Computer Vision and Image Understanding 148 (2016) 111–122

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Human fall detection in videos via boosting and fusing statistical features of appearance, shape and motion dynamics on Riemannian manifolds with applications to assisted living

Yixiao Yun*, Irene Yu-Hua Gu

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

ARTICLE INFO

Article history:
Received 14 April 2015
Accepted 3 December 2015

Keywords:
Human fall detection
Riemannian manifolds
Boosting
Elderly care
Assisted living

ABSTRACT

This paper addresses issues in fall detection from videos. It is commonly observed that a falling person undergoes large appearance change, shape deformation and physical displacement, thus the focus here is on the analysis of these dynamic features that vary drastically in camera views while a person falls onto the ground. A novel approach is proposed that performs such analysis on Riemannian manifolds, detecting falls from a single camera with arbitrary view angles. The main novelties of this paper include: (a) representing the dynamic appearance, shape and motion of a target person each being points moving on a different Riemannian manifold; (b) characterizing the dynamics of different features by computing velocity statistics of their corresponding manifold points, based on geodesic distances; (c) employing a feature weighting approach, where each statistical feature is weighted according to the mutual information; (d) fusing statistical features learned from different manifolds with a two-stage ensemble learning strategy under a boosting framework. Experiments have been conducted on two video datasets for fall detection. Tests, evaluations and comparisons with 6 state-of-the-art methods have provided support to the effectiveness of the proposed method.

1. Introduction

Recent decades have witnessed the rapid growth of aged population in most countries. According to statistics, falling on the ground is one of the most vital risks for this age group which may lead to bone fracture, coma, and even death [1]. In these cases, emergent medical attentions are necessary after the fall. Since many people in this age group live alone, it can be difficult for them to seek help immediately, especially when severe injury or unconsciousness occur due to the fall. Automatic surveillance systems have drawn increasing research interests recently, aiming at automatically detecting falls and triggering alarms.

Many existing methods are based on wearable devices with motion sensors, such as accelerometers and gyroscopes [1], which produce reasonably good results in fall detection. However, users could feel uncomfortable after wearing the device for a long time, or sometimes forget to wear them. Besides, wearable devices consume batteries that need to be replaced or recharged frequently. Visual monitoring hence has some advantages, where users are freed from wearing devices with minimum disturbance to their daily lives. Although it has been argued that video surveillance may cause the exposure of personal privacy, the problem can be effectively avoided by not directly using the raw video but only extracting numerical features from it, where the personal identity is not revealed.

Existing work: 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. [2] use an omni-camera that is mounted on the ceiling to capture image sequences, and a fall is detected if the aspect ratio of the bounding box exceeds a predetermined threshold. Qian et al. [3] employ two bounding boxes, one for the whole body, the other for the lower part of the body. Variations of these two boxes are calculated as the features for fall detection based on SVM classification. Debard et al. [4] extract 4 features from the bounding box to describe a fall, including aspect ratio, torso angle, center speed and head speed. An SVM classifier is employed to detect a fall using these features. Charfi et al. [5] define 14 features based on the bounding box such as height and width, aspect ratio, and centroid coordinates of the box. Transforms (e.g., Fourier, wavelet) are applied to these features before fall detection through SVM and AdaBoost classification. The major drawback is insufficient description of the motion from using the bounding box, and the performance is also heavily dependent on view angles.

Another commonly adopted strategy is to exploit the wide spatial coverage of multiple cameras, or the depth information from depth

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

http://dx.doi.org/10.1016/j.cviu.2015.12.002
1077-3142/

cameras. Auvinet et al. [6] reconstruct a 3D volume of the person from 8 cameras based on camera calibration, and a fall is indicated if a large portion of the body volume is found near the ground for a certain period of time. Rougier et al. [7] calculate a cost between consecutive frames by shape matching, and feed the costs as criteria for shape deformations to a GMM classifier. The occurrence of a fall is decided by majority voting from 4 cameras. Ma et al. [8] obtain human silhouettes from depth images and learn curvature scale space (CSS) features from them. Actions are represented by a bag of CSS words, and classified by the extreme learning machine (ELM) into falls and other actions. Stone and Skubic [9] model the vertical state of a 3D object in each depth image frames, and segment the time series in on-ground state from those in vertical state. Then, an ensemble of decision trees is used to compute a confidence that a fall occurs before an on-ground state. 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 adopting manifold-based feature analysis from a single camera with arbitrary view angles, with the following motivations and contributions.

Motivations: instead of using bounding boxes, the focus here is shifted to the analysis of different features of the target person inside the box. Since it is a broadly accepted intuition that these image features vary drastically in camera views while the person falls onto the ground, better features could be obtained by studying their changing rate in a certain time interval. A suitable metric is preferred for measuring the rate. Riemannian geometry fulfills this requirement, given the assumption that these image features can be effectively described by low-dimensional data points on Riemannian manifolds. By converting the analysis of dynamic features of a target person in an arbitrary camera view to the study of velocity statistics on a unified manifold for different camera views, it is expected that these features are less sensitive to view angles. This could lead to a simple and effective solution, without combining multiple cameras.

