MỘT MÔ HÌNH HỌC SÂU CHO BÀI TOÁN PHÁT HIỆN NGƯỜI BỊ NGÃ

Phùng Thị Thu Trang[1*], Ma Thị Hồng Thu[2]
[1]Trường Ngoại ngữ - ĐH Thái Nguyên,
[2]Trường Đại học Tân Trào

TÓM TẮT

Té ngã là một trong những vấn đề nghiêm trọng đối với con người, chiếm tỷ lệ tử vong lên đến 25%, tỷ lệ này càng cao hơn đối với những người cao tuổi. Nhận dạng người bị ngã là một trong những bài toán quan trọng trong lĩnh vực thị giác máy tính. Những năm gần đây, thị giác máy tính đã đạt được tiến bộ ấn tượng khi mà học sâu thể hiện khả năng tự động học. Đã có nhiều mô hình học sâu dựa trên mạng nơ ron tích chập 3D (CNN) đã được đề xuất để giải quyết vấn đề này. Trong bài báo này, chúng tôi đề xuất một mô hình (2+1)D ResNet-18 giải quyết bài toán nhận dạng người bị ngã. Kết quả thử nghiệm cho thấy, (2+1)D ResNet-18 cho độ chính xác tốt hơn 0,87% trên bộ dữ liệu FDD và 1,13% trên bộ dữ liệu URFD so với các phương pháp được đề xuất gần đây.

Từ khóa: *Học sâu; mạng CNN; phát hiện người bị té ngã; mạng nơron; (2+1)D ResNet*

Ngày nhận bài: 05/8/2020; Ngày hoàn thiện: 13/11/2020; Ngày đăng: 27/11/2020

A DEEP LEARNING MODEL FOR FALLING DETECTION

Phung Thi Thu Trang[1*], Ma Thi Hong Thu[2]
[1]TNU – School of Foreign Languages,
[2]Tan Trao University

ABSTRACT

Falling is one of the most serious problems for humans, accounting for up to 25% of death rates, which is even higher for the elderly. Falling detection is one of the most important problems in computer vision. In recent years, computer vision has made impressive progress when deep learning demonstrates the ability to automatically learn. There have been many deep learning models based on 3D convolutional neural network (CNN) that have been proposed to solve this problem. In this paper, we propose a model which is called (2+1)D ResNet-18 to solve the falling detection task. The experimental results show that (2+1)D ResNet-18 gives 0.87% better accuracy on the FDD dataset and 1.13% on the URFD dataset than the recently proposed methods.

Keywords: *Deep learning; convolutional neural networks; falling detection; neural networks; (2+1)D ResNet*

Received: 05/8/2020; Revised: 13/11/2020; Published: 27/11/2020

* Corresponding author. *Email: phungthutrang.sfl@tnu.edu.vn*

1. Giới thiệu

Học máy, đặc biệt là học sâu, đã đạt được những thành tựu to lớn trong nhiều lĩnh vực gần đây. Mạng nơ ron hồi quy (RNN) và Mạng RNN cải tiến Long Short – Term Memory (LSTM) với ý tưởng rằng chúng có thể kết nối các thông tin trước đó với thông tin hiện tại, đã được áp dụng để giải quyết nhiều vấn đề trong nhận dạng giọng nói và xử lý ngôn ngữ tự nhiên (NLP) một cách hiệu quả. Cùng với sự phát triển của NLP, xử lý hình ảnh và thị giác máy tính cũng có những bước đột phá. Các mô hình được xây dựng dựa trên mạng nơ ron tích chập (CNN) đạt được nhiều thành tựu lớn. Ví dụ: Alex và các cộng sự [1] đã xây dựng một mạng gọi là AlexNet, mạng này đã chiến thắng trong cuộc thi phân loại hình ảnh (ImageNet) năm 2012. Trong các năm tiếp theo, rất nhiều mô hình dựa trên mạng tích chập đã được đề xuất chẳng hạn như ZFNet [2] năm 2013, GoogleNet [3] năm 2014, VGGNet [4] năm 2014, ResNet [5] năm 2015. Ngoài phân loại hình ảnh, mạng tích chập thường được áp dụng cho nhiều bài toán về hình ảnh như phát hiện đa đối tượng, chú thích hình ảnh, phân đoạn hình ảnh, v.v.

