A Trimodal Dataset: RGB, Thermal, and Depth for Human Segmentation and Temporal Action Detection

Christian Stippel[1 [0000−0003−0482−902X]], Thomas Heitzinger[1 [0000−0002−3129−5054]], and Martin Kampel[1 [0000−0002−5217−2854]]

Computer Vision Lab, TU Wien, Vienna, Austria
{christian.stippel, thomas.heitzinger,martin.kampel}@tuwien.ac.at

Abstract. Computer vision research and popular datasets are predominantly based on the RGB modality. However, traditional RGB datasets have limitations in lighting conditions and raise privacy concerns. Integrating or substituting with thermal and depth data offers a more robust and privacy-preserving alternative. We present TRISTAR[1], a public TRImodal Segmentation and acTion ARchive comprising registered sequences of RGB, depth, and thermal data. The dataset encompasses 10 unique environments, 18 camera angles, 101 shots, and 15,618 frames which include human masks for semantic segmentation and dense labels for temporal action detection and scene understanding. We discuss the system setup, including sensor configuration and calibration, as well as the process of generating ground truth annotations. On top, we conduct a quality analysis of our proposed dataset and provide benchmark models as reference points for human segmentation and action detection. By employing only modalities of thermal and depth, these models yield improvements in both human segmentation and action detection.

Keywords: segmentation · temporal action segmentation/detection · scene understanding · video understanding

1 Introduction

RGB data is one of the most commonly used modalities for computer vision datasets [25, 18]. However, this modality has a number of notable shortcomings. Firstly, RGB sensors are sensitive to lighting conditions, and their image quality can be compromised under non-optimal conditions. Secondly, the RGB modality can lead to the identification of individuals, posing potential privacy concerns in sensitive applications. Lastly, segmentation accuracy may suffer due to inadequate camera quality and intensity similarities between the foreground and background.

Integrating various modalities offers a more comprehensive and detailed scene representation: color modalities provide contour and texture details, depth data

[1] https://zenodo.org/record/7996570, https://github.com/Stippler/tristar

shows the scene geometry, and thermal imaging contributes temperature information.

This paper presents a unique trimodal dataset designed to address the limitations of existing single-modality datasets in semantic segmentation and action recognition. Our dataset comprises 101 registered sequences of RGB, thermal, and depth shots, captured in diverse office scenarios. The key dataset characteristics are listed in Table 1.

Table 1: Details of the Trimodal Dataset.

Content	Indoor Human Behavior
Modalities	Registered RGB, Depth, Thermal
Type of Data	Sequences
Resolution	640x480
Frame Rate	8.7 fps
#Offices	10
#Camera Angles	18
#Shots	101
#Frames	15,618
#Individuals	8
#Actions	14

Figure 1, shows samples of our trimodal dataset where we employ the registration methodology outlined by Stromayer et al. [31] to align the different modalities.

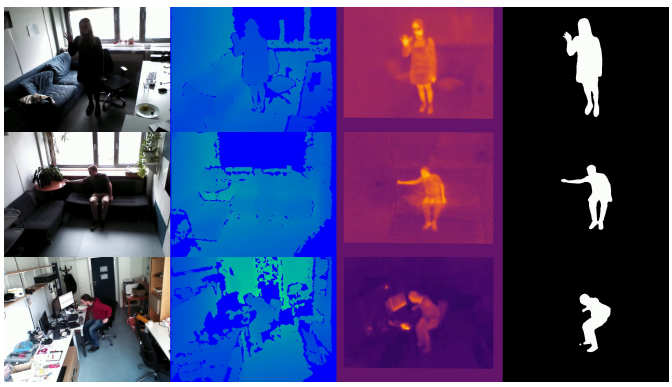


Fig. 1: Examples from our trimodal dataset, encompassing RGB, depth, thermal imaging, and human segmentation mask.

