Learning-based Human Fall Detection using RGB-D cameras

Szu-Hao Huang and Ying-Cheng Pan
Department of Industrial engineering and engineering management,
National Tsing Hua University, Hsinchu, Taiwan
shhuang@ie.nthu.edu.tw

Abstract

Automatic detection of human fall events is a challenging but important function of the real-time surveillance system. The goal of the proposed system is to develop a frame-by-frame fall detection system based on real-time RGB-D camera devices. The proposed system is composed of a complex off-line learning stage which combines several novel machine learning techniques and a series of on-line detection processes. A background subtraction method based on iterative normalized-cut segmentation algorithm is proposed to identify the pixel-wise human regions rapidly. The silhouettes are extracted to measure the pose similarity between different samples. Manifold learning algorithm reduces the feature dimensions and several discriminant analysis techniques are applied to model the final human fall detector. The experimental database contains 65 color video and corresponding depth maps. The experimental results based on a leave-one-out cross-validation testing show that our proposed system can detect the fall events effectively and efficiently.

1. Introduction

Traditional surveillance system [1] focused on the feasibility and efficiency of large-scale video recording and compression. In the health care applications, the elderly and patients will be watched by multiple digital cameras which have been installed in home environments previously. In order to decrease the costs of manual surveillance, an intelligent health-care system based on machine learning and computer vision techniques is necessary to assist the event detection and danger alerts.

With the development of 3D scanning technology, real-time consumer RGB-D cameras can be purchased in reasonable price, such as Microsoft Kinect and ASUS Xtion. The RGB-D camera includes a color image sensor and a depth maps capturing system. This kind of device can provide richer information than traditional image surveillance camera and potential benefits of low-cost, vision-based monitoring system design.

Previous fall detection system are mostly applied in the video surveillance system and the Hidden Markov Models [2-3] are used to model the human motion. In the action classification and recognition system, more discriminant feature representation methods may provide more information and achieve higher prediction accuracy. From the continuous video sequence, Rougier et al. [4] extracted 3D information and calculate the 3D trajectory of the user's head which is adopted for further analysis of 3D velocities characteristics. Auvinet et al. [5] used image analysis to reconstruct 3D human shape and position with a multiple cameras system.

In order to efficiently solve image processing and computer vision subproblems, hybrids methods which combine several different techniques are developed. Nait-Charif and McKenna [6] used an ellipse model as the tracker and adjusted the scenes with multiple sources of illumination by the particle filters. Hazelhoff et al. [7] adopted principal component analysis to determine the direction of the main axis of the body and the variances.

The stereo camera acquisition systems require equipment installation and calibration processes. The development of RGB-D camera provides more possibility for the application. Kepski and Kwolek [8] only adopted depth image sequence from Kinect and applied mean-shift clustering algorithm to accomplish reliable fall detection. Stone and Skubic [9] used the measurements of temporal and spatial gait parameters to facilitate an assessment method for developing a safely continue living environment.

However, the robustness of the previous systems highly depends on the precision of the pre-processes, such as head detection and human centroid estimation. In this paper, we proposed a silhouette-based method which can analyze the shape variations from large learning samples and achieve higher accuracy of fall detection.

In the off-line training stage of proposed system, several statistical machine learning techniques, including manifold learning and discriminant analysis, are proposed to generate the fall event detector. Various classification methods, such as linear discriminant analysis and AdaBoost algorithm, are integrated to achieve higher accuracy. In addition, an iterative normalized cut algorithm is proposed to extract the silhouette of human region in a more efficient way. Through the integration of RGB-D camera, computer vision, patter recognition and machine learning techniques, the proposed system can automatically detect the fall events from the video sequence and its corresponding depth maps. The experimental results, which are based on a leave-one-out cross-validation testing in 65 video datasets, show that our proposed system can detect the fall events effectively and efficiently.

2. Proposed Method

The proposed fall detection system consists of a series of statistical machine learning methods. The precise human silhouette can be extracted with the joint information of input video sequence and depth maps which are captured by RGB-D camera system. With the Hausdorff similarity measurements and a pre-learned action classifier, the proposed system can detect fall events in real-time surveillance system.

