Neurocomputing 207 (2016) 726–734

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Human fall detection in videos by fusing statistical features of shape and motion dynamics on Riemannian manifolds

Yixiao Yun*, Irene Yu-Hua Gu

Department of Signals and Systems, Chalmers University of Technology, Gothenburg SE-41296, Sweden

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
10 May 2016
Accepted 19 May 2016
Communicated by Marco Cristani
Available online 6 June 2016

Keywords:
Human fall detection
Riemannian manifolds
Dynamic shape and motion
Support vector machines
Elderly care
Assisted-living

ABSTRACT

This paper addresses issues in fall detection in videos. We propose a novel method to detect human falls from arbitrary view angles, through analyzing dynamic shape and motion of image regions of human bodies on Riemannian manifolds. The proposed method exploits time-dependent dynamic features on smooth manifolds based on the observation that human falls often involve drastically shape changes and abrupt motions as comparing with other activities. The main novelties of this paper include: (a) representing videos of human activities by dynamic shape points and motion points moving on two separate unit *n*-spheres, or, two simple Riemannian manifolds; (b) characterizing the dynamic shape and motion of each video activity by computing the velocity statistics on the two manifolds, based on geodesic distances; (c) combining the statistical features of dynamic shape and motion that are learned from their corresponding manifolds via mutual information. Experiments were conducted on three video datasets, containing 400 videos of 5 activities, 100 videos of 4 activities, and 768 videos of 3 activities, respectively, where videos were captured from cameras in different view angles. Our test results have shown high detection rate (average 99.38%) and low false alarm (average 1.84%). Comparisons with eight state-of-the-art methods have provided further support to the proposed method.

1. Introduction

There is an increasing demand for assisted-living in elderly care, due to the rapid growth of ageing population in the world. According to the statistics, falling down poses a high risk to the ageing group, as this could lead to bone fracture, stroke, and other health emergences [1,2]. In such cases, urgent medical attentions are needed. Since many persons in this ageing group live alone, it is difficult for them to seek help immediately. Automatic e-health care systems have drawn increasing interests and public awareness recently. Among many possibilities in such systems, detecting falls followed by triggering alarms is one of the main issues.

Previous work: Many existing methods use wearable devices with motion sensors like accelerometers and gyroscopes [1,2], and are able to produce reasonable results for the fall detection. For example, Bourke et al. [3] suggested a method to distinguish falls from activities of daily living (ADL) by using tri-axial accelerometer sensors mounted on the trunk and thigh of users, and by thresholding the resultant signals. Kwolek and Kepski [4,5] employed a tri-axial accelerometer to indicate falls if a measured acceleration value is higher than a predetermined threshold, and used depth maps obtained from a Kinect sensor to authenticate the fall alert. However, users can feel uncomfortable after wearing the devices for a long time, or forget to wear them sometimes, apart from requiring frequent battery charging. Visual monitoring hence provides some advantages when the privacy issue is properly handled.

Much effort has been made to detect human falls in videos. One way to address this problem is to analyze the bounding boxes that encompass the target person in each frame. Miaou et al. [6] used an omni-camera mounted on the ceiling to capture videos. A fall is then detected if the aspect ratio of bounding box exceeds a predetermined threshold. Qian et al. [7] employed two bounding boxes, one for the whole body, another for the lower body part. Variations of these two boxes are used as the feature for fall detection by using a SVM classifier. Debard et al. [8] extracted four features from the bounding box to describe a fall, including the aspect ratio, torso angle, speed of motion in box center and head. A SVM classifier is then employed to detect the fall using these features. Charfi et al. [9] defined fourteen features based on the parameters of bounding box, e.g. height, width, aspect ratio, and centroid of the box. Transforms (e.g., Fourier, wavelet) are then applied to these features before fall detection by using either a SVM or AdaBoost classifier. The major drawback is the insufficient description of human motion by using the bounding box alone. Furthermore, the performance is heavily dependent on the camera

* Corresponding author.
E-mail address: yixiao@chalmers.se (Y. Yun).

http://dx.doi.org/10.1016/j.neucom.2016.05.058
0925-2312/

view angles.

Another commonly adopted way is to use the wide spatial coverage of multiple cameras, or the depth information from depth cameras. Auvinet et al. [10] reconstructed a 3D volume of a person from eight cameras based on the camera calibration. A fall is indicated if a large portion of body volume is found near the ground for a certain duration of time. Rougier et al. [11] computed the cost between the consecutive frames by shape matching, and use a criterion for describing the shape deformation used in a GMM classifier. A fall is decided by a majority voting from four camera views. Hung and Saito [12] computed the occupied area of a person by multiplying the widths of the minimum bounding boxes from two orthogonal views. The information is then used to determine whether the person is standing, sitting, or lying on the ground. A fall is detected if the occupied area exceeds a threshold and the person remains on the ground for an extended time duration. Ma et al. [13] obtained human silhouettes from depth images and learn curvature scale space (CSS) features. Actions are represented by a bag of CSS words, and classified by the extreme learning machine (ELM) for the fall and other action classes. Stone and Skubic [14] modeled the vertical state of 3D object in each depth video frame, and segmented the time series between the states where or not a person is on the ground. An ensemble of decision trees is then used to compute the confidence that a fall occurs before a person is on the ground. It is worth noting the trade-off between the performance and complexity (or cost) in multi-camera or depth-camera methods.