The second row emphasizes a potential limitation of relying solely on the RGB modality: the similar RGB intensities in the foreground and background make it challenging to precisely distinguish the person from the couch.

Our key contributions include a novel trimodal dataset consisting of varying office settings, that provides a resource for researchers focusing on multi-modal data fusion and related tasks, combined with human segmentation maps to aid with the task of, semantic segmentation, action labels that describe a wide range of activities, enabling the development and evaluation of temporal action detection models.

In addition, we train benchmark models on the task of human segmentation and action recognition and provide them as baselines for further research. These baselines demonstrate the effectiveness of complementing or replacing RGB with depth and thermal modalities.

The selection of segmentation and temporal action segmentation as our primary downstream tasks is based on several considerations. First, these tasks represent diverse levels of complexity, allowing us to demonstrate the utility of our trimodal dataset across a wide array of challenges. Second, both tasks have substantial real-world applicability, encompassing use-cases from autonomous driving to assistive technologies, underscoring the practicality of our research. Third, they particularly benefit from multimodal data, with depth and thermal information enhancing performance by offering structural context and distinguishing capabilities. Finally, as these tasks are commonly used for benchmarking in computer vision, they enable a direct and meaningful comparison of our work against existing methodologies and datasets.

The remainder of this paper is structured as follows: Section 2 presents a review of existing datasets and their limitations; Section 3 provides an overview of the system setup, including sensor configuration and calibration; Section 4 details the construction, analysis, and evaluation of our trimodal dataset; Section 5 outlines potential tasks and applications that benefit from the dataset; and Section 6 concludes the paper and discusses possible future work.

2 Related Work

Our efforts to build a trimodal dataset are based on an understanding of the existing literature on datasets, their limitations, and current methodologies employed for multimodal datasets. To provide a broader context for our work, we now delve into a review of existing datasets and their accompanying methodologies.

2.1 Datasets

Some notable examples of datasets for semantic segmentation and action recognition include PASCAL VOC [8], COCO [18], ADE20K [36], and the Charades dataset [28]. While these datasets played a vital role in advancing semantic segmentation research and action classification, they primarily focus on RGB data

and lack the inclusion of additional modalities such as thermal and depth information.

Depth datasets provide information about scene geometry and can be used in a various domains, such as 3D reconstruction and scene understanding. Examples of outdoor depth datasets include KITTI [10] and Cityscapes [3], while the NYU Depth [30] dataset is a prominent example of an indoor depth dataset. Furthermore, IPT [13] is a depth dataset for tracking tasks in enclosed environments. Although these datasets provide depth information, they do not include thermal data, which can be crucial for addressing challenges posed by varying lighting conditions and ensuring privacy preservation. The follow-up MIPT dataset also includes a small number of depth and thermal sequences, but lacks RGB data [12].

Thermal imaging is acknowledged for its potential in a variety of computer vision tasks [12, 13, 17]. Unlike traditional RGB imaging, thermal imaging is less susceptible to illumination changes and provides additional information about the subject. Thermal data provides information of the temperature distribution, which proves particularly useful in scenarios with bad lightning and semantic segmentation of living objects. Several thermal image datasets have been introduced to promote research in this domain. One noteworthy example is the OSU Thermal Pedestrian Dataset [5], which includes a substantial number of pedestrian thermal images collected under different environmental conditions. However, this dataset is primarily used for pedestrian detection tasks rather than segmentation or action recognition. Another dataset, the Terravic Facial Infrared Database [21], contains both visible and thermal facial images. Kniaz et al. propose Thermagan for person re-identification and publish their dataset ThermalWorld alongside it [17]. Heitzinger et al. introduce an identity-preserving 3D human behavior analysis system that addresses privacy concerns in continuous video monitoring. They also release a public multimodal dataset composed of depth and thermal sequences, intended to support a variety of privacy-sensitive applications in ambient-assisted living and human security monitoring [12]. Brenner et al.'s survey [1] provides a systematic literature review of the fusion of RGB-D and thermal sensor data, highlighting the progress made in this area over the past decade. The PST900 dataset [27] is one resource that proposes long wave infrared (LWIR) imagery as a supporting modality for semantic segmentation using learning-based techniques. This dataset provides 894 synchronized and calibrated RGB and thermal image pairs with per-pixel human annotations across four distinct classes. In addition to presenting a unique dataset, the authors introduce a novel passive calibration target. Another notable resource is the InfAR action dataset [9], which focuses on action recognition using infrared data.

