Fall Detection: Leveraging Depth Information in Bayesian Networks

Marc Oliu[1,2] mao@milestone.dk
Rohat Bozyil[2] robo@milestone.dk
Mia Siemon[1,2] misi@milestone.dk
Alejandro Martinez-Senent[2] armt@milestone.dk
Sergio Escalera[1,3] sescalera@ub.edu
Thomas B. Moeslund[1] tbm@create.aau.dk
Kamal Nasrollahi[1,2] kn@create.aau.dk

Aalborg Universitet[1]
Fredrik Bajers Vej 7K, Aalborg Øst

Universitat de Barcelona and
Computer Vision Center[3]
Gran Via de les Corts Catalanes 585, Barcelona
Campus UAB, Edifici O, Cerdanyola del Vallès

Milestone Systems[2]
Banemarksvej 50, Brøndby

Abstract

Deaths due to accidental falls is a pervasive issue across the healthcare system, affecting both professional settings such as hospitals and nursing homes as well as private residences where elderly people might live either alone or without supervision during part of the day. With them being the second leading cause of accidental deaths, just behind traffic accidents, there is significant incentive to develop measures capable of reducing the impact of such accidents. An accurate Fall Detection system presents itself as a viable solution, whereas an automated video surveillance system is capable of raising an alert when such an event occurs. For such an approach to be commercially viable, it must both have a high degree of accuracy and be trainable with easily obtainable data. The first is necessary in order to reduce the number of false positives, and thus not burden the healthcare personnel with false alerts. The second would make it possible to implement the system without depending on datasets available for research purposes only, nor requiring a large time investment on creating a private dataset, with the privacy concerns this entails.

We address this by introducing an Anomaly Detection approach based on Bayesian Networks. Our approach models a given video frame based on the relationship between simple features extracted from the image, and does not require any kind of private information nor class labels to work. This makes our model both privacy-preserving and with low data requirements. Furthermore, the model can be trained in just a few seconds. We achieve results that far surpass the current state-of-the-art when compared to other unsupervised approaches.

1. Introduction

Falls are the second foremost cause of accidental death worldwide according to the World Health organization [15], with the likelihood of such events occurring increasing with age and being particularly pervasive among elderly people. Among said portion of the population, it is not only much more frequent, but the likelihood of a fall leading to serious injury or even death increases significantly. This is of particular concern to the healthcare system, where a series of issues arise because of it. First is the issue of monitoring large spaces such as hospitals and nursing homes, where it is important to quickly react to any such events not only to reduce the impact of such accidents, but also to reduce the number of lawsuits placed on the hospital or staff. Another point of concern are private residences, where elderly people often live alone during part or most of the day. In such cases where no one is available to provide help on short notice, a monitoring system would be capable of alerting the healthcare system to provide aid.

This issue has long been addressed by the Computer Vision community, but most such solutions have in place several restrictions that make it extremely difficult to implement a commercially viable solution. This stems mainly from the dependence of most models on labeled datasets, where each frame is assigned a binary label according to whether a fall is occurring or not. This poses a dual problem: On one hand, while such datasets are available for research purposes, they are restricted for commercial use. When implementing a commercial solution it then becomes necessary to create a private dataset, with all of the privacy and labeling concerns this entails. On the other hand, most modern approaches depend on deep neural networks trained specifically to tackle Fall Detection. This results in a large number of trainable parameters, necessitating large

scale datasets to train the model without over-fitting.

We solve the first issue by introducing an unsupervised classifier, making it possible to depend on unlabeled data to train the model. This also allows for in-situ training of the model: The video feed of the installed camera can be directly used to train the model, further increasing the accuracy by tailoring the model to the idiosyncrasies of the space that is being monitored. The second issue is solved through the use of off-the-shelf models to extract useful features from the frame, which is then combined using a Bayesian Network (BN). With BN being a type of probabilistic model with a reduced number of parameters, it is possible to drastically reduce the amount of data needed to train the model.

2. Related works

Fall Detection as a Computer Vision problem has seen significant interest over the years. Historically, approaches have been based on measuring simple low-level indicators such as head velocity [14, 18] and simple body shape and motion measurements [8, 13]. In contrast, most modern approaches depend on the training or fine-tuning of various Deep Neural Network architectures, resulting in models that take a significant amount of time to train and cannot be easily adapted to specific scenes and camera angles. These can be roughly separated into two main categories: Pose-based and detection based.

