Ski Fall Detection from Digital Images Using Deep Learning

YULIN ZHU

Auckland University of Technology, Auckland 1010 New Zealand

WEI QI YAN

Auckland University of Technology, Auckland 1010 New Zealand

Abstract. In this paper, we explore how to take advantage of computer vision to assist ski resorts and monitor the safety of skiers on the tracks. In order to quickly detect any falls or injures, and provide first aid for injured people, we make use of archived ski videos, which are employed to explore the possibility of skiers fall detection. Throughout combinations of visual object detection with human pose detection by using deep learning methods. Our ultimate goal of this project is to provide a way for ski safety monitoring which has potential applications for physical training. Our contribution in this paper is to propose a fall detection method suitable for skiers based on visual object detection, we have obtained 0.94 mAP accuracy in preliminary tests.

CCS CONCEPTS • Skiing • Deep learning • Visual object detection • Fall detection

ACM Reference Format:

Y. Zhu, W. Yan. Ski fall detection from digital images using deep learning. In ICCCV' 22: The 5th International Conference on Control and Computer Vision (ICCCV 2022), August 19-21, 2022. Xiamen, Fujian China.

1 INTRODUCTION

Skiing is popular in the world because of its unique interest. After a long time of development, it has become a fashion sport in winter. However, we should also pay special attention to the dangers in enjoying the funny brought by skiing. With the increasing populations of skiing, skiing is no longer just a recreational activity, but also has gradually developed into a professional game in the Winter Olympic Games. Skiing is mainly grouped into outdoor alpine skiing and indoor skiing. How to ensure skiing safety and reduce skiing injury is the focus of the safety measures of ski resort. In skiing, the number of people injuries has always been increased every year [1，2], which not only hinders the development of skiing skills, but also brings huge losses, even the lives of the injured people outdoors like high mountains. By quickly detecting falls or other unusual events, faster injury rescue and the first aid could be served timely.

Inspired by research outcomes that take use of weather forecasts to predict ski injuries, visual object detection is employed to detect and mark a skier's fall from a sequence of digital images or motion pictures. It is expected that digital videos from surveillance with computer vision could reduce the work pressure of security and rescuer, and shorten the rescue time.

In this paper, we take use of a sequence of video frames to detect fall actions in spatiotemporal domain. Our contribution is to detect incidental falls of outdoor skiers. Unlike indoor detection, outdoor ski scenes often have a larger white area and people are often harder to be detected. As the novelty of our research project, we combine computational methods of human skeleton detection and object detection together, which improves the accuracy from 0.76 mAP to 0.94 mAP. Due to the use of less training samples, we believe that this result can still be further improved.

The remaining part of this paper is organized as follows. Literature will be reviewed in Section 2. Our methodology is depicted in Section 3. Our result analysis is presented in Section 4. Our conclusion and future work are addressed in Section 5.

2 LITERATURE REVIEW

A systematic review on fall detection from digital images is classified into 3-fold, including fusion-based method, vision sensor systems, and physical devices [3]. Most algorithms [3] for fall detection are important in the context of real world. By limiting to accelerometer-based fall detection algorithms, the implementation and comparison have been proffered and analyzed recently [4]. These algorithms are working with various parameters and thresholds together. A study provided guidance on comparing accuracy of threshold-based detection approach with machine learning-based approach, including logistic regression, naïve Bayes, nearest neighbors, decision tree, and support vector machine [5]. By feeding the dataset into two classifiers, it is reported that the best machine-learning algorithm gives the highest accuracy than the high-performance threshold-based algorithms. In contrast, the methods measuring acceleration performs better to detect falls as they have lower error rates and higher detection rates [6].

It is crucial to detect falls automatically for both indoor and outdoor scenes. There have been numerous studies for fall detection using various methods, including wearable devices and visual-based systems. More research work has been conducted indoor fall detection from surveillance videos [7–15], in this paper, our focus is mainly on detection of outdoor activities [16–18].

A novel approach for outdoor fall detection aims to use an ellipse method to model the shape of the person from color images, estimate the person's activity by using a state vector and an observation vector from a single camera [18]. This approach overcomes the difficulties of illumination under an outdoor environment. However, it could not resolve the problems of detecting multiple people and multiple falls that simultaneously occur in the same scene.

It is reported that occlusions, computational complexity, and false alarm rate are considerable barriers for vision-based fall detection [7–15, 17–19]. One approach of fall detection is to measure velocity and inactivity based on the 3D bounding box of deformation [8]. Mastorakis et al. [8] proposed highly completed privacy prevention, the method on fall-like actions including lying down and crouching are wrongly detected. An ellipse method was later supplied to distinguish between "fall-down" and "fall-like" activities under a camera tilted against the horizontal. In Chen's approach [11], skeleton extraction and ellipse fitting are combined for various human shapes. Nevertheless, the challenge is to automatically model human postures with much reliable computations.

