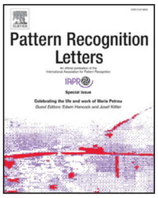
Pattern Recognition Letters 68 (2015) 103–110

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



On how to improve tracklet-based gait recognition systems☆

Manuel J. Marín-Jiménez[a,*], Francisco M. Castro[b], Ángel Carmona-Poyato[a], Nicolás Guil[b]

[a] *Department of Computing and Numerical Analysis, University of Cordoba, Campus de Rabanales, Cordoba 14071, Spain*
[b] *Department of Computer Architecture, University of Malaga, Campus de Teatinos, Malaga 29071, Spain*

ARTICLE INFO

Article history:
Received 17 April 2015
Available online 12 September 2015

Keywords:
Gait recognition
Tracklets
DTF
Metric learning
Binarization

ABSTRACT

Recently, *short-term dense trajectories features* (DTF) have shown state-of-the-art results in video recognition and retrieval. However, their use has not been extensively studied on the problem of gait recognition. Therefore, the goal of this work is to propose and evaluate diverse strategies to improve recognition performance in the task of gait recognition based on DTF. In particular, this paper will show that (i) the proposed RootDCS descriptor improves on DCS in most tested cases; (ii) selecting *relevant trajectories* in an automatic way improves the recognition performance in several situations; (iii) applying a *metric learning* technique to reduce dimensionality of feature vectors improves on standard PCA; and (iv) binarization of low-dimensionality feature vectors not only reduces storage needs but also improves recognition performance in many cases. The experiments are carried out on the popular datasets CASIA, parts B and C, and TUM-GAID showing improvement on state-of-the-art results for most scenarios.

1. Introduction

The task of identifying people based on the way they walk is known as *gait recognition*. Popular approaches for gait recognition are mainly based on features extracted from sequences of binary silhouettes [3,6,9]. In contrast, Castro et al. [4] presented recently a fully automatic system for gait recognition based on dense *tracklets* (i.e. short-term point trajectories) [13,21,22] showing an excellent recognition accuracy on a multiview setup. However, there is still room for improvement on the proposed approach. In particular, (a) instead of using all the tracklets obtained by dense sampling, it might happen that only a subset of them would be really discriminative; (b) the capability of representation of the low-level motion descriptors could be improved by properly post-processing them; (c) instead of performing dimensionality reduction on the video-level descriptors with Principal Components Analysis (PCA), which does not take into account category information, a semisupervised metric learning technique could offer a better representation in a new feature space; (d) rather than formulating gait recognition as a classification problem, defining it as a verification problem (i.e. are subject A and B the same one?) could broaden its applicability; (e) instead of using real-valued video-level descriptors, the transformation to binary descriptors would help to reduce the memory needs in large-scale databases; and (f) a better strategy to assign label identities to the test subjects could be defined on top of the previously used 'one-vs.-all' ensemble of binary classifiers to improve the recognition accuracy of the system.

Therefore, the main contribution of this paper is a thorough experimental evaluation of all the previously mentioned improvements on the popular datasets CASIA [25], parts B and C, and TUM-GAID [10]. In CASIA dataset, several challenging situations are evaluated, as carrying bags, wearing long coats, walking outside during night at different velocities, among others. While in TUM-GAID, we focus on people recorded on two different seasons (with the corresponding differences in clothing) plus carrying bags and wearing coating shoes. The experimental results will show that in most situations each of the proposed improvements help to increase the recognition performance of tracklet-based gait recognition systems, which have already shown state-of-the-art results [4]. Moreover, from our point of view, the findings of this study can be directly applied to more general human action recognition scenarios.

The rest of the paper is organized as follows. After presenting the related works, the proposed methodology is presented in Section 3. The experimental results are presented and discussed in Section 4. And, finally, Section 5 presents the conclusions.

2. Related works

Many research works have been published in recent years tackling the problem of gait recognition. For example, a complete survey on this problem can be found in [11]. One of the most successful approaches, proposed by Han and Bhanu [9], is called 'Gait Energy

☆ This paper has been recommended for acceptance by R. Davies.

[*] Corresponding author. Tel.: +34- 957218980; fax: +34- 957218360.
E-mail address: mjmarin@uco.es (M.J. Marín-Jiménez).

http://dx.doi.org/10.1016/j.patrec.2015.08.025
0167-8655/© 2015 Elsevier B.V. All rights reserved.

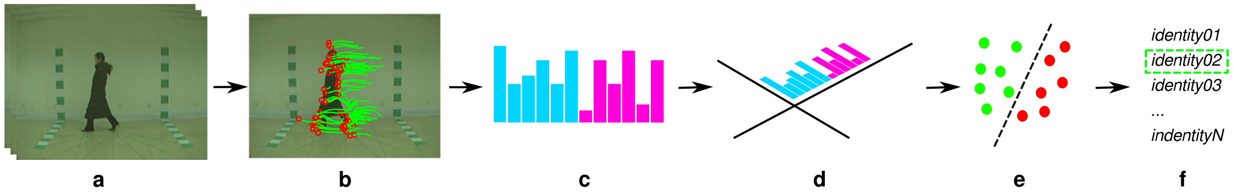


Fig. 1. Pipeline of the evaluated gait recognition system. (a) Input video. (b) Person-focused tracklets. (c) Pyramidal Fisher Motion descriptor. (d) Projected and compressed descriptor. (e) Classifier (e.g. SVM or NN). (f) An identity is selected from the classification scores.

