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] *Dept. 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*

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. ***Keywords:*** Gait recognition; Tracklets; DTF; Metric learning; Binarization

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 (Han and Bhanu, 2006; Fihl and Moeslund, 2009; Bashir et al., 2010). In contrast, (Castro et al., 2014) presented recently a fully automatic system for gait recognition based on dense *tracklets* (i.e. short-term point trajectories) (Wang et al., 2011; Jain et al., 2013; Wang and Schmid, 2013) 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

*Corresponding author: Tel.: +34-957218980; fax: +34-957218360
e-mail: mjmarin@uco.es (Manuel J. Marín-Jiménez)

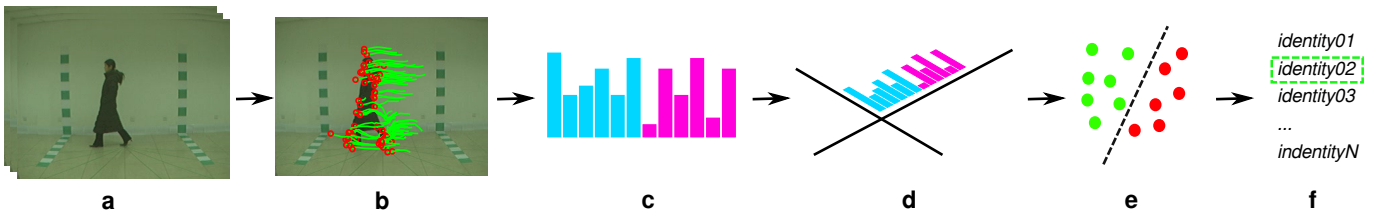

Fig. 1. Pipeline of the evaluated gait recognition system. (a) Input video. (b) Person-focused tracklets. (c) Pyramidal Fisher Motion descriptor. (d) Projected and compressed descriptor. (e) Classifier (e.g. SVM or NN). (f) An identity is selected from the classification scores.

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 (Yu et al., 2006), parts B and C, and TUM-GAID (Hofmann et al., 2014). In CASIA dataset, several challenging situations are evaluated, as carrying bags, wearing long coats, walking outside during night at different velocities, amongst 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 (Castro et al., 2014). 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 Sec. 3. The experimental results are presented and discussed in Sec. 4. And, finally Sec. 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 (Hu et al., 2004). One of the most successful approaches, proposed by (Han and Bhanu, 2006), is called 'Gait Energy 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 (Al-Maadeed et al., 2014). In contrast, tracklet-based methods have shown good robustness in diverse challenging situations (Castro et al., 2014) eliminating previous restrictions as fine-grained segmentation of people or 3D body reconstruction to be able to deal with curved trajectories (Iwashita et al., 2014).

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. (2012) 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 (Arandjelovic and Zisserman, 2012), it is proposed an improvement of SIFT descriptor for object retrieval, named 'Root-SIFT'. 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 (Castro et al., 2014).

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 where located in 'far' locations in the new space, whereas the gait descriptors of the same subject were projected in 'near' locations. In (Simonyan et al., 2013), 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 (Jégou et al., 2012), 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 (Ye et al., 2012), 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 missclassification happens. Therefore, we evaluate in this paper the way the method of (Ye et al., 2012) 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 (Castro et al., 2014): *(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 (Castro et al., 2014) that we will use as base. Then, the diverse improvements proposed in this paper are described.

3.1. Fisher Motion descriptor (Castro et al., 2014)

We summarize here the Fisher Motion (FM) approach proposed in (Castro et al., 2014).

Divergence-Curl-Shear (DCS) descriptor, for *dense trajectory features* (DTF) (Wang et al., 2011), was introduced in (Jain et al., 2013) for the problem of human action recognition. Afterwards, DCS was successfully applied to the problem of human gait recognition by (Castro et al., 2014). As described in (Jain et al., 2013), 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 (Jain et al., 2013) 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 (Wang et al., 2011) 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.

Person-focused tracklets. To remove tracklets that were not generated by people motion, (Castro et al., 2014) localize people in the image sequences and discard those tracklets that do not pass through the person region. A similar idea is used in (Wang and Schmid, 2013) in order to separate camera motion from people motion. Instead of using a gradient-based person detector, as in (Castro et al., 2014), 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 (KaewTraKulPong and Bowden, 2002) 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 (Castro et al., 2014), those bounding boxes allow to vertically split the person region into two halves to compute a FM descriptor per half. Therefore, our final gait descriptor will be the concatenation of both FMs.