Nhận dạng hoạt động người không những là chủ đề nghiên cứu quan trọng trong tính toán nhận biết ngữ cảnh mà còn là chủ đề đối với rất nhiều lĩnh vực khác. Ngã là một vấn đề nghiêm trọng ở người cao tuổi rất thường gặp, gây tàn phế và thậm chí gây tử vong, là nguyên nhân đứng thứ 5 gây tử vong ở người cao tuổi. Ngã là một yếu tố gây tử vong, thống kê ở bệnh viện có tới 25% các trường hợp nhập viện do ngã bị tử vong, trong khi chỉ có 6% tử vong do các nguyên nhân khác. Bài toán phát hiện người bị té ngã là một trong những bài toán phổ biến trong lĩnh vực nhận dạng hoạt động của con người, thu hút được nhiều sự chú ý của các nhà khoa học. Đây là một bài toán quan trọng và có ý nghĩa hết sức to lớn đối với vấn đề bảo vệ sức khỏe của con người. Nhiệm vụ đặt ra đối với bài toán này là cần đưa ra dự đoán một cách chính xác và trong thời gian thực khi gặp trường hợp người bị ngã để giảm thiểu thời gian người ngã nằm trên sàn từ sau thời điểm ngã đến khi được người chăm sóc phát hiện.

Trong bài báo này, chúng tôi đề xuất mô hình (2+1)D ResNet-18 dựa trên kiến trúc 3D ResNet từ [6] để giải quyết bài toán phát hiện người bị té ngã. Kết quả thử nghiệm cho thấy, mô hình của chúng tôi cho độ chính xác hơn 0,87% trên bộ dữ liệu FDD và 1,13% trên bộ dữ liệu URFD so với các phương pháp được đề xuất gần đây trong [7] và [8].

Bài viết được chia thành 5 phần. Sau phần giới thiệu, phần 2 trình bày một số nghiên cứu gần đây, phần 3 mô tả kiến trúc mạng (2+1)D ResNet-18, phần 4 trình bày các thử nghiệm trên hai bộ dữ liệu FDD và bộ dữ liệu URFD cũng như thảo luận về kết quả. Phần 5 khép lại với kết luận và tài liệu tham khảo.

2. Một số nghiên cứu gần đây

Hiện nay, có hai cách tiếp cận phổ biến để giải quyết bài toán nhận dạng hoạt động, bao gồm: nhận dạng hoạt động dựa trên thị giác máy tính và nhận dạng hoạt động dựa trên cảm biến. Đối với phương pháp nhận dạng hoạt động dựa trên cảm biến đòi hỏi người sử dụng phải luôn luôn mang các thiết bị cảm biến theo bên người, điều này đôi khi gây vướng víu và phiền toái đối với người sử dụng hoặc có nhiều người đôi khi còn quên không mang theo các thiết bị này bên mình.

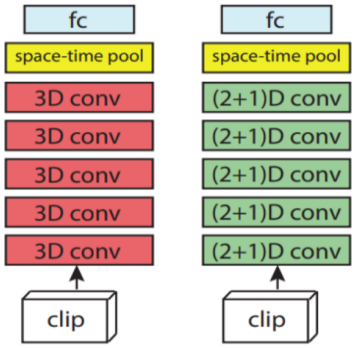
Các phương pháp nhận dạng hoạt động dựa trên thị giác máy tính thì tập trung vào việc theo dõi các dữ liệu video thu được từ camera, sau đó phân tích và đưa ra kết luận về hành động (trong bài báo này là phát hiện té ngã). Đa số các công bố theo cách tiếp cận này đều dựa trên học có giám sát. Nhiều hệ thống đều được xây dựng bằng cách trích chọn những đặc trưng từ các khung hình của video, sau đó áp dụng các kỹ thuật học máy để phân lớp. Ví dụ, Charfi cùng các cộng sự [9] đã trích xuất 14 đặc trưng từ hình ảnh dựa