Pose-based approaches can be considered as the successors of the more classical approaches, where more complex features are used to measure body shape. These first detect individuals in a scene and then extract their body pose. Q. Xu *et al.* [16] use OpenPose as an off-the-shelf model to detect individuals in a frame and extract their skeletons. A classifier is then trained to predict whether the skeleton configuration corresponds to the fall or non-fall category. Similarly, S. Juraev *et al.* [5] follow the same pipeline but using a transformer model. Each token encodes the body pose captured at a given frame, with the output token embeddings being used to classify each frame as either fall or non-fall. The same pipeline is followed by S. McCall *et al.* [7], except for the transformer-model being pre-trained on a general dataset of 2D poses and later fine-tuning it for Fall Detection.

Detection based methods, on the other hand, approach the task similarly to that of object detection, where a fallen person is viewed as just another object instance. Models falling in this category essentially push the decision of which features to consider relevant for the task to the object detector, significantly reducing the need for model design choices and increasing its flexibility at the cost of a higher model complexity, training times and data requirements. A. Bansal *et al.* [2] propose fine-tuning a Faster R-CNN architecture, pre-trained on the COCO dataset. Their model is fine-tuned to propose bounding boxes for the various individuals in the scene, providing a score for fall and non-fall instances. In essence, fall and non-fall instances are treated as separate objects to be localized. A similar approach was followed by Y. Ke *et al.* [6], but using a YOLOv5 detector instead.

Note that in both cases supervised training is required. This is not so for the method by J. Gonzalez *et al.* [3], the most similar method to our own, which considers the problem from the point of view of Anomaly Detection. Said approach trains a BN to model non-fall instances based on simple features extracted using off-the-shelf models, such as bounding boxes, relative location of the head and absolute position of a person relative to the image plane. This simple approach turns out to be extremely effective, achieving results on par with the state-of-the-art on the CAUCAFall dataset.

3. Proposed method

The proposed approach is based on BN, and more specifically is an extension of the Fall Detection approach proposed by J. Gonzalez *et al.* [3]. It can be considered as a combination of classical and modern techniques, where on one hand simple features are combined through a simple probabilistic model to describe sample likelihoods, while on the other hand complex deep architectures are used to extract said features.

One of the main sources of inaccuracy in the original approach is usage of the grid cell as a prediction target, resulting in the marginalization of both aspect ratio and head position during evaluation. Additionally, the 3D geometry of the scene is ignored, overlooking one of the potentially richest visual cues. Our main contributions are twofold, and aim at addressing the above mentioned issues:

- Use state probability to estimate the fall score
- Introduction of depth information

The overall pipeline of the model can be seen in Fig. 1. Similarly to the original approach, an off-the-shelf object detector is used to extract the bounding box of individuals and their head locations. The candidate locations of the person in the image are determined based on the overlap between the lower edge of their bounding box and a fixed-size grid cell overlaid on the image. These features are explained in more detail in Sec. 3.2. In addition, our model encodes the 3D angle between the floor and an individual, as well as the distance between them, through a combination of depth estimation and segmentation. This is explained in more detail in Sec. 3.1. Finally, Sec. 3.3 explains the changes to the computation of the likelihood for an observation.