Instead of using only bounding box or conventional ellipse methods, Fourier Temporal Features are proposed as a shape matching method to determine the silhouette of individuals in an image sequence [10, 30]. Another three-point representation based on human shape variation reaches high successful detection rate 90.50% and overcomes a false alarm rate 6.70% with the minimum computational complexity [7].

Over the past few years, machine learning methods, particularly associated with neural networks, which have played an essential role in designing pattern recognition methods [20]. The availability of deep learning is crucial in the recent success of pattern recognition such as continuous speech recognition and handwriting recognition. A better pattern recognition method was created with automated model training with less reliance on manually-designed deep nets. Lecun et al. [20] proposed a graph converter network (GTN) for global training of multimodule systems by using a gradient-based approach to minimize the overall costs. Two kinds of online handwriting recognition methods are introduced and the relevant experiments are carries out. The experiments show the advantage of global training and scalability of graph transformation networks. At the same time, a graph transformation network for reading bank checks is also introduced. It makes use of a character recognizer based on convolutional neural network (CNN) to provide accurate business and personal inspection records.

Various perceptual and machine learning methods have been proposed to monitor human activities automatically [21]. Visual information to identify human activities (computer vision method) is one of the most usual methods. However, these methods have limitations, such as being affected by indoor lighting conditions and obstacles. Jung & Chi pointed out a human activity classification model based on voice recognition and investigated its performance and limitations. In the project, 10 activity classes that are carried out in the room were selected, the corresponding images were collected for preprocessing, two-bit feature vectors are converted for model training and evaluations. The accuracy rate of the established neural network model is 87.20%. The problems encountered are based on the differences in data collection, such as the number of samples, the number of classes, and the objects to be detected and recognized. These reasons will affect the training and evaluation process of the classification model. From the experimental results, we see that it is possible to use deep learning models to classify various signal patterns of human activities.

Various CNN nets [22] including multilayer architecture with the low-level features on body limp parts are applied to overcome the barriers of posture estimation. A stacked hourglass network has been employed to determine the accurate pixel location of key points provided in an RGB image [23]. One of the issues is how to annotate visual objects with background, another problem is the possible exclusion of non-visible body joints.

You Only Look Once (YOLO) presents a unified real-time system by spatially separating bounding boxes [24]. Additionally, YOLO shows a higher computation speed compared to other highly ranked detection frameworks in feature extraction. Consequently, OpenPose successfully detects 2D pose in real-time with multiple people [25]. The combination of different methods is a critical role in fall detection. Therefore, our approach is to utilize the combination of YOLO and OpenPose to control the efficiency and reliability based on multiperson fall detection in video surveillance.

3 METHODOLOGY

CNNs are a class of convolution operations associated with Feedforward Neural Networks (FFNN), which are representative algorithms of deep learning [25]. A convolution unit usually contains a convolution layer, an activation function, and a pooling layer show as Fig. 1. CNNs have been used to extract features from digital images. The process of YOLO convolutions is to use a 3x3 or 7x7 convolution kernel to slide through the windows on the image successively, and extract the visual features of the images after the operations.

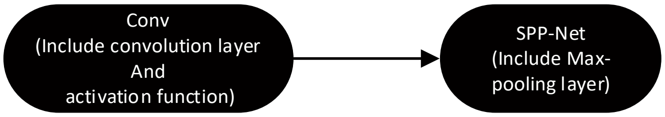


Figure 1: Part of CNN structure in YOLOv5

$$ReLU\ f(x) = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases} \tag{1}$$

For visual object detection, it usually includes object locating and classification [24]. Classification is a way to carry out visual feature matching through convolutions, and finally obtain an assigned label. Locating is the process to find the location of the object on the image. So there are two types of methods for visual object detection: one-stage detection methods and two-stage detection methods. One-stage detection methods conduct classification and localization information directly from the backbone network, which is relatively faster. Compared with one-stage detection methods, two-stage detection methods take one more steps of RoI using RPN (i.e., Region Proposal Network) [27]. The purpose is to obtain the object position accurately and improve the detection accuracy, but will slow down the detection speed (i.e., frames per second).

Because surveillance videos are dynamic, we expect to detect the people and easily judge whether they fall through digital image processing under the given speed. Therefore, we chose to highlight human body skeleton so that the one-stage detection algorithms are able to find the object and classify it easily and accurately.