trên đạo hàm bậc nhất và bậc hai, biến đổi Fourier và biến đổi Wavelet, sau đó nhóm tác giả sử dụng SVM để phân lớp các hình ảnh này. Zerrouki cùng các cộng sự đã xây dựng hệ thống nhận dạng té ngã bằng cách tính diện tích vùng cơ thể và góc của cơ thể, sau đó các đặc trưng này được đưa vào hệ thống phân loại khác nhau [10], SVM là phương pháp phân loại cho kết quả tốt nhất thời điểm đó. Vào năm 2017, cũng với nhóm tác giả này, họ đã mở rộng nghiên cứu bằng cách áp dụng thêm các hệ số Curvelet và sử dụng mô hình Markov ẩn (HMM) để mô hình hóa các tư thế cơ thể khác nhau [11].

Trong những năm gần đây, học sâu (deep learning) đã đạt được nhiều thành tựu to lớn trong lĩnh vực trí tuệ nhân tạo, đặc biệt là thị giác máy tính. Cùng với sự bùng nổ về sự phát triển phần cứng, các framework hỗ trợ, đã có rất nhiều mô hình học sâu được xây dựng để giải quyết bài toán phát hiện người té ngã. Chẳng hạn như Adrián cùng các cộng sự đã xây dựng, đề xuất mô hình sử dụng kiến trúc mạng VGG-16 để trích chọn đặc trưng và phân lớp [7]. Năm 2019, Sarah đã mở rộng phương pháp bằng cách sử dụng các hình ảnh đầu vào khác nhau cho mô hình VGG-16 [8]. Trong bài báo đó, họ đã sử dụng ba loại hình ảnh: ảnh RGB, ảnh optical flow (áp dụng optical flow để trích xuất ra hình ảnh chuyển động giữa các khung hình) và ảnh khung xương (áp dụng pose estimate để trích xuất ra hình ảnh khung xương của con người). Thêm vào đó, họ đã kết hợp sử dụng các hình ảnh này với nhau và kết quả cho thấy, với đầu vào gồm cả 3 loại hình ảnh trên thì mô hình của họ đạt kết quả cao nhất.

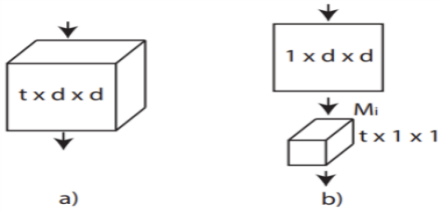
3. Đề xuất mô hình

Năm 2015, Kaiming He cùng các cộng sự đã đề xuất một mô hình mang tên ResNet [5]. Với kỹ thuật skip connection trong [5], ResNet đã có thể tránh được vấn đề vanishing gradient mà không làm giảm hiệu suất mạng. Điều đó giúp các lớp sâu ít nhất không tệ hơn các lớp nông. Hơn nữa, với kiến trúc này, các lớp trên nhận được nhiều thông tin trực tiếp hơn từ các lớp thấp hơn nên nó sẽ điều chỉnh trọng lượng hiệu quả hơn. Sau ResNet, một loạt các biến thể của phương pháp này đã được giới thiệu. Các thí nghiệm cho thấy những kiến trúc này có thể được đào tạo với các mô hình CNN với độ sâu lên tới hàng ngàn lớp. ResNet đã nhanh chóng trở thành kiến trúc phổ biến nhất trong lĩnh vực học sâu và thị giác máy tính.