To the best of our knowledge, only a single dataset exists that combines RGB, thermal, and depth data [22] for human segmentation. This dataset consists of 5,274 frames recorded in three shots in three distinct office scenes.

Given the scarcity and potential benefits of trimodal data, we motivate the creation of our dataset; we provide 101 different shots recorded in 10 offices

from 18 unique camera angles to provied a resource for researchers working on multi-modal data fusion and related tasks.

2.2 Methods

One of the earliest methods for human segmentation is the Histogram of Oriented Gradients (HOG) descriptor combined with a Support Vector Machine (SVM) for human detection, introduced by Dalal and Triggs [4]. However, this approach struggles with occlusions and variations in human appearance. To overcome these challenges, more recent works have leveraged the power of deep learning. Mask R-CNN, proposed by He et al. [11], extends Faster R-CNN [23] by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. DeepLabv3+ [2] is another method that employs an encoder-decoder structure with dilated convolutions and spatial pyramid pooling for semantic segmentation. Recent methods for segmentation have moved towards leveraging attention mechanisms and transformers, leading to the development of architectures such as Self-Attention Generative Adversarial Networks (SAGANs) [35], and Vision Transformers (ViTs) [7]. One particularly notable method is the Swin Transformer [19], which introduces a hierarchical structure with shifted windows to enable efficient self-attention over images. Finally, Segment Anything (SAM) paves the way for generalizable zero-shot segmentation that can be applied to RGB, thermal, or depth modalities [16]. These methods have primarily been developed and evaluated using RGB data. Their effectiveness with depth or thermal data is less explored, likely due to the scarcity of multimodal datasets that inclued these modalities.

In the domain of action recognition and classification long-term Recurrent Convolutional Networks (LRCNs) [6], 3D Convolutional Neural Networks (3D-CNNs) [14] and transformer-based approaches such as the Video Swin Transformer [20] are among the notable methods. These models primarily rely on RGB data, however, the exploration of action recognition in multimodal datasets combining RGB, depth, and thermal data is relatively limited, primarily due to the scarcity of such datasets. Although there are a few existing works like NTU RGB+D [26] that incorporate depth information, their ability to handle thermal data is limited.

In summary, while existing datasets and methods have significantly advanced semantic segmentation and action recognition, they predominantly focus on RGB data. The limited availability of multimodal datasets, particularly those combining RGB, thermal, and depth data, limits exploration into the potential benefits of these modalities.

3 System Overview

In light of the scarcity of RGB, depth and thermal datasets identified in the prior literature, we designed and implemented a system to capture and annotate data across these three modalities. This system includes a Compact Tri-Modal

Camera Unit (CTCAT) for data acquisition and implements a streamlined annotation process for data labeling [31]. Our approach is inherently scalable and capable of distributed operation, leveraging the novel zero-shot model Segment Anything [16] for effective object recognition and employing a distributed labeling system that allows for efficient, large-scale annotation tasks [32]. Annotators perform labeling on the RGB modality. However, to improve the accuracy of the labeling process, they are provided access to the corresponding thermal and depth data. These additional data are mapped to RGB using a color scale, as illustrated in Figure 1.