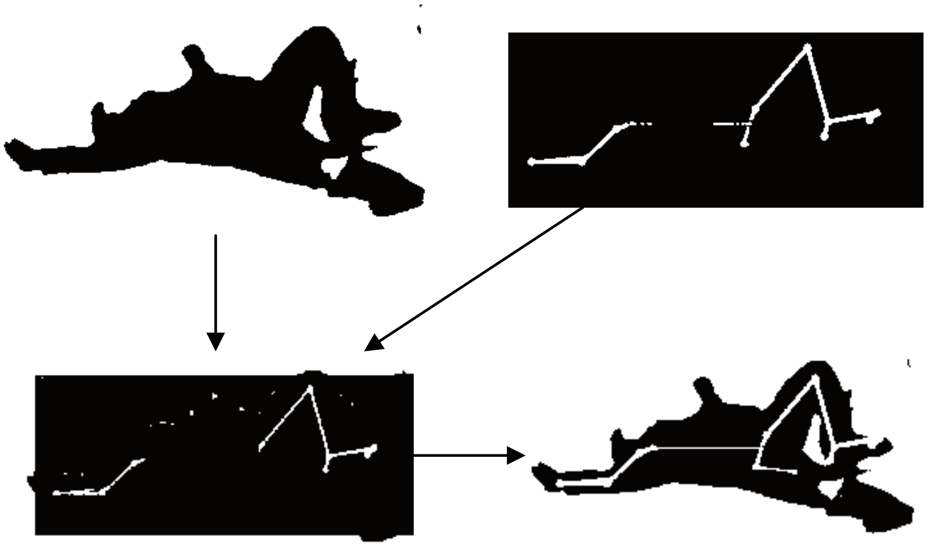
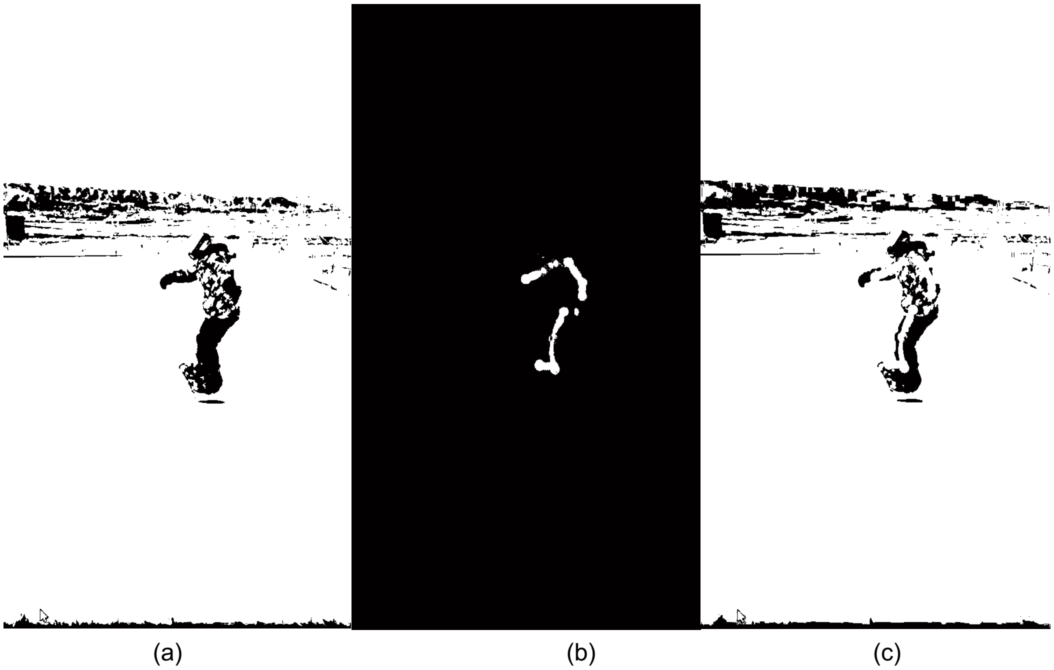


Figure 2: Mark skeleton

In order to investigate the influence of the input images to the model, three schemes were selected. For skiing scenes, it is relatively difficult to quickly identify the skier in a video, as the clothing and background are easily blended together. It is assumed that the detection accuracy of object detection is improved by applying human skeleton. In this project, we have two methods for the original image as shown in Fig. 2 by using OpenPose [25]. The first is to remove the background by redrawing the skeleton on a black background, because in the ski resort, human clothes are easily mixed with background, we take use of the color marked skeleton, keep the background in black, and reduce the disturbance of the clothes color, conduct a convolution that has more obvious characteristics. The second method is to attach the color skeleton directly to the original image, without the background and clothing. The goal is to focus on the actions of the characters and separate the players from the background as shown in Fig. 3.



(a) (b) (c)

Figure 3: The three types of input images (a) Original image (b) Black image (c) Color image

The marked skeleton from the result of OpenPose as shown in Fig. 4 along with the color images are used as the input, after first 10 layers convolution, VGG-19 outputs a set of feature maps, then the network has two separated branches, one branch is to predict the confidence, key points or human joints. The other is to predict the pixels direction in the skeleton of human body.