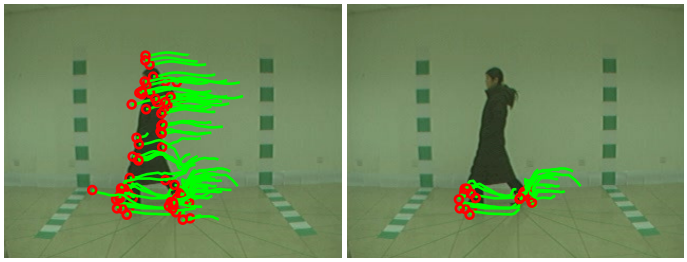

Fig. 2. Dense tracklets vs sparse tracklets. (a) Dense tracklets within the person region. (b) Sparse tracklets: only the 30% most discriminative tracklets are selected here. For this example, where the woman is wearing a long coat, the most representative tracklets are located around the lower legs.

Feature encoding. In order to build a video level descriptor from tracklet-based features, we use the approach of (Castro et al., 2014), the Fisher Motion (FM) descriptor. FM is based on the improved Fisher Vectors (FV) encoding (Perronnin et al., 2010). The FV encoding, that can be seen as an extension of the Bag of Words (BOW) representation (Sivic and Zisserman, 2003), builds on top of a Gaussian Mixture Model (GMM), where each Gaussian corresponds to a visual word. Whereas in BOW, an image is represented by the number of occurrences of each visual word, in FV an image is described by a gradient vector computed from a generative probabilistic model.

3.2. RootDCS

Inspired by the work of (Arandjelovic and Zisserman, 2012), where they propose *RootSIFT* for object retrieval, we propose *RootDCS*, which is an element wise square root of the $L1$-normalized DCS descriptors, for gait description. This is equivalent to use Hellinger distance for comparing DCS descriptors.

With this new regularization, the difference between larger bin values of the resulting kinematic features histogram with respect to the smaller bin values is reduced. This variance-stabilizing transformation improves the capability of representation of the computed descriptors.

3.3. Sparse tracklets

Dense tracklets are computed by defining a spatial grid at different scales as proposed by (Wang et al., 2011). However, not all the tracklets contribute in the same way to define the gait motion. We propose here a method to select a subset of the most discriminative tracklets in terms of gait motion. The key idea is to discard those tracklets whose mean velocity vectors are similar to the median velocity of the velocity field of the person, keeping in this way the most relevant ones, e.g. arm and leg swing. An example can be seen in Fig. 2.

We apply the procedure presented in Algorithm 1 to all the possible time intervals $\mathcal{T}$ of length L, where L is the selected length of tracklets (e.g. $L = 15$). Given the ordered list $\mathcal{L}$, different criteria can be applied to discard the most irrelevant tracklets. For example, keep just the top K tracklets; or, keep just the top percentage P of tracklets; or, discard all the tracklets whose energy E

is lower than a given threshold. During some preliminary experiments, we found that keeping a percentage P of the tracklets was satisfactory for most of our experiments. So, we will apply this criterion during our experiments (Sec. 4).

3.4. Metric learning and binarization for discriminative dimensionality reduction

Since a FM descriptor is usually a high-dimensional vector (e.g. 81408 dimensions with 128 Gaussians), and in order to be able to handle and store thousands of theses descriptors, a step of dimensionality reduction is needed. The most popular approach for unsupervised dimensionality reduction is Principal Components Analysis (PCA). However, since we are dealing with a problem where samples are labelled, we propose here the use of a semisupervised approach: metric learning (ML). Our goal is to learn a new feature space where samples of different classes were easily separable, w.r.t. the Euclidean distance. Therefore, we adopt here the approach of (Simonyan et al., 2013).

We learn a projection matrix W in such a way that samples of different subjects in the projected space have a large distance d_W, and samples of the same subject have a small distance d_W:

$$d_W^2(x, y) = \|Wx - Wy\|_2^2 \tag{1}$$

This is formulated as a non-convex objective function, optimized by following (Simonyan et al., 2013).

