estimation models should be fine-tuned on datasets with people to improve segmentation between body parts.

Overall, our work shows that it is feasible to learn depth features relevant for activity classification from just a traditional RGB video. We believe that this serves as a first step towards eliminating the need for depth cameras for HAR tasks and improving the versatility of models that depend on depth data.

REFERENCES

[1] Carreira, J., and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[2] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.

[3] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 4 (2017), 677–691.

[4] Farooq Bhat, S., Alhashim, I., and Wonka, P. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 4008–4017.

[5] Garg, R., B.G., V. K., Carneiro, G., and Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision – ECCV 2016* (Cham, 2016), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, pp. 740–756.

[6] Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 961–970.

[7] Hu, J.-F., Zheng, W.-S., Pan, J., Lai, J., and Zhang, J. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018).

[8] Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., and Shah, M. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding 155* (Feb 2017), 1–23.

[9] Imran, J., and Kumar, P. Human action recognition using rgb-d sensor and deep convolutional neural networks. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (2016), pp. 144–148.

[10] Ji, S., Xu, W., Yang, M., and Yu, K. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 1 (2013), 221–231.

[11] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. The kinetics human action video dataset, 2017.

[12] Kumar, A. C., Bhandarkar, S. M., and Prasad, M. Monocular depth prediction using generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018), pp. 413–4138.

[13] Li, W., Zhang, Z., and Liu, Z. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (2010), pp. 9–14.

[14] Liao, Y., Xie, J., and Geiger, A. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv.org 2109.13410* (2021).

[15] Ming, Y., Meng, X., Fan, C., and Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing 438* (2021), 14–33.

[16] Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. Ntu rgb+d: A large scale dataset for 3d human activity analysis, 2016.

[17] Shaikh, M. B., and Chai, D. Rgb-d data-based action recognition: A review. *Sensors 21*, 12 (2021).

[18] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *Computer Vision – ECCV 2012* (Berlin, Heidelberg, 2012), A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., Springer Berlin Heidelberg, pp. 746–760.

[19] Simonyan, K., and Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems* (2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27, Curran Associates, Inc.

[20] Tan, M., and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[21] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks, 2015.

[22] Xu, Z., Vilaplana, V., and Morros, J. R. Action tube extraction based 3d-cnn for rgb-d action recognition. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* (2018), pp. 1–6.

[23] Yang, X., Zhang, C., and Tian, Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. pp. 1057–1060.

[24] Zhao, C., Sun, Q., Zhang, C., Tang, Y., and Qian, F. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences 63*, 9 (Jun 2020), 1612–1627.