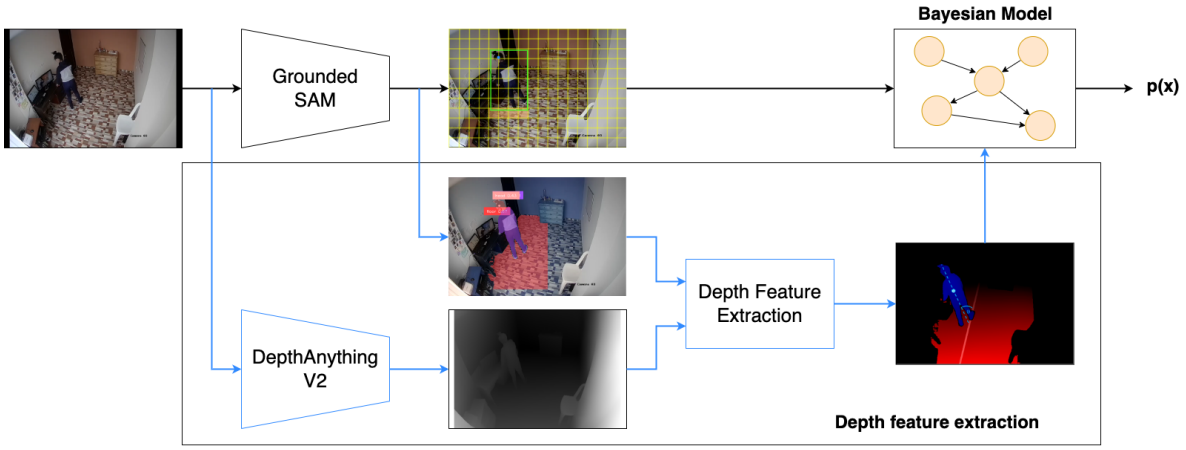

Figure 1. Evaluation pipeline for the proposed approach. Depth feature extraction is introduced by our model, and consists of depth estimation, object segmentation, and feature extraction through linear and SVD fitting of the point clouds of the individual and ground, respectively. Extraction of bounding boxes, head location and absolute positioning are the same as in the original approach [3].

3.1. Depth information

Simple features such as the bounding box ratio and head position provide limited information with regards to distinguishing some fall cases from certain camera angles, such as falling forward or backward, from non-fall cases such as sitting or crouching. As such, a significant improvement to the original approach is the introduction of depth-based features capable of distinguishing such cases. More specifically, we compute two pieces of information: the angle of an individual relative to the ground plane, and the distance between the ground and the centroid of the individuals point cloud. These are illustrated in Fig. 2

In order to extract these two variables, we use an off-the-shelf monocular depth estimator (Depth Anything V2 [17]) to estimate a depth map of the image. We then use the Grounded SAM pipeline [12], consisting of Grounding DINO 1.5 [11] and Segment Anything V2 [10], to obtain the bounding boxes of individuals and their heads, as well as the pixel segmentation of both individuals and the ground plane. The segmentation information is then combined with the depth map in order to obtain the point cloud of the individuals and ground plane.

We then calculate the centroid of each individual along with the best fit 3D line through linear regression, which corresponds to the main axis of the individual. The parameters of the ground plane are extracted through Singular Value Decomposition (SVD). The covariance matrix of the point cloud corresponding to the ground plane is extracted and the two Principal Components of the distribution computed. These correspond to two orthogonal vectors along the plane, from which the plane parameters are easily obtained. With these intermediate variables extracted, we obtain the two target parameters in the following manner:

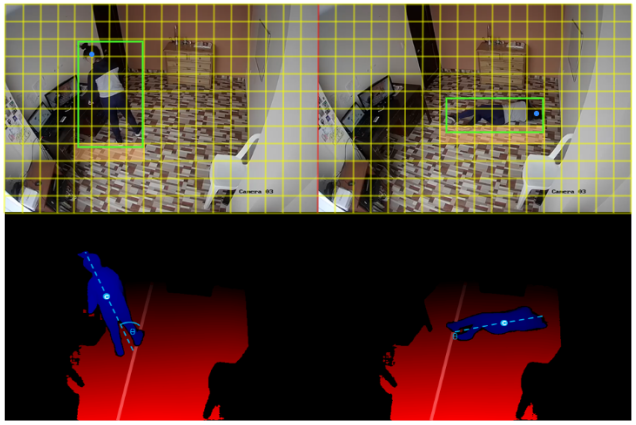

Figure 2. **Top:** We extract three variables from the RGB information of a frame. The bounding box of an individual (green box), the head position (light blue dot) and the grid cells locations of the bounding box (marked in orange). **Bottom:** We extract two continuous variables from the depth information of a frame. The distance between an individuals centroid (light blue dot) and the ground plane (red area), and the angle θ between the individual and the floor plane.

For the distance between the ground and the centroid of the individual, we compute the distance between the centroid and ground plane. For the angle between the individual and ground plane, we first project the main axis onto the ground plane, then compute the cosine similarity between the main axis and said projection.

Finally, we condense both of those parameters into a single discrete variable that serves as a node for the BN. To do so, we first compute the multivariate normal distribution of these two variables over all samples in the training set. In order to assign a discrete value to a sample, we com-