Although the descriptors obtained after ML are already *small*, we can reduce even more the storage needed by the gait descriptors, by applying a binarization stage. The goal of binarization is to map a low-dimensional real valued descriptor Γ to a binary code $B \in \{0, 1\}^q$ with the bit length q (where $q \geq m$). We follow the simple, but powerful, idea of (Jégou et al., 2012): compute a random $q \times q$ matrix R; apply QR-decomposition to matrix R to obtain an orthogonal matrix Q; and, keep the first m columns of Q to define the projection matrix U. To compress a target descriptor Γ, the following equation is applied: $B = \text{sign}(U\Gamma)$, where function $\text{sign}(v)$ returns 1 if $v > 0$ and 0 otherwise. This representation enables the comparison of gait descriptors by using the Hamming distance, which can be efficiently computed on state-of-the-art computers, in conjunction with a nearest neighbor classification framework.

Algorithm 1 Sparse tracklets

Input: $\mathcal{T}$ = time interval of length L; $\mathcal{S}$ = set of tracklets in $\mathcal{T}$

procedure SortTracklets($\mathcal{T}$, $\mathcal{S}$)
  for each t in $\mathcal{T}$ **do**
    for each $\mathbf{p}_i^t$ **do**
      $\mathbf{v}_i^t \leftarrow velocity(\mathbf{p}_i^t)$ ▷ Velocity of each point at t
    end for
    $\hat{\mathbf{v}}^t \leftarrow median(\{\mathbf{v}_i^t\})$ ▷ Median velocity at time t
    for each $\mathbf{v}_i^t$ **do**
      $\bar{\mathbf{v}}_i^t \leftarrow \mathbf{v}_i^t - \hat{\mathbf{v}}^t$
      $m_i^t \leftarrow |\bar{\mathbf{v}}_i^t|$ ▷ Magnitude of residual
    end for
  end for
  for each $\int_j$ in $\mathcal{S}$ **do**
    $E_j \leftarrow \frac{1}{L} \sum_{t=1}^{L} m_j^t$ ▷ *Energy* of each tracklet
  end for
  $\mathcal{L} \leftarrow sortdesc(\mathcal{S}, E)$ ▷ Sort tracklets given E
end procedure

Output: $\mathcal{L}$ = set of sorted tracklets.

In addition, as discussed in (Gordo et al., 2013), the binarization step implies a better spread of the vector energy (i.e. variance) along the dimensions, in contrast to vectors only PCA- or ML-compressed.

3.5. Rescoring classification scores

Since this is a multiclass problem, we use, as in (Castro et al., 2014), a *'one-vs-all'* ensemble of binary SVMs (Osuna et al., 1997) with linear kernels. Therefore, the identity label for a target test sample is given by the classifier that returned the highest classification score. Although the strategy of taking the maximum over the scores offers most of the time satisfactory results, it shows a clear limitation in cases where several classifiers return similar scores for the same target sample – it might happen that the third best score corresponded to the right identity.

To deal with those situations, very common in gait recognition, we propose the use of the *Rank Minimization* (RM) method of (Ye et al., 2012). RM is a *late fusion* method, originally proposed for the problems of object categorization and video event detection, whose aim is to combine the outputs of different classifiers to obtain a final and more robust decision on a set of test samples. Let's summarize the method below.

After obtaining classification scores from different models, usually trained on different features, we want to combine those scores in order to improve the classification capability of each individual model, obtaining a (hopefully) better score. Let $\mathbf{s} = [s_1, s_2, ..., s_m]$ be a confidence score vector of a model on m samples. A pairwise relationship matrix T is constructed from $\mathbf{s}$ as: $T_{jk} = \text{sign}(s_j - s_k)$

Given n models, the robust late fusion method of Ye et al. aims at optimizing the following problem:

$$\begin{aligned} &\min_{\hat{T},E_i} \|\hat{T}\|_* + \lambda \sum_{i=1}^{n} \|E_i\|_1, \\ &\text{s.t. } T_i = \hat{T} + E_i, i = 1, ..., n; \hat{T} = -\hat{T}^\top \end{aligned} \tag{2}$$

Where T_i is the pairwise relationship matrix of the i-th model, E_i is a sparse matrix associated to the i-th model, $\hat{T}$ is the estimated rank-2 pairwise relationship matrix consistent among the samples and models, and λ is a positive tradeoff parameter to be cross-validated. Such optimization problem is solved by inexact Augmented Lagrange Multiplier method. As described in (Ye et al., 2012), given the estimated matrix $\hat{T}$, and assuming that $\hat{T}$ is generated from $\hat{\mathbf{s}}$ as $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$, the new score vector $\hat{\mathbf{s}}$ is computed as

$$(1/m)\hat{T}\mathbf{e} = \arg\min_{\hat{\mathbf{s}}} \|\hat{T}^\top - (\hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top)\|_F^2, \tag{3}$$

treating $(1/m)\hat{T}\mathbf{e}$ as the recovered $\hat{\mathbf{s}}$ after the late fusion of the input scores.