Hình 1. *Sự khác nhau giữa 3D CNN và (2+1)D CNN*

Trong [6], các tác giả đã đề xuất mô hình 3D ResNet để giải quyết bài toán phân loại hành động. Tuy nhiên, các mô hình ở trong [6] đều rất sâu và phức tạp, đồng thời chúng được huấn luyện trên các bộ dữ liệu lớn. Do đó, các mô hình 3D Resnet này không phù hợp với bài toán nhận dạng té ngã mà trong bài báo này đang xét đến. Để giảm độ phức tạp của mô hình 3D CNN, trong [12], các tác giả đã trình bày kỹ thuật kết hợp 3D CNN với 2D CNN và sử dụng (2+1)D CNN. Qua thử nghiệm cho thấy, việc sử dụng (2+1)D CNN cho kết quả tốt hơn hẳn so với chỉ sử dụng 3D CNN và kết hợp 3D CNN với 2D CNN. Hình 1 mô tả sự khác nhau giữa hai kiến trúc 3D CNN và (2+1)D CNN. Trong đó, mỗi khối 3D conv đều được thay thế bằng các khối (2+1)D conv.



Hình 2. *So sánh khối 3D convolution thông thường với khối (2+1)D convolution*

Hình 2 mô tả sự khác nhau giữa hai khối 3D conv và (2+1)D conv. Trong đó, với khối 3D conv thì kích thước hạt nhân thường được sử dụng sẽ có dạng t ×d × d còn trong khối (2+1)D conv, phép tích chập 3D này sẽ được tách thành hai phép tích chập nhỏ hơn với phép tích chập thứ nhất có kích thước hạt nhân là 1 × d × d và phép tích chập thứ hai sẽ có kích thước hạt nhân là t × 1 × 1. Với (2+1)D conv, thì số lượng tham số và chi phí tính toán được giảm đi đáng kể so với khối 3D conv thông thường. Trong [12], các tác giả đã chứng minh rằng (2+1)D conv hoạt động tốt hơn 3D conv.

Toàn bộ kiến trúc mô hình (2+1)D ResNet-18 được trình bày như trong bảng 1. Trong đó, Conv1, Conv2_x, Conv3_x, Conv4_x là các tầng tích chập với x thể hiện rằng tầng đó được lặp lại nhiều lần và có sử dụng kỹ thuật skip connection. Đầu ra của tất cả các tầng tích chập mặc định đều được đưa vào tầng Batch Normalization và ReLU. Ở cột tham số, 7 × 7 × 7; 64 thể hiện rằng tầng tích chập đó có kích thước hạt nhân là 7 × 7 × 7 và số lượng bộ lọc là 64. Với khối MaxPool, *k* đại diện cho kích thước hạt nhân và *s* là bước nhảy. Khối FC đại diện cho tầng Fully Connected, trong tầng này chúng tôi sử dụng hàm sigmoid để đưa ra dự đoán phân lớp cho video clip đầu vào.

4. Thử nghiệm và các kết quả

4.1. Các bộ dữ liệu và thiết lập

Trong bài báo này, chúng tôi sử dụng hai bộ cơ sở dữ liệu là FDD và URFD để tiến hành thử nghiệm và so sánh kết quả mô hình chúng tôi đã đề xuất với các công bố gần đây.

Bộ dữ liệu FDD được xây dựng năm 2013. Bộ dữ liệu này bao gồm các video được quay lại ở hai địa điểm là phòng cà phê và phòng ở nhà. Tất cả các video trong bộ dữ liệu được quay lại bởi một camera duy nhất và được thiết lập có độ phân giải hình ảnh là 320 × 240 pixel và tốc độ khung hình là 25 fps. Các diễn viên trong mỗi video đều thực hiện các hoạt động bình thường ở nhà và ngã tại mỗi thời điểm khác nhau, các hoạt động này đều được thực hiện một cách ngẫu nhiên. Địa chỉ website của bộ dữ liệu FDD là http://le2i.cnrs.fr/fall-detection-dataset?lang=fr.

Bộ dữ liệu URFD được Bogdan Kwolek cùng các cộng sự xây dựng năm 2014 [13] nhằm mục đích nhận dạng người bị ngã thông qua các loại thiết bị khác nhau như camera, gia tốc kế, Microsoft Kinect (trong bài báo này, chúng tôi chỉ sử dụng các video được quay từ camera trong bộ dữ liệu mà không sử dụng thông tin từ các thiết bị khác). Bộ dữ liệu bao gồm 70 videos với 30 videos chứa các hành động ngã khác nhau và 40 videos còn lại chứa những hoạt động bình thường được diễn ra hàng ngày, chẳng hạn như: ngồi, đi lại, cúi người, v.v. Địa chỉ tải xuống bộ dữ liệu URFD tại http://fenix.univ.rzeszow.pl/mkepski/ds/uf.html.