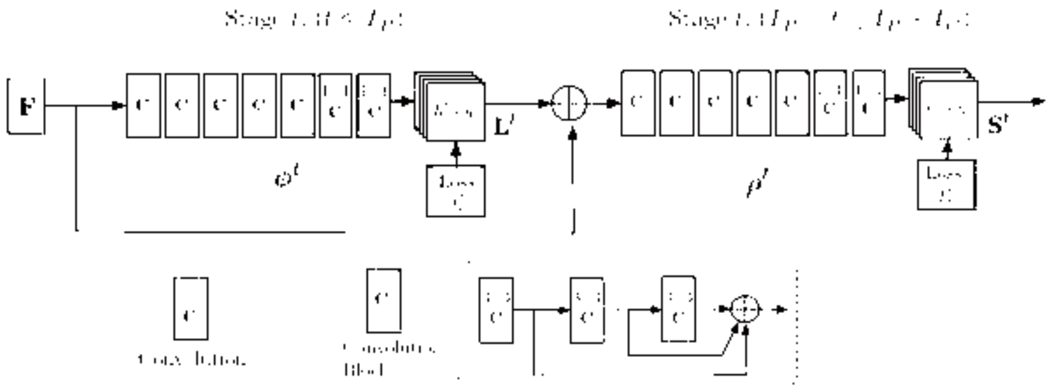
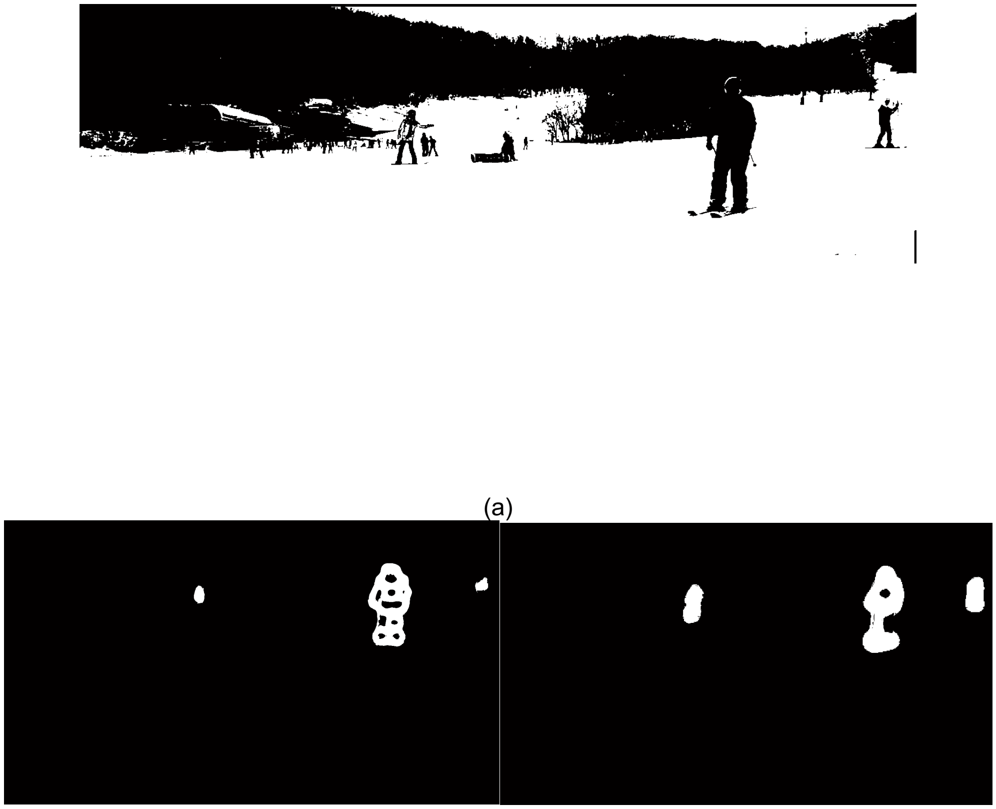


Figure 4: The network structure of OpenPose

In each subsequent stage, the input of the network will be employed for the predictions from the previous stage with original image to connect with feature map, we apply them to produce the refined prediction of human joints and connect them to our human body.



(a)

(b) (c)

Figure 5: The result of two-branch multistage CNN (a) Original image (b) Background heat map (c) Superposition of heat map

Greedy inference algorithm was harnessed to parse the confidence maps, and quickly map these joints to different individuals based on the detected joints and joint connectivity regions. The result is shown in Fig. 5

Based on the inspiration of segmentation and detection of small objects in large images, the goal of our project is to improve the accuracy of tiny object detection in the case with a small number of samples. If we are simply use of the original images for model training, overfitting or underfitting may occur. In order to avoid such issues, we firstly identify human skeleton and take use of vivid colors to mark it. The purpose is to give the same object with two labels for complementary purpose. For example, we firstly identify players, then find whether the people have made a specific action so as to improve the detection accuracy. It is also important to select an appropriate number of convolution kernels, which will determine the width of feature map. Pertaining to YOLOv5, there are four network structures with various network depths and widths [23]. As a comparison, we chose YOLOv5s and YOLOv5m which have different numbers of convolution kernels and residual units, respectively.