So let's see how we adapt the previously explained method to our problem. In our *'one-vs-all'* ensemble of N binary SVMs, we have one classifier specialized in a single identity. Let us name it c^i, where i represents the i-th identity. During test time, from each binary classifier c^i, we obtain a vector $\mathbf{s}^i$ of m scores (one per test sample). So, we can compute a pairwise relationship

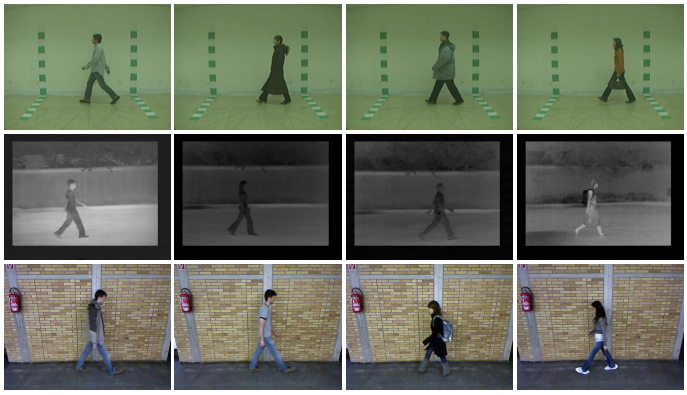

Fig. 3. Evaluated datasets. (top) CASIA-B. People recorded from the 90º camera walking indoors. Three situations are included: normal walking ('nm'), walking with a coat ('cl') and walking with a bag ('bg'). (middle) CASIA-C. People recorded outdoors during night with an infrared camera. Four situations are included: normal walking ('fn'), slow walking ('fs'), fast walking ('fq') and walking with a bag ('fb'). (bottom) TUM-GAID. People recorded from the same viewpoint walking indoors in two seasons. Three situations are included: normal walking ('N','TN'), walking with a bag ('B','TB') and walking with coating shoes ('S','TS').

matrix T^i from each $\mathbf{s}^i$. Then, we can obtain from T^i a new vector $\hat{\mathbf{s}}^i$ of identity-specialized scores, where the original scores have been re-scored taking into account the pairwise relationships represented in T^i. This process is repeated independently for each identity-specialized classifier i, thus, obtaining a new set of scores $\mathcal{S} = \{\hat{\mathbf{s}}^1, \hat{\mathbf{s}}^2, ..., \hat{\mathbf{s}}^N\}$, where N is the number of possible identities. To decide the final identity of each test sample, we seek the maximum over its scores in $\mathcal{S}$, assigning as identity the one of the identity-specialized classifier that generated such combined score.

4. Experimental results

The goal of the following experiments is to evaluate how the alternatives proposed above (Sec. 3) might improve the recognition performance of a GRS based on DTF, e.g. Pyramidal Fisher Motion. In particular, we will compare: (a) DCS vs. RootDCS, (b) FM vs. sparse-FM, (c) PCA vs. ML, and (d) low-dimensionality vectors vs. binarized vectors.

After describing the datasets where the experiments are conducted, the rest of this section presents the defined experiments and results.

4.1. Databases

We perform our experiments on "CASIA Gait Dataset", parts B (CASIA-B) (Yu et al., 2006), C (CASIA-C) (Tan et al., 2006), and TUM-GAID dataset (Hofmann et al., 2014). In CASIA-B 124 subjects perform walking trajectories in an indoor environment (top row of Fig. 3). The action is captured from 11 viewpoints. Three situations are considered: normal walk (*nm*), wearing a coat (*cl*), and carrying a bag (*bg*). For our experiments, and without loss of generality, we will focus on profile viewpoints, i.e., 90^o. In CASIA-C 153 subjects perform walking trajectories in an outdoor environment during night (middle row of Fig. 3). The action is captured from a single viewpoint with an infrared camera. Four situations are considered: normal walk (*fn*), fast walk (*fq*), slow walk (*fs*) and carrying a bag (*fb*). In both CASIA sets, video resolution is 320×240 pixels. In TUM GAID 305 subjects perform