3.1 Sensor Setup

We utilize the Compact Tri-Modal Camera Unit (CTCAT) described by Strohmayer et al. [31]. The CTCAT combines three types of cameras: RGB, a structured light depth camera with an operational range of 0.6-8m, and a 160x120 uncooled radiometric thermal camera which allows us to capture across all three modalities at a rate of up to 8.7 fps. Although each camera has its unique resolution, we rescale all images to a standardized resolution of 640x480 pixels for consistency within the dataset. To align the RGB, thermal, and depth cameras, a custom-made, heated checkerboard calibration pattern is used.

Figure 2 illustrates our camera setup, a close-up view of the CTCAT unit and a sample trimodal shot, capturing the scene's diverse modalities. Mounted on a tripod, the camera and its accompanying portable monitor are powered by an affixed battery pack. Typically positioned on tables or countertops at a height between two and three meters, the setup allows for comprehensive capture of office scenes. Before recording, we optimize the camera perspective using the narrow field of view of the thermal sensor. Individuals are then filmed performing various actions, guided by instructions provided by a designated person. The action list includes tasks like picking up a glass, drinking, and typing.

Fig. 2: The camera setup with the CTCAT unit and the captured scene.

3.2 Ground Truth Generation

We employ pretrained YoloV7 and YoloV8 models [34], [15] to detect bounding boxes of humans based on the RGB modality of our trimodal dataset. These bounding boxes serve as input for the Segment Anything tool [16], allowing us to obtain preliminary human masks. The initial results obtained from Segment Anything form the foundation for the manual labeling process. A team of twelve annotators undertakes the task of labeling a total of 15,618 frames on the RGB modality. The generated masks are also used for the corresponding depth and thermal frames. While the labeling takes place directly on the RGB modality, the annotators are also provided access to the corresponding thermal and depth modalities, registered and color mapped, to use as reference in cases where the person is not clearly distinguishable from the background. To facilitate this large-scale task, we utilize a self-hosted Label Studio instance [32] which allows multiple annotators to work simultaneously. Figure 3 illustrates the human segmentation annotation process using Label Studio.

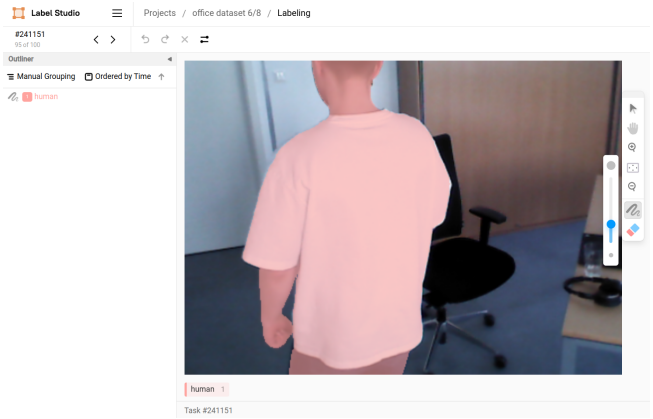


Fig. 3: Illustration of the manual human segmentation annotation process using Label Studio.

Action labeling is conducted with dense per-frame labeling of 14 classes, categorized into actions, states, transitions between states, and location of the person on the RGB modality. The labeling process is performed using a spreadsheet with the columns: original file name, person, shot, frame, actions, transitions, state, and location. If multiple labels were applicable within a single column, they are delimited with a space. The objective for this task is temporal action segmentation or action detection, which entails identifying the specific actions occurring within a given frame.

4 Trimodal Dataset

Building upon the groundwork laid out in the sensor setup, we now present our own multimodal dataset. Our dataset encompasses an array of office scenes recorded using a trimodal sensor arrangement integrating RGB, thermal, and depth data.

4.1 Dataset Design

In order to construct our multimodal dataset, we drew inspiration from notable contributions in the field. The Charades Dataset [29], shows humans in indoor environments and is densely labeled with activities. Meanwhile, the work of Palmero et al. [22] highlights the potential of multimodal datasets, albeit with a more limited scope and volume of data.