4 RESULT ANALYSIS

In this paper, image-based object detection will be conducted to detect skier falls. The method of comparative experiment is applied to assert whether the model having skeleton will have a higher accuracy rate. Our program was run on a laptop equipped with Intel i7 6700 CPU, Nvidia GTX960M GPU. The Operating System (OS) is based on Ubuntu 20.04 LTS, plus Anaconda 3 with Python 3.6. In order to mark body skeleton, we utilized OpenPose V1.6.0, and modified YOLOv5 to reach the result we expect. The CUDA Drive and cuDNN have been installed to accelerate the model training and visual object detection. In this paper, we take use of the videos from YouTube and selfie footages.

We are use of OpenPose to extract skeletons through the way of extracting each frame, generate three groups of images, which contain color skeletons with background. There are 282 images which were obtained for each group, including 252 frames extracted from the YouTube videos and 30 images from our own video. The training set includes 85.00% frames extracted from the YouTube videos, which contributed to 100 normal/stand samples and 124 fall samples. The validation set has 15.00% images and another 15.00% images from our footage.

In this paper, we consider three experimental scenarios. We selected ski videos from YouTube and some videos taken by myself. The videos were split into three groups: Original video, skeleton video with black background, and the video with skeleton marked on the original video. By extracting the frames of three videos and manually annotated, we obtained a total of 1,500 labeled images for our model training.

Throughout the model training by using images labelled with skeletons, we see that visual object detection is a fast method to detect whether skiers fall. The detection process is fast and the results are accurate. As the highest test results, the accuracy rate reached up to 0.94 mAP.

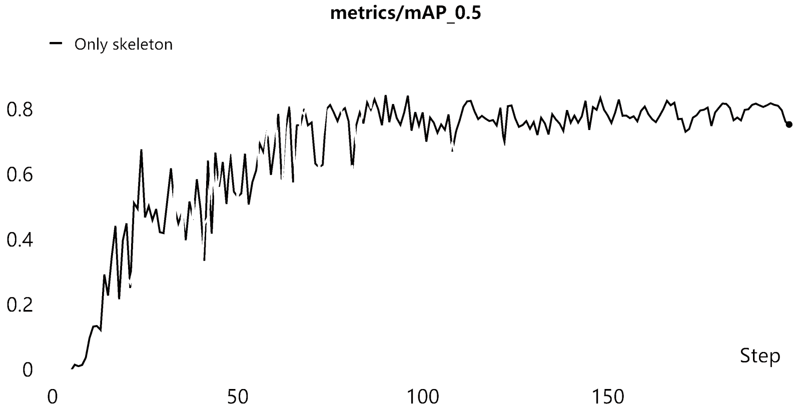


Figure 6: The mean average precision over 50.00%

We compare and analyze the training results. In the design of our experiments, we carried out two image processing methods based on the same set of images by using a black background displaying only the skeleton, another is to attach the skeleton to the original image. The images are shown in Fig. 6, the results were obtained by using different images with the same training methods. After the computations with 100 epochs, the curve becomes relatively converged.

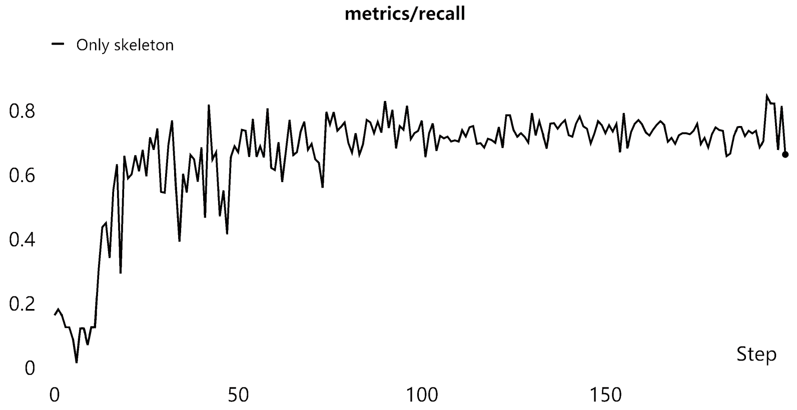
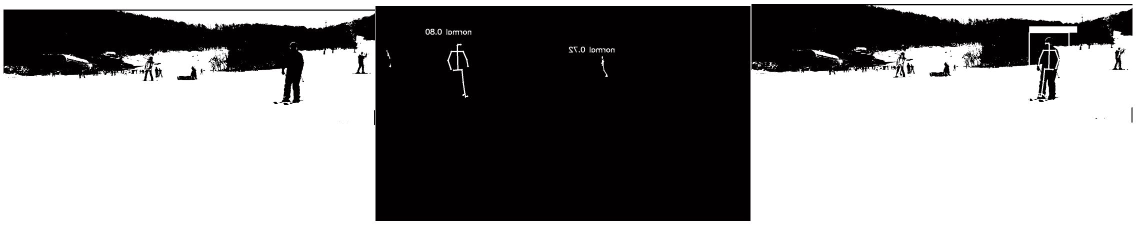