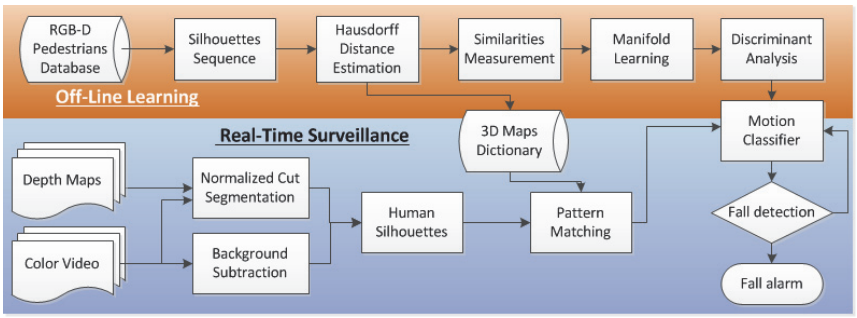

Figure 1. System Flow Chart

2.1 System Overview

Figure 1 illustrates the flow chart of our proposed fall detection system. The proposed system can be divided into two major stages, named as off-line learning stage and real-time surveillance stage. The goal of the off-line learning stage is to build the 3D maps dictionary and action classifier with various machine learning techniques. Collecting the Hausdorff lookup tables of all the learning pedestrian samples, 3D maps dictionary can be applied to estimate the silhouette similarity rapidly. Action classifier is designed to predict the label of each testing sample through the manifold space transformation and discriminant analysis methods. In the practical system design, various static or dynamic classification models are applied to form the different action classifiers to examine the discrimination.

In the real-time surveillance stage, each testing sample captured by RGB-D camera system is composed of a color image and the corresponding depth map. The human region segmentation based on iterative normalized-cut algorithm is proposed to extract the boundary information and generate the silhouette of testing human subject. With the simple pattern matching to the pre-learned Hausdorff dictionary, the testing silhouette can be transformed to a lower-dimensional space with similarity measurements. Finally, a fall event which is detected by action classifier in transformed space will be send to the remote monitoring center. The details of the image processing procedures and statistical learning algorithms are described in the following sub-sections.

2.2 RGB-D Camera System

In the proposed system, a consumer depth sensor, named as Microsoft Kinect, is adopted to construct the RGB-D images acquisition system. With the “Kinect for windows” software development kit released by Microsoft Corporation, the color image and depth maps could be captured simultaneously in a real-time processing system. Figure 2 shows the sample output images of Kinect acquisition system which includes an additional skeleton tracking results and human segmentation in depth map. There are two camera systems integrated in Kinect device, including a traditional RGB camera and a monochrome CMOS depth sensor combined with an infrared laser projector. The depth sensor can capture 3D video data under various light conditions with the structured light technology.

Figure 2. The color image, tracking human skeleton, and depth maps captured by Kinect

The learning datasets and testing samples in experiments are extracted from a RGB-D pedestrian database which includes depth maps and video sequences captured by Kinect system. This database can be divided into two main categories: walking people in normal condition and fall event video sequence. Totally 65 video sequences with 2,584 frames are recorded from 10 different human subjects. Several samples are shown in the Figure 3.

Figure 3. Learning samples in RGB-D database

2.3 Iterative Normalized Cut Segmentation

The human silhouette is defined as the contour of human region in color video. The proposed system has two alternative sub-systems to calculate the silhouette. The first sub-system is a simple background subtraction method which can be applied in the fixed network camera system. With a pre-defined background image without people and a threshold method, the pixels of foreground can be viewed as the human region.

In pan-tilt camera system or a moving platform, such as robots, we proposed a human region segmentation algorithm based on normalized-cut energy minimization. In traditional normalized-cut algorithm, the image segmentation problem could be formulated as a graph optimization problem. Each pixel can be viewed as a graph node and the linked edges can be modeled by the color similarity and spatial relationship. This weighted graph can be represented as G = (V,E,W), where V is the collection of pixel nodes, E represents the connection edges with neighbor nodes, and W recorded the weights between nodes which can be calculated by pixel similarity.

The weight Wi,j between pixel i and j is defined as the follows:

$$W_{i,j} = e^{\frac{-|I(i)-I(j)|}{\sigma_I}} * \begin{cases} \frac{-\| x(i)-x(j) \|_2^2}{\sigma_x} & if \ \| x(i)-x(j) \|_2^2 < r \\ 0 & otherwise, \end{cases} \quad (1)$$

In addition, an iterative algorithm based on region growing and shrinking framework is proposed as shown in Figure 4.