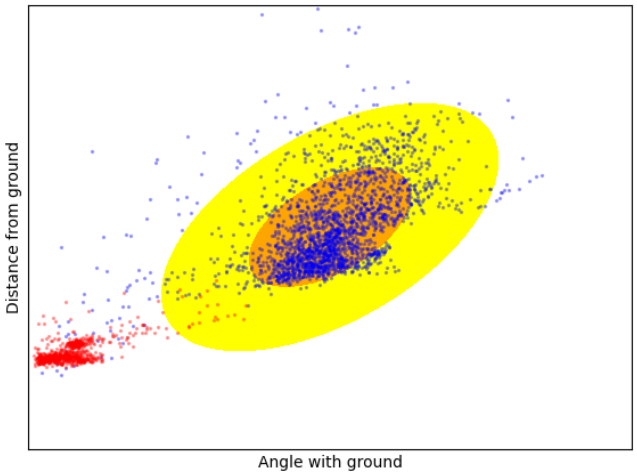

Figure 3. Multivariate normal distribution of the two variables dependent on depth information, as well as the three categorical values for the resulting categorical variable. **White:** PDF below 0.025. **Yellow:** PDF in between 0.025 and 0.5. **Orange:** PDF above 0.5. Non-fall training samples are marked in blue and fall test samples in red.

pute its Probability Density Function (PDF) based on the extracted distribution, then threshold the probability based on manually selected threshold values. In our experiments, we obtained the best results for three discrete values with thresholds at 0.025 and 0.5. These are outlined in Fig. 3 and correspond to the white, yellow and orange areas. Negative test samples are marked in red. Note that while this distribution is not a precise classifier, with many false positive samples falling within the white low probability region, those are for samples with a high angle or distance relative to the ground. Such cases are easily identified by other variables in the BN, such at the aspect ratio of the bounding box and head position.

3.2. Bayesian model

The overall model consists of a BN based on the approach by J. Gonzalez *et al.* [3]. The BN consists of five variables, with the first four being the same as in the original work. The connectivity between those is shown in Fig. 5 for both the original model and the one we propose. For our model, we maintain the same logical relationship between variables and establish a logical relationship between the depth-based variable and the aspect ratio and head position. This is due to the depth information encoding basic information on the angle of the person, which is indirectly reflected on the values for both the aspect ratio and head position. The information encoded by each variable is the following.

Scene is a categorical variable representing the scene being displayed. Each different scene or camera perspective corresponds to a different value.

GridCell represents the grid cell to which the detected individual is assigned. An input image is divided into a regular grid, with each cell corresponding to a different value for the categorical variable.

AspectRatio represents the aspect ratio of the bounding box for an individual. It consists of two possible discrete values. Portrait when the height of the bounding box is larger than its with, and landscape when the width is larger than the height.

HeadPosition encodes the coordinates of the center of the individuals head relative to the body. It is a categorical variable with two possible values representing whether the had position is either above or below the central point of the body.

Depth encodes information on the orientation of the individual relative to the ground and the distance between the individuals centroid and the ground. The information is obtained and encoded according to Sec. 3.1.

3.3. State probability

The original approach determined the likelihood of a sample belonging to a specific grid cell given the observed values for the other nodes of the BN. Given that each observation can fall on a number of grid cells (each grid cell that intersects with the lower border of an individuals bounding box), the authors computed the probability of each of the intersecting grid cells and then obtained the average probability.

Our approach, on the other hand, provides the grid cell information as an additional input observation, computing the overall probability of the BN state for each of the potential grid cells. The maximum probability is then kept as the final probability score for the sample. These changes are introduced due to two main reasons. Firstly, we are training the model as an anomaly detection approach, meaning that only non-fall samples are available during training. As such, the state probability of the BN given a sample represents how well said sample fits with the non-fall data distribution. Secondly, given that any of the valid variable configurations has a high probability, the sample itself follows the training distribution. As such, taking the maximum state probability is expected to be much more accurate than computing the average.

4. Experimental setup

We consider two datasets for our experiments: CAUCAFall [4] and High-Quality Fall Simulation Dataset (HQFSD) [1]. Both datasets are used for an ablation study where we aim to determine how each of our two main contributions, namely the use of state probability for anomaly detection and the introduction of depth information, affect