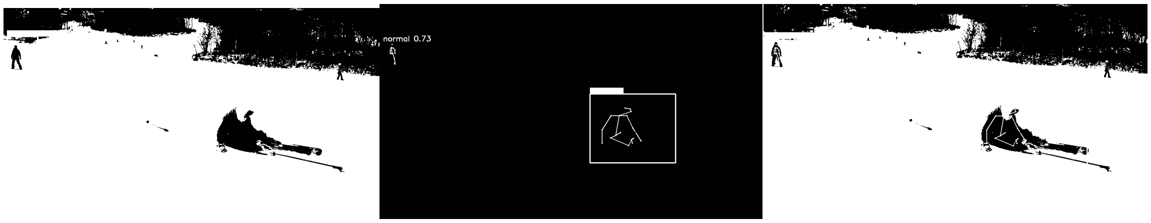
Figure 7: The recall rates during model training

From the results, we see the method of displaying the skeleton to the original image achieves the best results, the model is able to extract more features. By comparing the recall rates, we see that in Fig. 7, marking human body in advance is beneficial to improve recall rates, the skeleton is able to provide visual features in the feature map compared with the same clothes.



(a) (b) (c)

Figure 8: The prediction on different models with normal samples (a) Original image (b) Black image (c) Color image



(a) (b) (c)

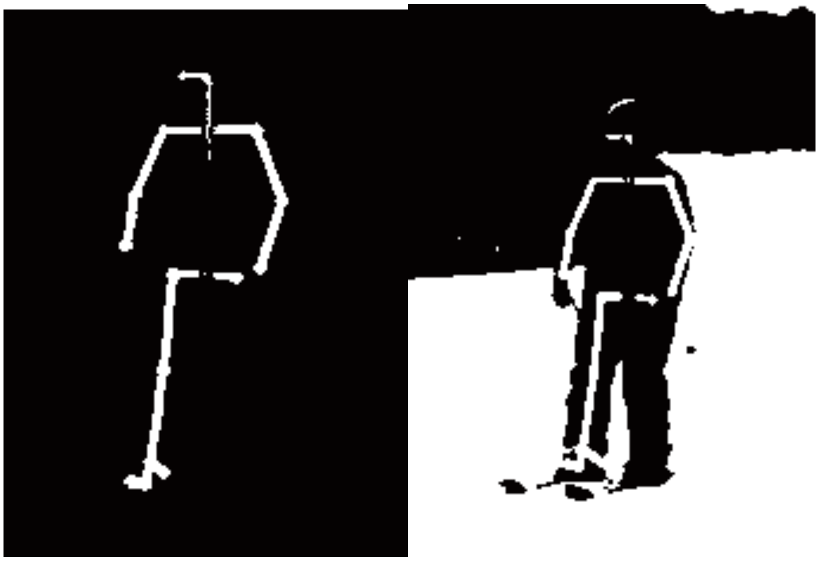
Figure 9: The predictions on various models with fall samples (a) Original image (b) Black image (c) Color image

Pertaining to the generalization ability of the three models in the case of a small number of samples, we compared two images in Fig. 8 and Fig. 9, which were not utilized in the training set. Through comparisons, we see that the model trained with original image has not prediction bounding box in the input samples, the other samples have prediction bounding box but with low confidence. Compared with other two methods, the predictions are well. The accuracy is improved by removing the background and keeping only the skeleton.

Figure 10: The comparison between black background and original image

In this paper, we find that not all skiers are marked, even if their skeletons have been identified, but no matter their postures are normal or fall, they still did not get a corresponding identification. By adjusting the threshold value, we see that they have a lower confidence, the reason may be that the object is too far, too small, overcrowding or occlusion as shown in Fig. 10.

More samples are obtained through altering image enhancement algorithms. For example, in both versions of YOLOv4 and YOLOv5, mosaicing is employed for data enhancement through randomly scaling, clipping, and rearranging.