Bảng 1. *Kiến trúc mô hình (2+1)D Resnet-18*

Tên khối	Tham số	Lặp	Kích thước đầu ra
Tầng Input			(16,224,224,3)
Conv 1	7 × 7 × 7, 64	1	(16,112,112,64)
MaxPool	*k*=(3,3,3) *s*=(1,2,2)	1	(16,56,56,64)
Conv2_x	1 × 3 × 3, 128 3 × 1 × 1, 128	2	(8,28,28,128)
Conv3_x	1 × 3 × 3, 256 3 × 1 × 1, 256	2	(4,14,14,256)
Conv4_x	1 × 3 × 3. 512 3 × 1 × 1. 512	2	(2,7,7,512)
Global Spatial Pool		1	(2,512)
Flatten		1	(1024)
FC		1	(1)

Mô hình của chúng tôi được đào tạo từ đầu với hàm tối ưu hóa là Adam. Các video huấn luyện được chia thành nhiều clip có độ dài 16 khung hình và mỗi khung hình có kích thước là $224 \times 224 \times 3$. Kích thước mỗi batch là 16 clips. Tỷ lệ học tập được khởi tạo là 0,001 và giảm đi 10 lần nếu trong 10 epoch liên tiếp mà mô hình không cải thiện được độ chính xác trên tập kiểm thử. Tất cả các mô hình đều được huấn luyện với 100 epochs và độ chính xác được tính trên tập ảnh thử nghiệm. Để đánh giá chính xác hiệu suất của mô hình, chúng tôi sử dụng phương pháp five-fold cross validation và so sánh kết quả của mô hình với các phương pháp đã được đề xuất gần đây trong [7] và [8] về cả độ chính xác, lượng tham số sử dụng cũng như số phép toán thực hiện.

4.2. Phương pháp đánh giá

Từ quan điểm của việc học có giám sát, phát hiện té ngã có thể được coi là một bài toán phân loại nhị phân mà trên đó một bộ phân loại phải quyết định xem chuỗi các khung video đầu vào có nhãn là ngã hay không. Phương pháp phổ biến nhất để đánh giá hiệu suất của bộ phân loại như vậy là recall (hoặc sensitivity), specificity và độ chính xác (accuracy). Ba phương pháp đánh giá chúng tôi sử dụng được xác định như sau:

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Trong đó:

- TP - true positives: số lượng video clip được gán nhãn là ngã và dự đoán của mô hình cũng là ngã.

- FP - false positives: số lượng video clip được gán nhãn là không phải sự kiện ngã trong khi dự đoán của mô hình là ngã.

- TN - true negatives: số lượng video clip được gán nhãn là không phải sự kiện ngã và dự đoán của mô hình cũng là không phải sự kiện ngã.

- FN - false negatives: số lượng video clip được gán nhãn là ngã trong khi dự đoán của mô hình là không phải sự kiện ngã.

4.3. Kết quả và so sánh

Trong bảng 2, chúng ta có thể thấy, mô hình (2+1)D ResNet-18 cho kết quả tốt nhất về độ đo Specificity và Accuracy. Cụ thể, (2+1)D ResNet-18 hơn 3-streams trong [8] 1,28% về mặt Specificity và hơn 0,87% về mặt Accuracy. Về phép đo Recall, mô hình của chúng tôi kém hơn 0,8% so với Pose Estimation trong [8].

Đối với bộ dữ liệu URFD, các kết quả được trình bày như trong bảng 3. Có thể thấy, (2+1)D ResNet-18 hơn 1,29%, 0% và 1,13% so với phương pháp tốt nhất hiện có trong [7] và [8], tương ứng trên 3 phép đo Specificity, Recall và Accuracy.