For our dataset we select an office environment, as it provides a variety of scenarios, activities, and lighting conditions. The selection of action labels is based on their occurrences in a real office setting. Table 2 presents a comprehensive list of the various actions, states, transitions, and locations that are represented in our dataset.

Table 2: List of Actions, States, Transitions, and Locations used for Labeling.

Label	Items
Action Classification	`put_down`, `pick_up`, `drink`, `type`, `wave`
State	`sit`, `walk`, `stand`, `lie`
Transitions	`get_down`, `get_up`
Location	`out_of_view`, `out_of_room`, `in_room`

Furthermore, our dataset includes different types of office spaces, namely open office environments, meeting rooms, and individual offices, each offering a different layout and set of interactions with surrounding objects. Figure 4 illustrates the variety of office locations and lighting conditions covered in our dataset.

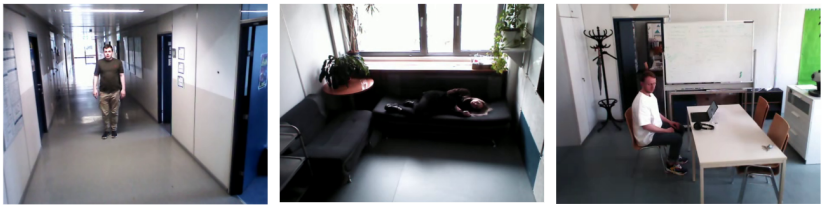


Fig. 4: Variety of office locations and lighting conditions in the dataset.

Lighting conditions, a crucial factor in visual data, are also varied in our dataset ranging from bright daylight, and artificial lighting, to low-light conditions. Finally, our dataset encompasses object classes commonly found in office environments ranging from office furniture such as desks, chairs, and cabinets to electronic devices like computers, and phones, as well as various personal items.

4.2 Dataset Analysis

Our dataset comprises 15,618 annotated frames, recorded from 18 unique camera angles. It documents the actions of eight individuals within various office settings, as highlighted by the distribution in Figure 5.

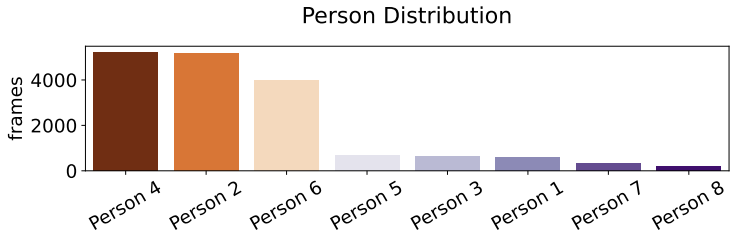
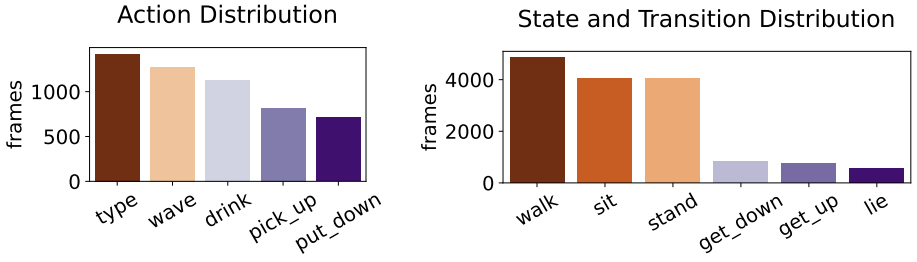


Fig. 5: Distribution of individuals in our dataset.

The action labels captured in our dataset showcase a diverse range of office activities. Actions such as `type`, `wave`, `drink`, `pick_up`, and `put_down` appear 1,420, 1,276, 1,129, 821, and 710 times, respectively as shown by Figure 6a. In terms of states, labels such as `walk`, `sit`, `stand`, and `lie` are featured 4,855, 4,065, 4,036, and 578 times, respectively. Transition labels, namely `get_down` and `get_up`, occur 821 and 770 times, respectively. The states and transitions happen exclusively and their distribution is shown in Figure 6b.