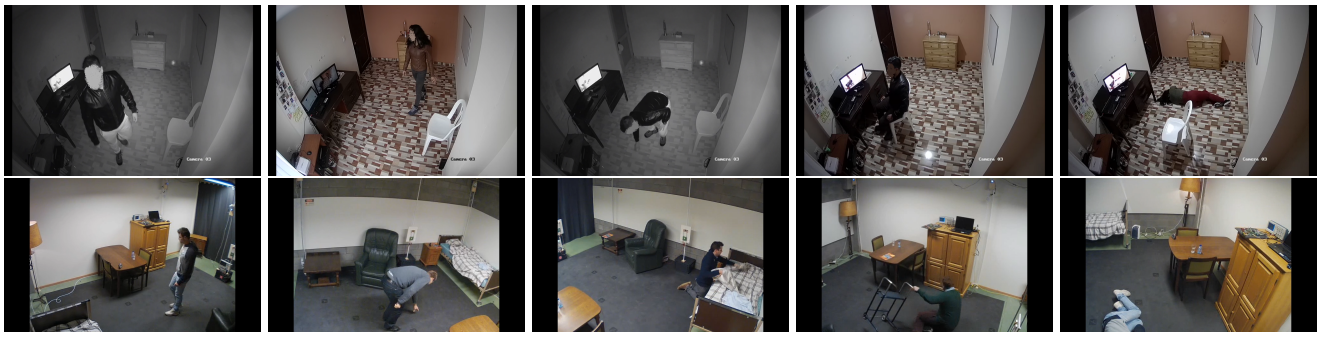

Figure 4. Samples from the CAUCAFall (**top**) and HQFSD (**bottom**) datasets, displaying various ADL and individuals. CAUCAFall displays wide variations in lighting conditions, while HQFSD is recorder from multiple points of view and displays a wider array of ADL.

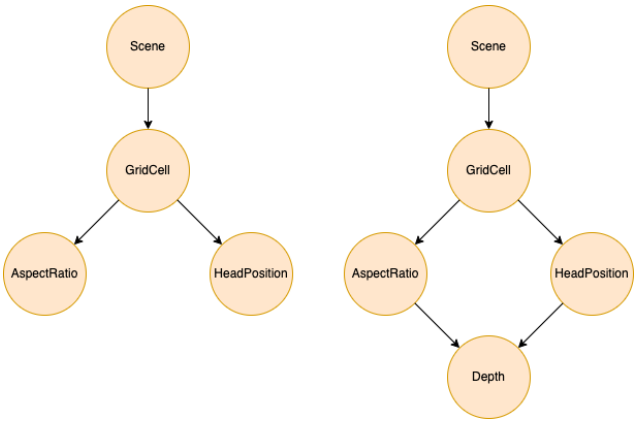

Figure 5. **Left:** BN used by the baseline model. **Right:** BN introducing angular information on the individuals orientation and distance between the ground and individuals centroid.

the overall performance of the classification model. Furthermore, we perform a state-of-the-art comparison on the CAUCAFall dataset.

4.1. Datasets

Two datasets are considered for our experiments, samples of which can be seen in Fig. 4. The CAUCAFall dataset [4], a small-scale video-based Fall Detection dataset consisting of video recordings of 10 subjects and consisting of 7388 fall and 12366 non-fall frames. This is a posed dataset where the subjects fall in different manners as well as performing some simple daily activity actions. More specifically, there are 5 types of falls (fall backwards, fall forward, fall left, fall right, fall sitting) as well as 5 simple Activities of Daily Living (ADL) that might result in false positive classifications (hop, kneel, pick up object, sit down, walk). Regarding the individuals, there are variations in terms of age, height, gender and clothing. As for the scene, it consists of a single indoors room recorded from the same perspective using a HIKVISION IR camera in AVI format at 23 frames per second (FPS). The space contains occlusions and variations inn lighting conditions (natural, artificial, night). We prioritize this dataset because it offers a robust representation of various scenarios commonly encountered in hospitals, presenting realistic results obtained from the experiments.

The second dataset is HQFSD [1]. Similarly to the previous one, this dataset consists of posed video recordings of 10 different subjects both performing ADL and falling. It consists of a total of 55 fall and 17 ADL scenarios recorded within the same room using 5 different cameras placed at different locations, for a total of 275 video recordings. The recordings are of variable length, spanning between $0:50$ and $4:58$ minutes for fall sequences, and between $11:38$ and $35:50$ minutes tor ADL. The dataset varies in terms of falling speeds, moving objects, subject ages, and the ADL being performed right before the fall.

Contrary to the previous dataset, HQFSD is annotated with the falling frames (from when someone loses balance until they reach the ground), instead of the fall itself. We re-annotate the dataset for Fall Detection by using the last timestamp of the original annotations as the beginning of the fall sequence, and manually annotate the end of the sequence as the moment when the person stands back up again. Our model is based on unsupervised learning, predominant in the field of Anomaly Detection, and there is no standard partitioning for this dataset. Due to this, we choose to split the dataset into a training partition consisting of the ADL sequences, and a validation partition with all of the fall sequences.