Bảng 2. *So sánh (2+1)D Resnet-18 với các nghiên cứu được công bố gần đây về độ chính xác trên bộ dữ liệu FDD*

Mô hình	Kiến trúc	Specificity	Recall	Accuracy
VGG + optical flow [7]	VGG-16	97,0	99,0	97,0
RGB [8]	VGG-16	79,02	100,0	80,52
Optical Flow [8]	VGG-16	96,17	99,9	96,43
Pose Estimation [8]	VGG-16	60,15	100,0	63,01
3-streams (OF+PE+RGB) [8]	VGG-16	98,32	99,9	98,43
(2+1)D Resnet-18	**Resnet**	**99,6**	**99,2**	**99,3**

Bảng 3. *So sánh (2+1)D Resnet-18 với các nghiên cứu được công bố gần đây về độ chính xác trên bộ dữ liệu UCFD*

Mô hình	Kiến trúc	Specificity	Recall	Accuracy
VGG + optical flow [7]	VGG-16	92,0	**100,0**	95,0
RGB [8]	VGG-16	96,61	**100,0**	96,99
Optical Flow [8]	VGG-16	96,34	**100,0**	96,75
Pose Estimation [8]	VGG-16	93,09	94,41	93,24
3-streams (OF+PE+RGB) [8]	VGG-16	98,61	**100,0**	98,77
(2+1)D Resnet-18	**Resnet**	**99,9**	**100,0**	**99,9**

5. Kết luận

Trong bài báo này, chúng tôi đã đề xuất một mô hình học sâu mang tên (2+1)D ResNet-18 dựa trên kiến trúc của ResNet để nhận dạng người bị té ngã từ dữ liệu video. Kết quả thử nghiệm cho thấy, mô hình đạt hiệu suất tốt hơn các mô hình đã được công bố gần đây.

Trong tương lai gần, chúng tôi đang có kế hoạch cải thiện độ chính xác của mô hình, Mặt khác, chúng tôi sẽ áp dụng mô hình trên cho các bài toán khác trong lĩnh vực thị giác máy tính và xử lý hình video.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1106-1114.

[2]. M. D. Zeiler, and R. Fergus, "Visualizing and Understanding Convolutional Networks," *European Conference on Computer Vision*, Springer, 2014, pp. 818-833.

[3]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[4]. K. Simonyan, and A. Zisserman, "Very deep Convolutional Networks for large-scale Image Recognition," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 1-14.

[5]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[6]. K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3d CNNs retrace the history of 2d CNNs and Imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546-6555.

[7]. A. Nú˜nez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based Fall Detection with Convolutional Neural Networks," *Wireless communications and mobile computing*, vol. 2017, pp. 1-16, 2017.

[8]. S. A. Cameiro, G. P. da Silva, G. V. Leite, R. Moreno, S. J. F. Guimarães, and H. Pedrini, "Multi-stream Deep Convolutional Network using High-level Features applied to Fall Detection in Video Sequences," in *International Conference on Systems, Signals and Image Processing*, 2019, pp. 293-298.

[9]. I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki, "Definition and Performance Evaluation of a robust SVM based Fall Detection Solution," in *8th International Conference on Signal Image Technology and Internet Based Systems*, 2012, pp. 218-224.

[10]. N. Zerrouki, F. Harrou, A. Houacine, and Y. Sun, "Fall Detection using Supervised Machine Learning Algorithms: A comparative study," in *8th International Conference on Modelling, Identification and Control (ICMIC)*, IEEE, 2016, pp. 665-670.

[11]. N. Zerrouki, and A. Houacine, "Combined Curvelets and Hidden Markov Models for Human Fall Detection," *Multimedia Tools and Applications*, vol. 77, no. 5, pp. 6405-6424, 2018.

[12]. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450-6459.

[13]. B. Kwolek, and M. Kepski, "Human Fall Detection on Embedded Platform using Depth Maps and Wireless Accelerometer," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 489-501, 2014.