Cross-modal Assisted Training for Abnormal Event Recognition in Elevators

Xinmeng Chen*[1], Xuchen Gong*[1], Ming Cheng[1], Qi Deng[2], Ming Li[+1]

[1]Data Science Research Center, Duke Kunshan University, [2]Technology Asia and Escalator (CRD), KONE Elevators

*Both authors contributed equally to this research , [+] Corresponding author

Abstract

- We collect and propose our multimodal dataset investigating passenger safety and inappropriate elevator usage for the task of abnormal event recognition in elevators.
- We present the RGBP framework to utilize multimodal data to enhance unimodal test performance. Experimental results show RGBP improves the unimodal inference performance on the Elevator RGBD dataset by 4.71% (acc.) and 4.95% (F1 score) with respect to the pure RGB model.
- Our RGBP framework outperforms two other methods for "multimodal training and unimodal inference": MTUT [1] and the two-stage method based on depth estimation.

Related Work

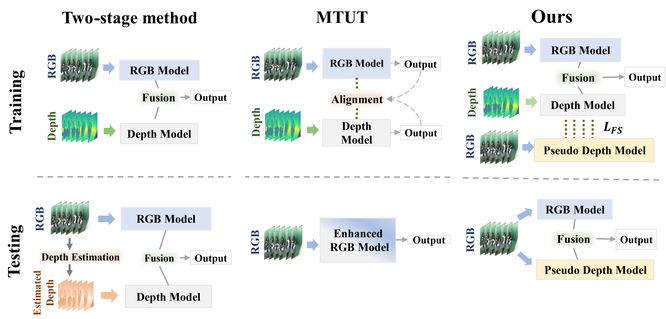


Figure 1: Comparison of our RGBP framework with two other frameworks that use depth and RGB data in the training stage and only RGB data in the inference stage. a) a two[1]stage framework that involves depth estimation. b) MTUT c) RGBP (proposed).

- Abnormal event recognition: detect particular scenarios in elevators, such as violent behaviors and passenger emergencies.
- Late fusion and intermediate fusion [2] have been employed to boost performance.
- For Multimodal Training and Unimodal Inference, the method MTUT encourages a model of one modality to utilize the learned information from multiple modalities.

Data

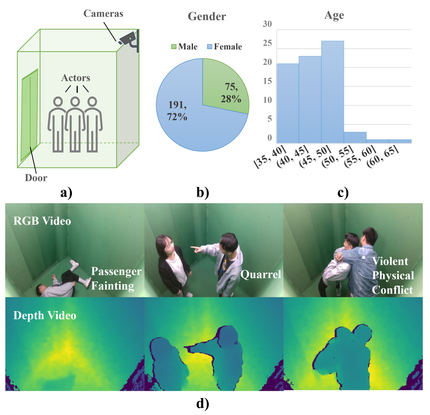
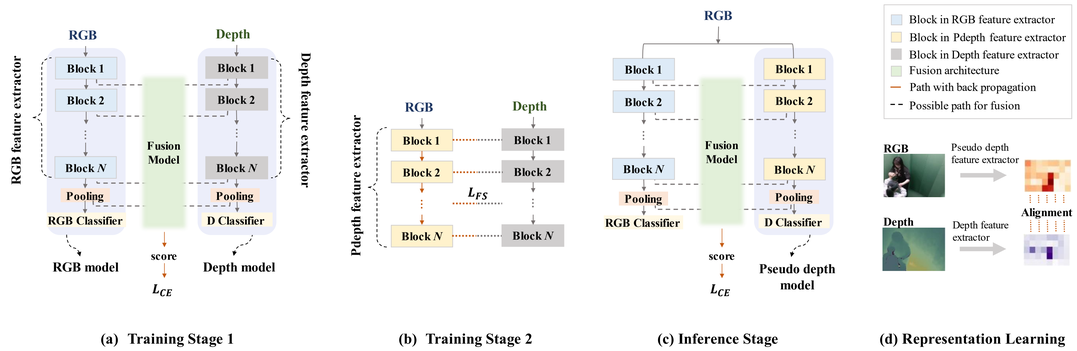


Figure 2: Data collection of the Elevator RGBD Dataset. (a) The elevator cabin for data collection, composed of a green screen as the background, an RGBD camera, and a door. (b) The gender distribution of the actors. (c) The age distribution of the actors. (d) Some examples of RGB and depth frames in the Elevator RGBD Dataset.

Table 1: The 5 events in the Elevator RGBD Dataset and their corresponding number of samples.

Event Name	Sample Number
Normal	14896
Passenger fainting	2128
Quarrel	2128
Force opening the door	2128
Violent physical conflict	4256

Method



(a) Training Stage 1 (b) Training Stage 2 (c) Inference Stage (d) Representation Learning

Figure 3. The RGBP framework works in three stages: two training stages and one inference stage. In training stage 1, the pre-trained RGB and depth model are fused, with only the fusion model trainable. In training stage 2, the frozen depth model supervises the pseudo depth feature extractor for learning depth data representations. Then, to assemble the model for testing, for the RGBD model trained in (a), its depth model is replaced by the pseudo depth model trained in (b). Then, the resulting model RGBP is ready for inference.