4.2. Evaluation metrics

We use multiple metrics when evaluating the performance of our model. Accuracy is the most straightforward, determining the overall classification accuracy of the model given the classification threshold maximizing the number of correctly classified frames on the overall dataset. A related metric is the Area Under the ROC Curve (AUC). This met-

ric represents the probability of the model ranking a positive sample higher than a negative one, given two randomly chosen posive and negative samples. An ideal model would have an AUC score of 1, while a completely random one would have an AUC of 0.5. This metric is theshold agnostic, and gives an impression of the overall discriminative prowess of the model, as opposed of measuring the performance for a given threshold like the accuracy metric does.

The problem of Fall Detection in video sequences can also be seen from the perspective of object tracking, where we aim to find the sequence of bounding boxes corresponding to a person on the ground. For this kind of problem two commonplace anomaly detection metrics are the Region-Based Detection Criterion (RBDC) and Track-Based Detection Criterion (TBDC) [9]. RBDC is a measure of the number of correctly detected anomalies in a per-frame level, where a fall prediction is considered correct if the intersection-over-union of the detected fall is grater than a certain threshold. TBDC measures the detection of anomalies over time, with the fraction of frames where the anomaly is correctly detected must surpass a given threshold.

Other metrics used during both the ablation study and comparison with the state-of-the-art are those associated with the statistical analysis of binary classification models, namely: precision, recall, specificity and F1-score.

5. Results

As previously discussed, we propose two main improvements over the original approach [3]. An ablation study is shown in Tab. 1. From the results we can see that just by computing the state probability of the sample, as opposed to trying to predict the grid cell ID like the original work did, we obtain a significant jump in accuracy of 6.04%. The AUC and RBDC scores also increased by 2.3% and 0.21% respectively. When looking at the individual samples, we can see that most of that gain comes from the samples 'Fall sitting' and 'Fall forward', both experiencing large jumps in accuracy. On the other hand, other samples like 'Fall backwards (10)' and 'Fall left (4)' experienced a minor but significant drop in accuracy despite their AUC score remaining the same or even improving. This is due to the classification threshold being selected at the global level and being shared by all samples, meaning that the relative scores of positive and negative frames might have actually improved within that video, but not relative to all frames within the full dataset.

When further introducing depth information to the model, we obtain another significant increase of 3.33% on the overall accuracy, with the AUC increasing by 0.34% and RBDC marginally by 0.07%. At the individual sample level, its interesting to notice that the individual accuracies correspond to the maximum of either the original approach and the one based on state probability, with there being further marginal gains in therms of both AUC and RBDC. This implies that the additional information mainly adds robustness to the model as well as providing further gains in terms of being capable of distinguishing between fall and no fall samples: No-fall samples obtain higher likelihood scores and those of fall samples decreases, resulting in a better sorting of the samples that increases the AUC score, but that doesn't necessarily translate to higher accuracy gains. While the overall response of the classifier has indeed improved, the optimal threshold maximizing the overall classification accuracy retains the same performance. This is further highlighted by the TBDC score, which remains unchanged after introducing depth information. This implies that while there is an increase on the number of correctly classified frames, this does not translate to a higher number of correctly classified fall sequences. In other words, most of that improvement comes from better classifying individual frames on already correctly predicted fall sequences.

In Tab. 2 we compare our method against previous works. We can see that our approach achieves state-of-the-art results on CAUCAFall, outperforming both the previous approach it is based on [3] as well as all other approaches trained on said dataset, while still maintaining a competitive speed. In fact, the results obtained approach the maximum possible accuracy, leaving little margin for future improvement. It is important to notice that our approach is slower than other similar methods due to the complexity of the feature extraction step. This is mainly due to the use of Grounded SAM for instance segmentation. The model could be heavily optimized by using a custom person and floor segmentation model, which is something to consider as future work.

The ablation of our model on the HQFSD dataset is shown in Tab 3. This dataset is much more challenging than the previous one, with multiple individuals occasionally appearing on a single frame and some of the camera views not always fully capturing the individual, which can happen even during fall instances. This is reflected by the much poorer overall performance on this dataset.