(a) (b)

Figure 11: Marked the skeleton on the original image (a) Skeleton (b) Marked skeleton

In this paper, we are use of YOLO models for cross-validation. Through the input of original images marked with color skeleton, we test the model based on the samples that are not appeared in training set, we find that the model shows well generalization ability. We notice that this model contains two kinds of inputs: Original image and the skeleton image, it also shows that the model has the ability of feature fusion. For skiing videos or images, it’s difficult to detect human bodies. By splitting the detection process into two steps, it assists us to improve the accuracy of fall detection. In the first step, the “bottom-up” method was employed to detect all the key points of human body in the image and then map these key points to individuals. The second step is to take use of the deep learning method of visual object detection in end-to-end way, so as to find human bodies and identify posture much accurately as shown in Fig. 11.

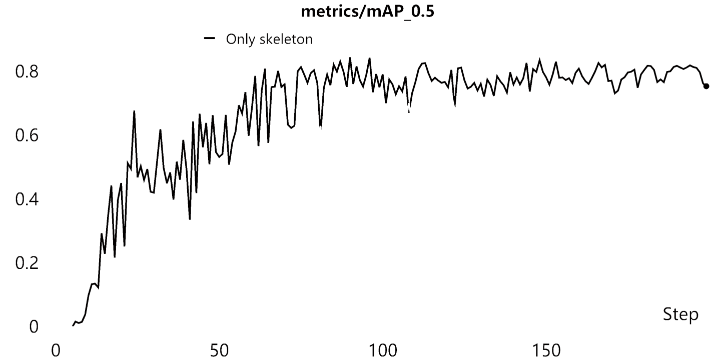


Figure 12: The comparison between original images and skeletons during training

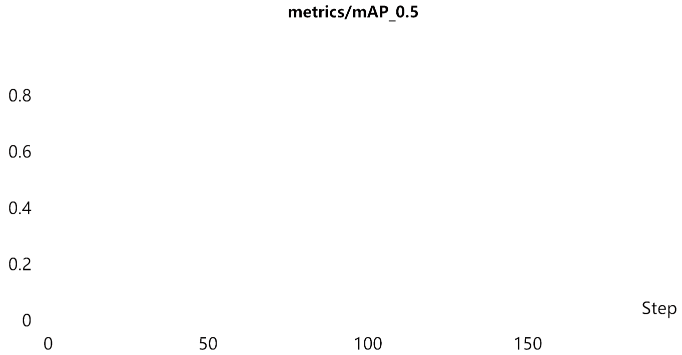


Figure 13: The comparison between original images and the images marked with skeletons during training

By analyzing the experimental result, we see that the skeleton recognition method introduced in this paper is able to greatly improve the accuracy of visual object detection. Through our comparison, we see that the best accuracy rate of object detection by using the original image is 0.76 mAP, the best accuracy rate of object detection using only the skeleton is 0.85 mAP as shown in Fig. 12, the best accuracy rate of object detection using the original image and skeleton together is 0.94 mAP as shown in Fig. 13.

5 CONCLUSION AND FUTURE WORK

In future, continuous improvement is still needed. For example, in our experiment, skiers overlap and crossover occurred when clash occurred. Such scenes can be well classified but not easy to detect. At the same time, it is necessary to take into account the density of video recorder equipment, which may lead to the detection of small objects. In this paper, we are use of object detection incorporating skeleton detection as a possible solution to skier fall detection by extracting video frames from surveillance footages. As a contribution to this project, we propose to detect human skeleton at first and mark the skeleton on the video frames as a method of image enhancement, we improve the accuracy by using visual object detection to identify fallen skiers. The object detection accuracy is only 0.76 mAP, after the original image is marked with skeleton, the accuracy is improved to 0.94 mAP. In addition, the model has good generalization ability and is able to detect visual objects. In this project, we find a new way to reduce human labor, which allows ski resorts to deploy surveillance cameras that provides high-definition videos so as to improve the coverage of monitoring and reduce the blind area of injure happenings.

Using deep learning to assist outdoor rescue [28, 29, 30] is a worthy research trial. There are a lot of injuries during skiing time, such as speeding, crashing, foreign inclusion, and so on. Based on the methods from deep learning and computer vision as well as intelligent surveillance, we are confident to provide much effective solutions [31,32,33,34].

REFERENCES

[1] Gerhard Ruedl, Martin Kopp, Renate Sommersacher, Tomas Woldrich, and Martin Burtscher. 2013. Factors associated with injuries occurred on slope intersections and in snow parks compared to on-slope injuries. Accident Analysis & Prevention, 50:1221–1225.

[2] Michael Koehle, Rob Lloyd-Smith, and Jack Taunton. 2002. Alpine ski injuries and their prevention. Sports Medicine, 32(12):785–793.

[3] Lingmei Ren and Yanjun Peng. 2019. Research of fall detection and fall prevention technologies: A systematic review. IEEE Access, 7:77702–77722.