1.Initialize vector x.
   (a) Set initial human regions by depth sensor.
   (b) Set $x_i = 1$ for all the pixels i in the current human region (inside the boundary).
   (c) Set $x_i = -1$ for all the pixels i in the background region.
2. Calculate the weight matrix W by pixel simularity.
3. From x, determine y = P+bN. Where vector P_i is set to 1 when x_i is 1 and set to 0 otherwise, while the vector N_i is set to -1 when x_i is -1 and set to 0 otherwise. The value b is defined as:
   $b = \sum_{x_i>0} d_i / \sum_{x_i<0} d_i$.
4. Calculate $Term_{PP} = P^T WP$, $Term_{PN} = P^T WN$, $Term_{NN} = N^T WN$.
5. Calculate $E = \frac{y^T Wy}{y^T Dy}$ with $Term_{PP}$, $Term_{PN}$, $Term_{NN}$, b.
6. Iterative updating
   (a) Set all the spatial neighboring pixels outside the current boundary as the set of growing candidates.
   (b) For each pixel i in the growing candidate set,
   · Set $x_i' = 1$ (x_i should be -1 because pixel i is originally outside the boundary).
   · Set vector M and estimate constant b' as follows:
   $$M_i = \begin{cases} -1 & \text{when } x_i = 1 \text{ and } x_i' = -1 \quad \text{(region shrinking)} \\ 1 & \text{when } x_i = -1 \text{ and } x_i' = 1 \quad \text{(region growing)} \\ 0 & \text{when } x_i = x_i' \end{cases}$$
   Calculate $M^T WP$, $M^T WN$, and $M^T WM$.
   · Calculate $E' = \frac{y'^T Wy'}{y'^T Dy'}$ with $Term_{PP}$, $Term_{PN}$, $Term_{NN}$, $M^T WP$, $M^T WN$, $M^T WM$, and b'.
   · If E' > E, update vector x and the following terms
   $Term_{PP} = Term_{PP} + 2M^T WP + M^T WM$
   $Term_{PN} = Term_{PN} + M^T WP + M^T WN + M^T WM$
   $Term_{NN} = Term_{NN} + 2M^T WN + M^T WM$
   b = b', E = E', and $x_i = x_i'$.
   (c) Set all the spatial neighboring pixels inside the current boundary as the set of shrinking candidates, then apply the updating process similar to step 6(b) but $x_i' = -1$ and $x_i = 1$.
   (d) If E is not converged, repeat step 6(a), (b) &(c).
7. Output vector x as the segmentation result.

Figure 4. Iterative normalized cut human region segmentation algorithm

2.4 Discriminant Analysis and AdaBoost

Linear discriminant analysis (LDA) or related fisher's linear discriminant is a supervised learning technique which finds a linear combination of features to separate two or more classes of training samples with labels. The basic idea of LDA is highly similar to principal component analysis (PCA) in that they both try to explain the data with feature space transformation and solve an eigen-system. With a projection process, LDA attempts to maximize the ratio estimated from between-class and within-class data variance. In addition, LDA has been widely used in various applications which need to classify samples into multiple categories or reduce the feature dimensions of each sample.

Several discriminant analysis methods are developed while these methods use different measurements of the distance. In our experiment, we adopted four kinds of discriminant analysis, including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), diagonal linear discriminant analysis (DLDA), and diagonal quadratic discriminant analysis (DQDA).

Adaboost, short for adaptive boosting, is a meta algorithm which can cooperate with other machine learning techniques and improve their performance. Freund and Schapire [10]. In the AdaBoost algorithm, WeakLearn is a function or an algorithm which performs the hypothesis to classify learning samples into different categories by considering the current sample weights. The word weak means the hypothesis is not expected to be very powerful. In our proposed system, WeakLearn is defined as a binary function of single feature value. The basic idea is that the attributes of WeakLearn are easy to calculate and a little bit better than random guess. Then the AdaBoost learning algorithm will compose many weak classifiers to form a final strong classifier.

3. Experimental Results

The experimental datasets consist of 2,584 image samples from 65 video. We applied a leave-one-out error measurement for training and testing framework. In other words, the image samples from 65 video are collected as learning data and the other one is adopted to be the testing video. Two different experiments are designed to prove the effectiveness of our proposed fall detection system.

3.1 System Comparisons

Previous fall detection systems can be divided into the RGB camera system [2-7] and RGB-D camera system [8-9]. Various algorithms are proposed to achieve efficient fall detection tasks. However, it is difficult to compare the system performance in numerical experimental results because the problem settings and testing datasets are totally different. Hence, we tried to list the characteristics of our proposed system to show the progress of our design in fall detection problem.

• Frame-by-frame decision scheme: The proposed system can suggest the per-frame decision of the fall event. It is an important attribute to assist the design of electronic travel aid devices. In addition, it also achieves near real-time alarms of fall events.

• Usage of simultaneous color images and depth maps: Early works usually adopts multiple view information which is captured from multiple cameras to increase the detection accuracy. The proposed decision support system is based on the color images and the corresponding depth maps. The rich 3D information in depth image can decrease the complexity of human region segmentation.

• Single RGB-D camera: Comparing to other related works based on RGB-D cameras, our system only adopted single sensor for the ease of implementations and environmental construction. The combination of the depth information and color images from single view is adequate to determine the target event.

• Computational efficiency: The time consumption of our proposed system is constrained by real-time design. The improvement of iterative normalized-cut and feature reduction with manifold learning dramatically decrease the computational complexity and increase the executional efficiency.