(a) Distribution of actions.

(b) Distribution of states and transitions.

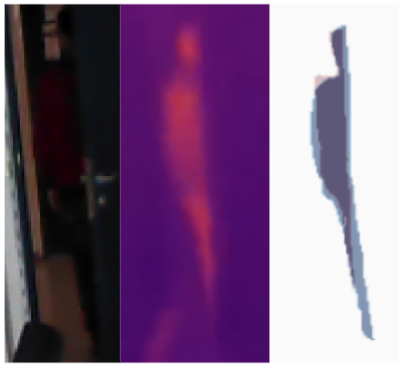
Fig. 6: Distributions of actions, states, and transitions in our dataset.

Finally, location labels `in_room`, `out_of_view`, and `out_of_room` are marked in 14,211, 1,163, and 234 instances, respectively.

Considering human segmentation, our dataset includes approximately 14,000 instances where humans are segmented within the frames. These instances, captured under varying scenes and lighting conditions, provide a dataset for human segmentation.

4.3 Dataset Quality Evaluation

To assess the consistency of the human segmentation labeling process in our dataset, 1106 frames are annotated twice. The quality of the labels is quantified through the calculation of the Jaccardian Index, which measures the overlap between two sets. The average Jaccardian Index is found to be **0.948**, indicating a high level of consistency between the different labelers. A further investigate of the frames with the 20 lowest Jaccard Indices, reveals two sources of errors. The first case, as depicted in Figure 7a, occurs when the labeled area is very small. This typically happens when the person is far away from the camera or mostly occluded. The second case, as shown in Figure 7b, arises from suboptimal labeling quality.



(a) First case: small labeled area due to a person leaving the room.

(b) Second case: discrepancy due to sloppy labeling.

Fig. 7: Visualization of RGB, thermal modality, and human segmentation masks. Violet, indicates an overlap between the two annotator's labels. Red and blue regions, signify areas of disagreement.

For labels, we compute the label agreement by comparing the labels given in both versions for the same frames. A match is considered when the same set of labels for a frame are provided, irrespective of the sequence. The results indicate a strong agreement for actions (92.4%), transitions (98.0%), states (94.7%), and locations (99.1%), confirming the robustness of our labeling process.

Figure 8 presents a confusion matrix for state transitions, providing insights into the instances where discrepancies occurred between the two labels. Of particular interest are the transitions from the `get_down` state to either the `walk` or `stand` states. These transitions often cause confusion due to the difficulty in pinpointing the exact moment when an individual begins to sit down. As the process of transitioning from a standing to a sitting or lying position is relatively brief, it results in a higher degree of error.

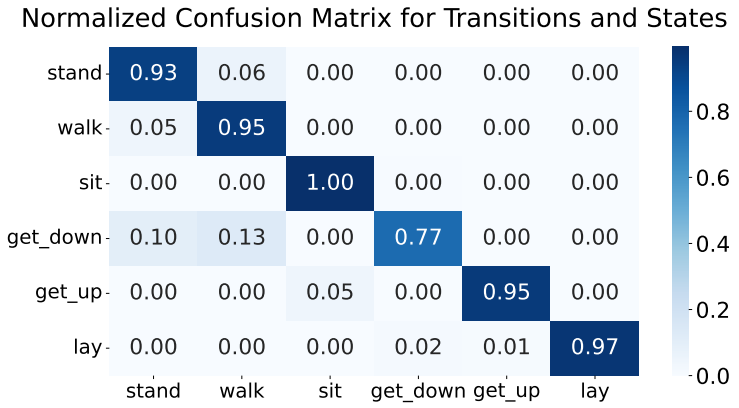


Fig. 8: Confusion matrix for state transitions in our double-labeled dataset.