Image'. This descriptor provides a spatio-temporal representation of the human gait from a sequence of binary silhouettes of people walking. Although it achieves a good representation of the human motion, since it relies on silhouettes, very good image conditions are required for background segmentation, or the application of sophisticated techniques for silhouette extraction [1]. In contrast, tracklet-based methods have shown good robustness in diverse challenging situations [4] eliminating previous restrictions as fine-grained segmentation of people or 3D body reconstruction to be able to deal with curved trajectories [12].

One of the ideas covered in this work is the selection of tracklets of interest. A similar idea has been previously addressed by Chakraborty et al. [5] for the problem of human action recognition with STIP-based descriptors. They propose a method that imposes local and temporal constraints on a set of detected STIPs in order to achieve robustness to camera motion and background clutter, showing outstanding results in state-of-the-art datasets in that moment. In that sense, we propose a novel approach to suppress uninformative tracklets.

In [2], it is proposed an improvement of SIFT descriptor for object retrieval, named 'RootSIFT'. The key idea is to normalize SIFT descriptors with the squared root operator in order to make directly possible a Hellinger distance computation on the descriptors. Their experimental results showed that this modification helped to boost the experimental results on the object retrieval task. We adopt this finding for our previously proposed Pyramidal Fisher Motion (PFM) descriptor [4].

The dimensionality of the PFM descriptors is generally large, therefore, a dimensionality reduction step is usually applied before the learning stage. PCA is, without any doubt, the most popular approach for unsupervised dimensionality reduction. However, we could use some prior information to learn a more discriminative projection space where gait descriptors of different individuals were located in 'far' locations in the new space, whereas the gait descriptors of the same subject were projected in 'near' locations. In [18], several metric learning techniques are proposed and evaluated in the context of face recognition, achieving excellent results. We extrapolate such idea to our problem, learning a discriminative projection matrix for compressing PFM descriptors.

An effective method to generate compact binary descriptors from real-valued ones is described in [14], allowing to reduce storage requirements in large-scale problems. In addition, the use of binary descriptors allows to compare them very quickly by using the Hamming distance, what is very convenient within a nearest neighbor framework, as they show with their experimental results. Therefore, a similar idea is applied to gait recognition in this paper.

In [24], the problem of information fusion on multimodal problems is addressed by proposing a late fusion approach that seeks a shared rank-2 pairwise similarity matrix which is used to re-score confidence values for the problems of object categorization and video event detection. Although in this paper we will use a single modality (i.e. tracklet-based features), during the final identity decision step, the common approach is to directly take the label corresponding to the binary classifier that returned the highest score for the target sample. However, there are situations where the difference between some scores is very low, compared to the others, and a misclassification happens. Therefore, we evaluate in this paper the way the method of [24] can help to solve some ambiguities, boosting the final accuracy of the gait recognition system.

3. Methodology

In this section, we present the diverse proposals that we suggest to improve the accuracy and efficiency of a DTF-based gait recognition system (GRS). The GRS that we will consider contains the same stages used in [4]: (i) dense trajectories detection and description; *(ii)* people detection and tracking; *(iii)* video-level representation; *(iv)* feature vector compression; and *(v)* video classification. This pipeline is summarized in Fig. 1.

Firstly, we overview the 'Fisher Motion' descriptor of [4] that we will use as base. Then, the diverse improvements proposed in this paper are described.

3.1. Fisher Motion descriptor [4]

We summarize here the Fisher Motion (FM) approach proposed in [4].

Divergence-Curl-Shear (DCS) descriptor, for *dense trajectory features* (DTF) [21], was introduced in [13] for the problem of human action recognition. Afterwards, DCS was successfully applied to the problem of human gait recognition by Castro et al. [4]. As described in [13], the divergence is related to axial motion, expansion and scaling effects, whereas the curl is related to rotation in the image plane. The magnitude of the shear is computed from the hyperbolic terms as described in the original paper. Then, those kinematic features are combined in pairs to get the final motion descriptors. We refer the reader to [13] for further details.

A natural alternative to DCS descriptor would be the concatenation of histograms of oriented gradients (HOG), histograms of optical flow (HOF) and motion boundary histograms (MBH), as proposed in [21] for human action recognition. However, our own preliminary experiments on gait data indicated that DCS outperforms such alternative descriptor in most cases. Therefore, we decided to focus our experimental study on DCS.

3.1.1. Person-focused tracklets

To remove tracklets that were not generated by people motion, Castro et al. [4] localize people in the image sequences and discard those tracklets that do not pass through the person region. A similar idea is used in [22] in order to separate camera motion from people motion. Instead of using a gradient-based person detector, as in [4], we use in this paper background subtraction to delimit the pixels that cover the target person. For that purpose, we learn a Gaussian Mixture Model from F video frames (e.g. 40 frames). We use the implementation of [15] included in Matlab.

After segmentation, for each video frame, we fit a rectangle (i.e. bounding-box) to the foreground region. Then, the sequence of bounding-boxes is smoothed along time and possible gaps are filled by interpolation. As in [4], those bounding boxes allow to vertically split the person region into two halves to compute a FM descriptor