[4] Fabio Bagala, Clemens Becker, Angelo Cappello, Lorenzo Chiari, Kamiar Aminian, Jeffrey M Hausdorff, Wiebren Zijlstra, and Jochen Klenk. 2012. Evaluation of accelerometer-based fall detection algorithms on real-world falls. PloS One, 7(5):e37062.

[5] Omar Aziz, Magnus Musngi, Edward J Park, Greg Mori, and Stephen N Robinovitch. 2017. A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials. Medical & Biological Engineering & Computing, 55(1):45–55.

[6] James T Perry, Scott Kellog, Sundar M Vaidya, Jong-Hoon Youn, Hesham Ali, and Hamid Sharif. 2009. Survey and evaluation of real-time fall detection approaches. In International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET), pages 158–164.

[7] Homa Foroughi, Baharak Shakeri Aski, and Hamidreza Pourreza. 2008. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In International Conference on Computer and Information Technology, pages 219–224.

[8] Georgios Mastorakis and Dimitrios Makris. 2014. Fall detection system using Kinect's infrared sensor. Journal of Real-Time Image Processing, 9(4):635–646.

[9] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. 2015. A survey on vision-based fall detection. In ACM International Conference on Pervasive Technologies Related to Assistive Environments, pages 1–7.

[10] Ahmad Lotfi, Suad Albawendi, Heather Powell, Kofi Appiah, and Caroline Langensiepen. 2018. Supporting independent living for older adults; employing a visual based fall detection through analysing the motion and shape of the human body. IEEE Access, 6:70272–70282.

[11] Yie-Tarng Chen, Yu-Ching Lin, and Wen-Hsien Fang. 2010. A hybrid human fall detection scheme. In IEEE International Conference on Image Processing, pages 3485–3488.

[12] Weiguo Feng, Rui Liu, and Ming Zhu. 2014. Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera. Signal, Image and Video Processing, 8(6):1129–1138.

[13] Shengke Wang, Long Chen, Zixi Zhou, Xin Sun, and Junyu Dong. 2016. Human fall detection in surveillance video based on PCANet. Multimedia Tools and Applications, 75(19):11603–11613.

[14] Kaibo Fan, Ping Wang, and Shuo Zhuang. 2019. Human fall detection using slow feature analysis. Multimedia Tools and Applications, 78(7):9101–9128.

[15] Jia Lu, Minh Nguyen, Wei Qi Yan. 2020. Human behaviour recognition using deep learning. International Conference on Image and Vision Computing New Zealand.

[16] Felix Busching, Henning Post, Matthias Gietzelt, and Lars Wolf. 2013. Fall detection on the road. In International Conference on e-Health Networking, Applications and Services (Healthcom 2013), pages 439–443.

[17] Ziqi Yu and Wei Qi Yan. 2020. Human action recognition using deep learning methods. International Conference on Image and Vision Computing New Zealand.

[18] Myeongseob Ko, Suneung Kim, Mingi Kim, and Kwangtaek Kim. 2018. A novel approach for outdoor fall detection using multidimensional features from a single camera. Applied Sciences, 8(6):984.

[19] Weidong Min, Song Zou, and Jing Li. 2019. Human fall detection using normalized shape aspect ratio. Multimedia Tools and Applications, 78(11):14331–14353.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324.

[21] Minhyuk Jung and Seokho Chi. 2020. Human activity classification based on sound recognition and residual convolutional neural network. Automation in Construction, 114:103177.

[22] Jia Lu. Deep Learning Methods for Human Behavior Recognition. PhD Thesis, Auckland University of Technology, New Zealand.

[23] Zeqi Yu. Deep Learning Methods for Human Action Recognition. Masters Thesis, Auckland University of Technology, New Zealand.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788.

[25] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1):172–186.

[26] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. 2018. Recent advances in convolutional neural networks. Pattern Recognition, 77:354–377.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.

[28] Balmukund Mishra, Deepak Garg, Pratik Narang, and Vipul Mishra. 2020. Drone surveillance for search and rescue in natural disaster. Computer Communications, 156:1–10.

[29] Xiaoxu Liu, Wei Qi Yan. 2021. Traffic-light sign recognition using capsule network. Multim. Tools Appl. 80(10): 15161-15171.

[30] Naresh Kumar and Nagarajan Sukavanam. 2018. Motion trajectory for human action recognition using Fourier temporal features of skeleton joints, Journal of Image and Graphics, Vol. 6, No. 2, pp. 174-180.

[31] Wei Qi Yan. 2021. Computational Methods for Deep Learning: Theoretic, Practice and Applications. Springer.

[32] Wei Qi Yan. 2019. Introduction to Intelligent Surveillance Surveillance Data Capture, Transmission, and Analytics. Springer.

[33] Chen Pan, Jianfeng Liu, Wei Qi Yan. 2021. Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

[34] Chen Pan, Wei Qi Yan. 2020. Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944.