5 Benchmarking

In our evaluation, we train a UNet and DeepLabV3 for human segmentation and an action detection model. To enable comparisons across different modalities, we employ z-normalization based on each modality's training set mean and variance. We adjust the size of each model's input channels and concatenate the normalized frames as per the combination of modalities in use. For example, combining depth and thermal data results in an input channel size of two, while employing all modalities increases the input channel size to five. These adjustments aim to assess the performance and impact of individual and combined modalities in multimodal learning tasks.

5.1 Split

The benchmark performance of models trained on our dataset requires a training, validation, and test split. The structure of this split is on the shot level rather than the frame level. This ensures that every frame within a shot is assigned exclusively to a single set (training, validation, or test set). Additionally, the selection of shots minimizes the overlaps in terms of offices and subjects

within training, validation, and test shots to prevent information leakage. The distribution the training, validation, and test sets, encompasses 63.77%, 18.23%, and 18.00% of the total data. This corresponds to 9,959 frames in the training set, 2,848 frames in the validation set, and 2,811 frames in the test set.

5.2 Human Segmentation

For the human segmentation task, we adopt the UNet [24] and DeepLabv3 [2] architectures, fine-tuning a model pretrained on the COCO dataset [18] for our trimodal data. We implement an early fusion technique, normalizing the input frames from each modality to a standard distribution via z-normalization, and then concatenating the respective modalities to form the multimodal input. To accommodate these inputs, the DeepLabv3 model-originally designed for RGB inputs-has an input channel added for thermal and depth modalities. When including the RGB modality, we copy the original RGB weights to the new input layer. Following this, the model undergoes ten epochs of training with a learning rate of 0.0001. The model yielding the lowest validation loss during this process is selected for testing. Interestingly, as revealed by Tables 3a and 3b, the best performing combination—yielding the highest Intersection over Union (IoU)—excludes the RGB modality. This finding underscores the value of thermal modality in achieving clear human visibility, especially in test scenes with RGB clutter.

Table 3: Results for Segmentation using UNet and DeepLabv3 on the test set. The input layers channel are updated to accommodate the concatenation of the models.

(a) Results for UNet.

RGB	Depth	Thermal	Loss	IoU
–	–	✓	0.040	0.659
–	✓	–	0.055	0.580
–	✓	✓	**0.020**	**0.775**
✓	–	–	0.147	0.356
✓	–	✓	0.062	0.673
✓	✓	–	0.071	0.553
✓	✓	✓	0.025	0.726

(b) Results for DeepLabv3.

RGB	Depth	Thermal	Loss	IoU
–	–	✓	0.041	0.660
–	✓	–	0.045	0.622
–	✓	✓	**0.023**	**0.806**
✓	–	–	0.050	0.586
✓	–	✓	0.041	0.670
✓	✓	–	0.086	0.494
✓	✓	✓	0.048	0.619

5.3 Action Detection

Our approach for temporal action detection builds upon the method presented in [33]. We employ the same early fusion technique as in the segmentation task to fuse the multimodal inputs. The model is initialized with random weights. Its

input is a set of eight frames at a time: the first seven frames serve as temporal context, and the eighth frame is the prediction target. The model architecture includes four 3D convolution pooling blocks with ReLU for feature extraction, global average pooling, and two Multi Layer Perceptrons (MLPs) for classification. One MLP classifier with softmax and cross-entropy loss is utilized for mutually exclusive state, transition, and location labels. The second MLP with sigmoid and binary cross-entropy loss is employed for action labels, accounting for the possibility of simultaneous actions. As demonstrated in Table 4, the combination of depth and thermal modalities yields the highest performance for action classification. This finding underlines the results from the segmentation tasks, reinforcing the advantages of utilizing non-RGB modalities, particularly in complex scenes, for robust action recognition.

Table 4: Results for Temporal Action Detection using custom 3D Convolution Architecture on the test set.