The use of the state probability score more significantly increases the overall performance of our model, increasing both the precision and specificity of the model even if that comes at the cost of a slight decrease in recall. Overall, it can be seen from the AUC score that the classification model is much better behaved when compared to the baseline. On the other hand, the addition of depth features provides a marginal increase to the performance of the model. For the CAUCAFall dataset, we had already seen that the bulk of the improvement in accuracy due to depth information came from the additional features proving robustness to the model, but no additional information capable of increasing the model performance above that of the best pre-

	Original Method				State probability				Depth Information			
	Accuracy	AUC	RBDC	TBDC	Accuracy	AUC	RBDC	TBDC	Accuracy	AUC	RBDC	TBDC
Fall left (4)	98.59	98.21	99.57	100	84.04	98.51	99.57	100	**99.59**	**99.82**	**99.91**	**100**
Fall backwards (4)	97.93	98.26	99.35	100	97.93	99.14	99.62	100	**97.93**	**99.90**	**99.95**	**100**
Fall backwards (10)	99.27	98.72	99.27	99.27	87.64	98.72	99.27	99.27	**99.27**	**98.72**	**99.27**	**99.27**
Fall right (1)	97.39	97.32	97.39	97.39	97.39	97.32	97.39	97.39	**97.39**	**97.32**	**97.39**	**97.39**
Fall sitting (2)	56.96	93.31	96.84	96.84	96.84	94.43	96.84	96.84	**96.84**	**94.43**	**96.84**	**96.84**
Kneeling (3)	100	-	-	-	95.33	-	-	-	**100**	-	-	-
Fall forward (4)	42.01	70.06	95.37	98.82	96.45	82.67	97.24	98.82	**96.45**	**82.67**	**97.24**	**98.82**
Sit down (5)	100	-	-	-	98.08	-	-	-	**100**	-	-	-
Pick up object (6)	98.52	-	-	-	98.03	-	-	-	**98.52**	-	-	-
Hopping (7)	100	-	-	-	100	-	-	-	**100**	-	-	-
Average	89.13	92.83	98.78	99.23	95.17	95.13	98.99	99.23	**98.50**	**95.47**	**99.06**	**99.23**

Table 1. Ablation of model performance between the original approach, our improved baseline using the state probability of the sample, and our model further introducing angular information. Best results are marked in bold or underlined, with the results for the best overall model marked in bold.

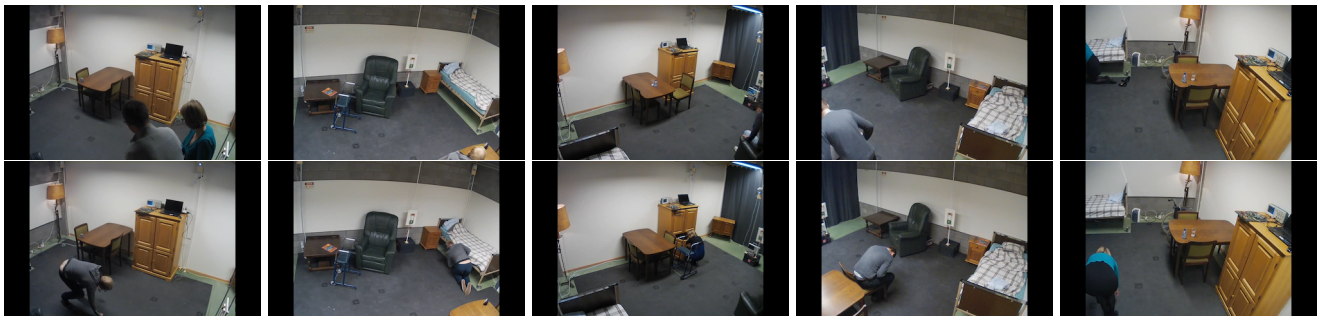

Figure 6. False positive prediction examples on the HQFSD dataset, displaying the two main failure modes of our model as seen from all five camera angles. **Top:** Clipping of the individuals is one of the main sources of error. **Bottom:** Extreme body poses such as exaggerated crouching and unusual sitting and crawling poses is also prevalent in the fall sequences of the dataset.

	Accuracy	Precision	Recall	FPS
Y. Ke *et al.* [6]	-	82.20	76.10	**71**
A. Bansal *et al.* [2]	93.20	95.12	97.50	-
J. Gonzalez *et al.* [3]	90.33	91.84	92.45	23
Q. Xu *et al.* [16]	97.25	-	97.25	-
Ours	**98.50**	**97.55**	**99.08**	2

Table 2. Comparison with other state-of-the-art Fall Detection approaches on CAUCAFall. Best results are marked in bold, second best are underlined.

	AUC	Prec.	Recall	Specif.	F1
baseline [3]	69.38	66.64	**99.53**	43.57	79.83
state probability	80.12	68.28	96.14	49.42	**79.85**
depth	**80.87**	**68.29**	96.14	**49.44**	**79.85**

Table 3. Ablation study on HQFSD. **Prec:** Precision. **Specif:** Specificity. Best results are marked in bold, second best are underlined.

dictions of both previous models (baseline and probability models). Similarly, the additional gains for this dataset are also marginal.