3.2 Classification Algorithms

The accuracy of different classification algorithms is compared in this section, which include four discriminant analysis methods and AdaBoost classification. Two accuracy indexes are defined in order to measure the detection hit rates. Among the 2,584 frames from 65 videos, the average accuracy of frame decisions is measured with the manual labels. The other index records the frame accuracy of the worst testing video. In discriminant analysis, the dimension of the datasets are all reduced to lower-dimension space (dimension=5) through Isomap. The features are also reduced to 5 in Adaboost learning to achieve fair comparisons.

Table 1. Frame accuracy comparisons between different classification algorithms

	Average	Worst Case
LDA	0.8861	0.5263
QDA	0.9350	0.7368
DLDA	0.8669	0.3889
DQDA	0.9291	0.5789
AdaBoost	0.9247	0.7143

Among the four discriminant analysis methods, quadratic discriminant analysis (QDA) shows the best results. The performance of QDA and AdaBoost are both with high stability. The average frame accuracy of the two methods is higher than other methods, and the estimation accuracy can still achieve at least 70% in the worst case. The performance of diagonal linear discriminant analysis (DLDA) is the worst among all the methods.

3.3 Manifold Dimension Reduction

QDA and Adaboost show higher detection accuracy than previous experiments. Here we adopted both of the methods to analyze the different dimension settings of manifold space learning. In QDA, the samples are projected to different dimension spaces through Isomap, and then the samples are classified by using discriminant analysis methods. We compare the accuracy while using different manifold dimension settings. The comparison of different dimensions can also be examined by AdaBoost. In this experiment, we use the dataset which is reduced to 10-dimensional Isomap space, and by using simple binary weak learners through different iteration training the multi-dimensional strong classifier can be generated.

Table 2. Frame accuracy comparisons of the manifold dimensions

Dimension	QDA	AdaBoost
1	0.8994	0.9320
2	0.9181	0.9320
3	0.9282	0.9320
4	0.9318	0.9320
5	0.9350	0.9247
6	0.9370	0.9354
7	0.9384	0.9348
8	0.9376	0.9363
9	0.9364	0.9336
10	0.9332	0.9354

High-dimensional Isomap preserves rich information in QDA. However, the difficulty of classification also increases due to the influence of noise. In this experiment, average frame accuracy increases as long as addition of dimension and the maximum is reached as dimension is 7. Similar circumstances show in AdaBoost, the maximum accuracy is reached in 8 iterative learning. In 140 decisions of whether human subject falls or not, only six of the data are judged incorrectly, which shows high applicability in practical software designing.

4. Conclusion

In this paper, a learning-based fall detection system based on RGB-D surveillance is proposed as a fundamental research of health care system. The proposed method includes a hybrid learning algorithm and a real-time detector. The off-line learning stage adopted various statistical machine learning techniques, includes manifold space learning, discriminant analysis and AdaBoost binary classification. With the iterative normalized-cut human segmentation and Hausdroff distance measurements, a silhouette map dictionary can be generated for rapid similarity measurement. The real-time surveillance system can detect the fall event frame-by-frame with the silhouette dictionary and action classifier. The experimental results show that our proposed method can achieve near 94% frame accuracy.

Acknowledgments

This work was supported by the Advanced Manufacturing and Service Management Research Center (AMSMRC), National Tsing Hua University.

References

[1] N. M. Barnes, et al.: “Lifestyle monitoring: technology for supported independence,” Computing and Control Engineering *Journal*, pp 169–174, 1998.

[2] B. T¨oreyin, et al.: “HMM based falling person detection using both audio and video,” *In Proc. IEEE Int. Workshop Hum.-Comput. Interaction*, pp. 1–4, 2005.

[3] D. Anderson, et al.: “Recognizing falls from silhouettes,” *In Proc. Int. Conf. IEEE EMBS*, pp. 6388–6391, 2006.

[4] C. Rougier, et al.: “Monocular 3-D head tracking to detect falls of elderly people,” *In Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.,* pp. 6384–6387, 2006.

[5] E. Auvinet, et al.: “Fall detection using multiple cameras,” *In Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2554–2557, 2008.

[6] H. Nait-Charif , et al.: “Activity summarization and fall detection in a supportive home environment,” *In Proc. 17th ICPR*, vol. 4., pp. 323–326, 2004.

[7] L. Hazelhoff, et al.: “Video-based fall detection in the home using principal component analysis,” *In Proc. Adv. Concepts Intell. Vision Syst.*, vol. 1, pp. 298–309, 2008.

[8] M. Kepski, et al.: “Human Fall Detection by Mean Shift Combined with Depth Connected Components,” *Lecture Notes in Computer Science*, vol. 7594, pp. 457-464, 2012.

[9] E.E. Stone, et al.: “Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment,” *In Proc. 5th Int. Conference on Pervasive Computing Technologies for Healthcare,* pp. 71-77, 2011.

[10] Y. Freund, et al.: “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *Journal of Computer and System Science*, Vol. 55, pp. 119-139, 1997.