In this paper, we propose a novel scheme for fall detection. The method employs a manifold-based analysis of dynamic shape and motion from videos using a single camera with arbitrary view angles.

Motivation: Instead of using parameters of bounding boxes, our focus is on the analysis of dynamic shape and motion of human body inside the box. Based on the observations that the shape and motion vary significantly when a person is falling, better features could be extracted from the rate of changes in shape and motion during a certain time interval. A suitable metric is required for measuring the rate. The Riemannian geometry fulfills such a requirement, under the assumption that many image features, including shape and motion, can be effectively represented by points on a low-dimensional Riemannian manifold. Therefore, we treat dynamic shape (or motion) as connected points on a Riemannian manifold. It is worth mentioning that such a description on the manifold is not dependent on different camera views. By studying the velocity statistics of dynamic shape and motion of human body on the Riemannian manifold, the features are expected to be less sensitive to the camera view angle. This can lead to a simple and effective solution, without exploiting camera geometry and requiring camera calibration.

Contributions: The main contributions of the proposed method are: (a) the dynamic features of a falling person are represented as points moving on the Riemannian manifold; (b) a human fall event is characterized by a feature vector of velocity statistics, and is correspondent to a moving point on the manifold by using the geodesics; (c) the statistical feature vectors of shape and motion on the two manifolds are weighted and combined according to their mutual information; (d) extensive tests were conducted on three video datasets (containing more than 1200 videos with different complexities), where the results achieved are comparable to state-of-the-art methods, including the ones based on multi-camera calibration and multi-modal information.

The remainder of this paper is organized as follows: Section 2 briefly reviews the background theory. Section 3 gives the big picture of the proposed method and then describes the main steps. Section 4 describes the classification-based fall detection scheme. Section 5 shows experimental results and comparisons on three video datasets for fall detection. Finally, Section 6 concludes the paper.

2. Theoretical background

This section briefly reviews the Riemannian geometry [15] and the unit n-sphere [16], for the sake of mathematical convenience in the subsequent sections.

2.1. Riemannian geometry

A *manifold* is a topological space, consisting of a complete set of low dimensional subspaces embedded in a high dimensional space. It is locally similar to the Euclidean space. For nonlinear manifolds, the usual Euclidean calculus and conventional statistics in vector spaces may not apply. A *differentiable manifold* equipped with a globally defined differential structure allows one to perform calculus on the manifold using special metrics. A *Riemannian manifold* is a differentiable manifold, where the tangent space at each manifold point has an inner product that varies smoothly from point to point.

The *geodesic* is defined as the shortest curve between the two points on a manifold. The *geodesic distance*, the length of the geodesic, is the distance measure between two points on the manifold.

2.2. The unit n-sphere

The unit n-sphere, $\mathcal{S}^n$, is an n-dimensional sphere with a unit radius, centered at the origin of the $(n+1)$-dimensional Euclidean space. An example for $n=2$ case is illustrated in Fig. 1. A unit n-sphere is defined as

$$\mathcal{S}^n = \{\mathbf{p} \in \mathbb{R}^{n+1}: \|\mathbf{p}\| = 1\}, \tag{1}$$

and can be considered as the simplest Riemannian manifold after the Euclidean space [17]. It inherits a Riemannian metric from embedding in $\mathbb{R}^{n+1}$. Under this metric, the geodesic distance $\rho(\mathbf{p}, \mathbf{q})$ between the two manifold points $\mathbf{p}, \mathbf{q} \in \mathcal{S}^n$ is the great-circle distance between them:

$$\rho(\mathbf{p}, \mathbf{q}) = \arccos(\mathbf{p}^T\mathbf{q}), \tag{2}$$

where $\arccos(\cdot)$ is the inverse cosine function [16]. It is worth noting that the great-circle distance between the two points is unique.

The unit n-sphere finds its connection to some computer vision tasks where the extracted feature vectors of objects are often normalized by the ℓ^2 norm. The descriptors thus lie on a unit n-sphere $\mathcal{S}^n$ for some n. In the cases where feature vectors are normalized block-wisely, the radius of the underlying sphere is not unit. However, since any n-dimensional sphere centered at the origin is homeomorphic to $\mathcal{S}^n$, it turns out that they share exactly

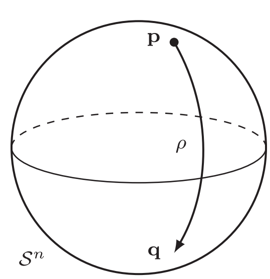

Fig. 1. Example of an n-sphere $\mathcal{S}^n$ $(n=2)$ embedded in an $(n+1)$-D space $\mathbb{R}^{n+1}$. $\mathbf{p}$ and $\mathbf{q}$ are the manifold points, $\mathbf{p}, \mathbf{q} \in \mathcal{S}^n$. The geodesic ρ is the shortest curve between $\mathbf{p}$ and $\mathbf{q}$ on the manifold.