RGB	Depth	Thermal	Loss	Accuracy	Precision	Recall
–	–	✓	2.367	0.903	0.796	0.620
–	✓	–	2.504	0.889	0.749	0.577
–	✓	✓	**2.347**	**0.907**	**0.813**	0.626
✓	–	–	2.659	0.876	0.704	0.537
✓	–	✓	2.346	0.904	0.799	0.623
✓	✓	–	2.465	0.897	0.758	**0.629**
✓	✓	✓	2.349	0.901	0.783	0.618

6 Conclusion

In this work, we have introduced a novel trimodal dataset that combines RGB, thermal, and depth data captured in diverse office environments. One finding from our experiments is the superior performance achieved by utilizing the depth and thermal modalities, even surpassing the combination of RGB, depth, and thermal data. This finding underlines the role that these less traditionally utilized modalities can play in enhancing the robustness and performance of machine learning models, particularly in environments with varying lighting conditions. Future research could exploit the temporal characteristics of our dataset. While UNet and DeepLabV3 architectures mainly focus on spatial features, incorporating temporal information could provide a richer context for the segmentation task. Leveraging Transformer models, which have demonstrated capability of capturing temporal dependencies in data, could also be considered.

Acknowledgments

This work was partly supported by the Austrian Research Promotion Agency (FFG) under the Grant Agreement No. 879744.

Bibliography

[1] Brenner, M., Reyes, N.H., Susnjak, T., Barczak, A.L.: Rgb-d and thermal sensor fusion: A systematic literature review. arXiv preprint arXiv:2305.11427 (2023)

[2] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)

[4] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. Ieee (2005)

[5] Davis, J., Keck, M.: A two-stage approach to person detectionin thermal imagery. In: Proceeding of Workshop on Applications of Computer Vision (WACV) (2005)

[6] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: 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)

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[8] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)

[9] Gao, C., Du, Y., Liu, J., Lv, J., Yang, L., Meng, D., Hauptmann, A.G.: Infar dataset: Infrared action recognition at different times. Neurocomputing **212**, 36–47 (2016)

[10] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)

[11] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

[12] Heitzinger, T., Kampel, M.: A foundation for 3d human behavior detection in privacy-sensitive domains. In: 32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021. p. 305. BMVA Press

(2021), https://www.bmvc2021-virtualconference.com/assets/papers/1254.pdf
[13] Heitzinger, T., Kampel, M.: Ipt: A dataset for identity preserved tracking in closed domains. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8228–8234. IEEE (2021)
[14] Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
[15] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (Jan 2023), https://github.com/ultralytics/ultralytics
[16] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
[17] Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G., Mizginov, V.: Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 606–624 (2018)
[18] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
[19] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
[20] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
[21] Miezianko, R.: Terravic research infrared database. IEEE OTCBVS WS Series Bench (2005)
[22] Palmero, C., Clapés, A., Bahnsen, C., Møgelmose, A., Moeslund, T.B., Escalera, S.: Multi-modal rgb–depth–thermal human body segmentation. International Journal of Computer Vision **118**, 217–239 (2016)
[23] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
[24] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
[25] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
[26] Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)

[27] Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: Pst900: Rgb-thermal calibration, dataset and segmentation network. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 9441–9447. IEEE (2020)
[28] Sigurdsson, G.A., Divvala, S., Farhadi, A., Gupta, A.: Asynchronous temporal fields for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 585–594 (2017)
[29] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 510–526. Springer (2016)
[30] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. ECCV (5) **7576**, 746–760 (2012)
[31] Strohmayer, J., Kampel, M.: A compact tri-modal camera unit for rgbdt vision. In: 2022 the 5th International Conference on Machine Vision and Applications (ICMVA). pp. 34–42 (2022)
[32] Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2022), https://github.com/heartexlabs/label-studio
[33] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
[34] Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7464–7475 (2023)
[35] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
[36] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)