Both the baseline model and our own to a lesser extent suffer from low specificity. By looking at some visual examples of the failure modes, as seen in Fig. 6, we find two main modes of failure in our model. First is the clipping of individuals for a significant period of time, resulting in incorrectly generated features such as bounding boxes and depth-based distance and angular features. Second are some extreme poses, such as exaggerated crouching and unusual sitting poses. Such cases are difficult to distinguish without using more complex features like body pose information nor making use of temporal information in our model, and result on a low specificity score.

6. Conclusion

In this work we have shown the feasibility of introducing summarized depth information to Bayesian Networks, allowing us to outperform the previous state-of-the-art for Fall Detection when compared to other Anomaly Detection approaches. Not only do we surpass the previous state-of-the-art on the CAUCAFall dataset by a wide margin, but we do so with a simple model that can be trained in a matter of seconds.

This represents a significant milestone for potential commercial applications of Fall Detection: The low parametric complexity of the model means that the approach can be trained on datasets of limited size, while the the usage of Anomaly Detection makes it unnecessary to label the training data. Furthermore, given the simple, anonymous nature of the extracted features, data can be stored for training without needing to worry about privacy. This would allow for training of the model by using data from the camera after installation, better adapting to that particular scene and point of view.

7. Acknowledgements

This work is supported by Milestone Research Program at Aalborg University, industrial post grant 1045-00036B, and Innovation Fund Denmark.

References

[1] Greet Baldewijns, Glen Debard, Gert Mertes, Bart Vanrumste, and Tom Croonenborghs. Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. *Healthcare technology letters*, 3(1):6–11, 2016. 4, 5

[2] Aayushi Bansal, Rewa Sharma, and Mamta Kathuria. A vision-based approach to enhance fall detection with fine-tuned faster r-cnn. In *International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, pages 678–684. IEEE, 2023. 2, 7

[3] Jacobo Gonzalez de Frutos, Mia Sandra Nicole Siemon, and Kamal Nasrollahi. Fall detection in hospital rooms with probabilistic graphical models. In *International Conference on Image Processing Theory, Tools and Applications*. IEEE, 2024. 2, 3, 4, 6, 7

[4] José Camilo Eraso Guerrero, Elena Muñoz España, Mariela Muñoz Añasco, and Jesús Emilio Pinto Lopera. Dataset for human fall recognition in an uncontrolled environment. *Data in brief*, 45:108610, 2022. 4, 5

[5] Sardor Juraev, Akash Ghimire, Jumabek Alikhanov, Vijay Kakani, and Hakil Kim. Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance. *IEEE Access*, 10:94249–94261, 2022. 2

[6] Yaojie Ke, Yinan Yao, Zhengye Xie, Hepeng Xie, Hui Lin, and Chen Dong. Empowering intelligent home safety: Indoor family fall detection with yolov5. In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 0942–0949. IEEE, 2023. 2, 7

[7] Sheldon McCall, Shina Samuel Kolawole, Afreen Naz, Liyun Gong, Syed Waqar Ahmed, Pandey Shourya Prasad, Miao Yu, James Wingate, and Saeid Pourroostaei Ardakani. Computer vision based transfer learning-aided transformer model for fall detection and prediction. *IEEE Access*, 2024. 2

[8] Shaou-Gang Miaou, Fu-Chiau Shih, and Chia-Yuan Huang. A smart vision-based human fall detection system for telehealth applications. In *Proc. Third ISATED Int. Conf. on Telehealth*, pages 7–12, 2007. 2

[9] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020. 6

[10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[11] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 3

[12] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[13] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Fall detection from human shape and motion history using video surveillance. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, volume 2, pages 875–880. IEEE, 2007. 2

[14] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. 3d head tracking for fall detection using a single calibrated camera. *Image and Vision Computing*, 31(3):246–254, 2013. 2

[15] WHO. Falls. https://www.who.int/news-room/fact-sheets/detail/falls. Accessed: 2024-11-13. 1

[16] Qingzhen Xu, Guangyi Huang, Mengjing Yu, and Yanliang Guo. Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications*, 540:123205, 2020. 2, 7

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3

[18] Miao Yu, Syed Mohsen Naqvi, and Jonathon Chambers. Fall detection in the elderly by head tracking. In *IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 357–360. IEEE, 2009. 2