Contributions: the appearance, shape and motion of a target person are represented each as points moving on a different Riemannian manifold. It is worth noting that each of them is a unified manifold for different camera views. Then, the characterization of the dynamics of these features are formulated as computing velocity statistics of moving points on the corresponding manifold, based on geodesic distances. Besides, a feature weighting scheme is adopted, where the weight of each statistical feature is obtained by computing the mutual information. Moreover, a boosting framework is used where a two-stage ensemble learning is conducted for fusing the statistical features learned from each manifold.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3 gives a big picture of the proposed method and illustrates each major step in detail. Section 4 describes the pre-processing of video data including foreground human detection and video event segmentation as automatic processes. Section 5 shows experimental results on two video datasets for human fall detection. Finally, Section 6 concludes the paper.

2. Review of related work

This section briefly reviews Riemannian geometry [10] including the space of symmetric positive definite matrices [10] and the unit n-sphere [11], and adaptive boosting [12], for the sake of mathematical convenience in subsequent sections.

2.1. Riemannian geometry

A manifold is an efficient low-dimensional representation of high dimensional data. It maintains important properties of the data such as topology and geometry [10]. In case of nonlinear manifolds which are not vector spaces, the usual Euclidean calculus and conventional

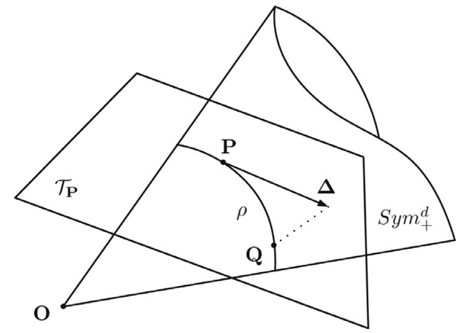

Fig. 1. Example of 2-D convex cone embedded in a 3-D space $\mathbb{R}^3$. $\mathbf{O}$ is the origin. $\mathbf{P}$ and $\mathbf{Q}$ are manifold points, i.e., $\mathbf{P}, \mathbf{Q} \in Sym^d_+$. $\mathcal{T}_{\mathbf{P}}$ is the tangent space at $\mathbf{P}$. $\mathbf{\Delta} \in \mathcal{T}_{\mathbf{P}}$ is the tangent vector whose projected point on the manifold is $\mathbf{Q}$. The geodesic ρ is the shortest curve between $\mathbf{P}$ and $\mathbf{Q}$ on the manifold.

statistics may not apply. However, a *differentiable manifold* equipped with a globally defined differential structure allows one to perform calculus on the manifold. Further, a *Riemannian manifold* is defined as a differentiable manifold where the tangent space at each point has an inner product that varies smoothly from point to point. That is, a Riemannian manifold possesses not only the differentiable structure that allows calculus to be done, but also a Riemannian metric that allows distances and angles to be measured on the manifold.

The *geodesic* is the shortest curve between the two points on a manifold. Geodesics correspond to straight lines in Euclidean spaces. Hence, *geodesic distance*, the length of the geodesic, is the most suitable distance measure between two points lying on a Riemannian manifold.

2.1.1. Space of symmetric positive definite matrices

The space of $d \times d$ symmetric positive definite (SPD) matrices (Sym^d_+) is an open convex cone

$$Sym^d_+ = \bigcap_{\mathbf{x} \in \mathbb{R}^d} \{\mathbf{P} \in Sym^d : \mathbf{x}^T \mathbf{P} \mathbf{x} > 0\} \tag{1}$$

whose strict interior is a Riemannian manifold [10]. To compute the statistics on Sym^d_+, the *affine-invariant* metric [13] and the *log-Euclidean* metric [14] are commonly used. These two metrics are mathematically equivalent, however, numerical results can differ. This paper uses log-Euclidean metric as it is computationally more efficient [14].

As shown in Fig. 1, *exponential map* ($\exp_{\mathbf{P}}(\cdot) : \mathcal{T}_{\mathbf{P}} \mapsto Sym^d_+$) and *logarithmic map* ($\log_{\mathbf{P}}(\cdot) : Sym^d_+ \mapsto \mathcal{T}_{\mathbf{P}}$) are a pair of operators mapping between the manifold Sym^d_+ and the tangent space at $\mathbf{P}$:

$$\exp_{\mathbf{P}}(\mathbf{\Delta}) = \exp(\log(\mathbf{P}) + \mathbf{\Delta}) = \mathbf{Q} \tag{2}$$

$$\log_{\mathbf{P}}(\mathbf{Q}) = \log(\mathbf{Q}) - \log(\mathbf{P}) = \mathbf{\Delta} \tag{3}$$

where $\mathbf{\Delta} \in \mathcal{T}_{\mathbf{P}}$ is the tangent vector whose projected point on the manifold is $\mathbf{Q}$ [15], $\exp(\cdot)$ is the matrix exponential [14], and $\log(\cdot)$ is the principal logarithm of a matrix which is defined as the inverse of the matrix exponential [14]. The geodesic distance between $\mathbf{P}$ and $\mathbf{Q}$ is computed from [15]

$$\rho(\mathbf{P}, \mathbf{Q}) = \| \log_{\mathbf{P}}(\mathbf{Q}) \| = \| \log(\mathbf{Q}) - \log(\mathbf{P}) \| \tag{4}$$

where $\| \cdot \|$ is the Frobenius norm of a matrix.

The space of SPD matrices is linked to region covariance [16] due to the fact that SPD cone is exactly the set of non-singular covariance matrices. Region covariance is an effective descriptor of object appearance features, and is shown to be robust and versatile for variations in illuminations, views and poses at modest computational cost by using integral images. Given an image region, let $\mathbf{f}$ be the d-dimensional feature vector for each pixel inside it. The features can