Feature Similarity Loss:

$$L_{FS} = \sum_{p}^{P} \frac{1}{D \times D} \lambda_p \left\| corr\left(F_p^{depth}\right) - corr\left(F_p^{pseudo}\right) \right\|_F^2 + \sum_{q}^{Q} \frac{1}{W \times H \times T \times C} \lambda_q \left\| F_q^{depth} - F_q^{pseudo} \right\|_F^2$$

Architecture Analysis

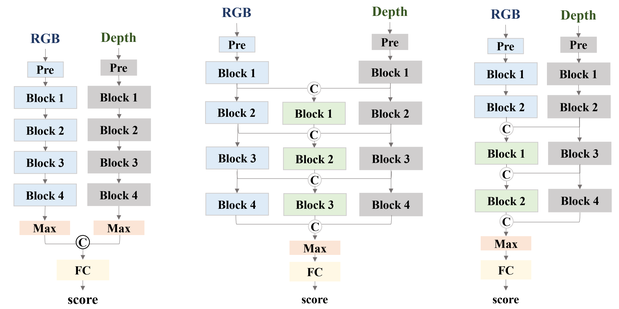


Figure 4: Three RGBD architectures we use to test our method. In this figure, the RGB (blue) blocks belong to the RGB feature extractor, and the depth (grey) blocks belong to the depth feature extractor.

Table 2: Comparison of the test accuracies of different architectures and methods.

Method	Acc.	# Parameters
Pure RGB	83.12	8.77 M
Pure Depth	83.04	8.63 M
RGB + Depth	86.71	17.54 M
$2\text{RGB}_{\text{Late}}$	86.28	17.55 M
$RGBD_{Late}$	**89.08**	17.54 M
$2RGB_{Center}$	83.80	266.01 M
$RGBD_{Center}$	88.37	266.00 M
$2RGB_{Left}$	85.74	25.79 M
$RGBD_{Left}$	87.85	25.78M

- Simply adding the output scores of the RGB and depth models enhances performance.
- Three RGBD models perform better than their corresponding two-stream RGB models.

Pseudo Depth Analysis

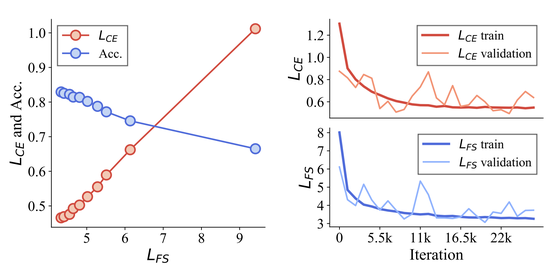


Figure 5. Correlation between L_{FS}, L_{CE}, classification accuracy and training iteration of the pseudo depth model. Left: L_{CE} and Acc. vs. L_{FS}. Right: L_{CE} vs. iteration and L_{FS} vs. iteration.

- On the training set, as L_{FS} decreases, pseudo depth model's classification loss decreases, and its classification accuracy increases. → L_{FS} provides effective supervision.
- The feature maps output by the four blocks in the pseudo depth model look more and more similar to that of the depth model.

Results

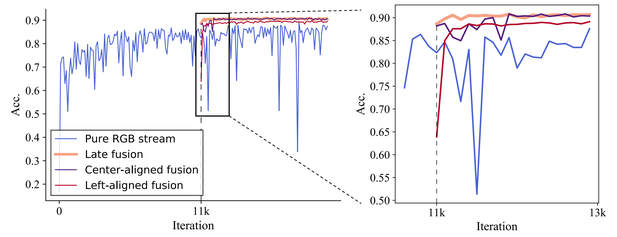


Figure 6. Comparison of validation accuracy of the pure RGB model with those of the late fusion mode ($RGBD_{Late}$), the intermediate center-aligned fusion model ($RGBD_{Center}$), and the intermediate left-aligned model ($RGBD_{Left}$).

Table 3: Comparison of our RGBP frameworks with other frameworks on the Elevator RGBD Dataset.

Method	Acc.	F1 score	Avg Time	FLOPs
MTUT	86.05	85.44	7.31ms	3.97G
DE_{Left}	70.56	69.72	1069.87ms	605.19G
DE_{Center}	85.17	84.64	1093.20ms	625.79G
DE_{Late}	85.40	84.91	1068.64ms	604.84G
$RGBP_{Left}$ (Ours)	86.42	86.18	15.57ms	8.29G
$RGBP_{Center}$ (Ours)	86.52	85.69	43.34ms	28.89G
$RGBP_{Late}$ (Ours)	**87.83**	**87.64**	14.95ms	7.94G

- Our RGBP framework improves the acc. by 1.78% with respect to the MTUT model and 2.43% with respect to DE_{Late}. Moreover, our RGBP framework improves the F1 score by 2.20% with respect to the MTUT model and 2.73% with respect to DE_{Late}.
- Using depth estimation during inference leads to accumulated error, with DE_{Late} and DE_{Center} being relatively more robust to the error in the estimated depth.

Conclusion

- Aimed for "multimodal training and unimodal inference," our proposed RGBP framework takes RGB data alone as input while make a prediction based on both the RGB features and the predicted depth features.
- This method utilizes rich training data to improve the model performance in the case of limited data, which is valuable in many practical applications.
- In future work, the technique for generating pseudo depth features from RGB data is expected to have applications beyond the elevator scenario and abnormal event recognition domain.

Reference

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. 2019. Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition With Multimodal Training. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 1165–1174.

[2] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Suk thankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 1725–1732.