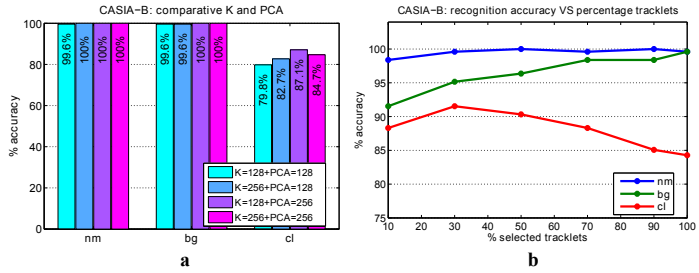

Fig. 4. Effect of selected parameters on the identification accuracy for CASIA-B. (a) Effect of dictionary size (K) and final descriptor length (PCA compression). (b) Effect of the percentage of selected tracklets (sparse) on the recognition accuracy. Each line corresponds to a different scenario. (Best viewed in digital format)

two walking trajectories in an indoor environment. The first trajectory is performed from left to right and the second one from right to left. In addition, for some subjects, two recording sessions were performed: firstly in January, where subjects wear heavy jackets and mostly winter boots, and secondly in April, where subjects wear different clothes. This subset is known as the *elapsed-time* scenario. Some examples can be seen in bottom row of Fig. 3. Hereinafter the following nomenclature is used to refer every of the four walking conditions considered in the dataset: normal walk (*N*), carrying a backpack (*B*), wearing coating shoes (*S*) and elapsed time (*TN-TB-TS*). Video is are recorded at a resolution of 640×480 pixels with a frame rate of approximately 30 fps.

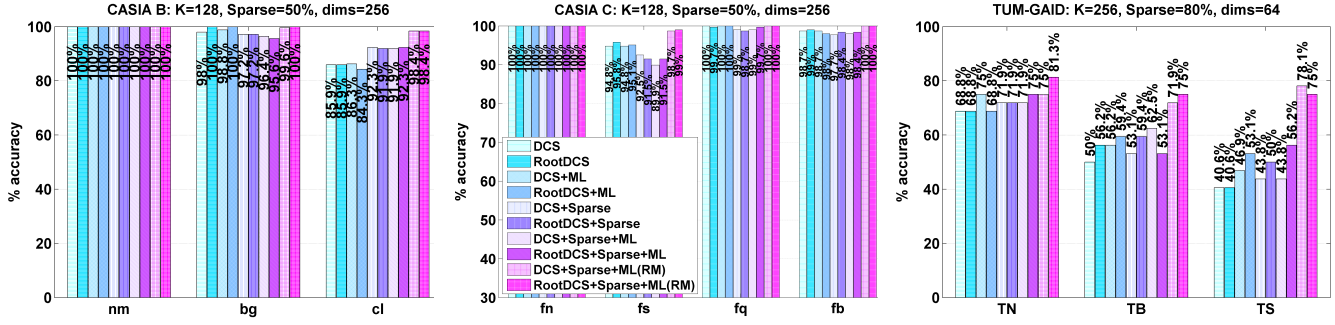

Fig. 5. Subject identification: comparison of combined techniques. Percentage of correct identification is represented. Bars are grouped per scenario. (left) Results on CASIA-B. (middle) Results on CASIA-C. (right) Results on TUM-GAID (elapsed-time). (Best viewed in digital format)

4.2. Identification task

Given a probe (test) sequence, the goal of *identification* is to assign an identity (label) to the subject performing the action, from a set of predefined identities (gallery). For this task, we train a set of N linearSVM (Osuna et al., 1997) in a *one-vs-all* fashion, where N is the number of possible identities. In some early experiments, we tried SVM with χ^2 and RBF kernels, but they did not show any improvement. For this task, we report the percentage of correct recognition: $Acc = 100 \cdot \frac{\#hits}{\#seqs}$; where $\#hits$ is the amount of correctly identified probe sequences, and $\#seqs$ is the total number of probe sequences.

In this set of experiments, firstly, we will study the effect of selecting different dictionary sizes and output dimensionality, and percentages of tracklets against using dense tracklets (Fig. 4). Then, we will compare how the combination of the different proposals contribute to the task of person identification (Fig. 5). And, finally, we will focus on the binarization of gait descriptors for people identification using a nearest neighbor (NN) strategy (Fig. 6). To allow a direct comparison with the state-of-the-art results, the

evaluation protocol used in these experiments follows the ones proposed by the authors of considered the datasets. For CASIA-B, always training on trajectories $\{1, 2, 3, 4\}$ of scenario 'nm', and testing on trajectories $\{5, 6\}$ of 'nm', $\{1, 2\}$ of 'bg' and $\{1, 2\}$ of 'cl'; and, for CASIA-C, always training on trajectories $\{0, 1\}$ of scenario 'fn', and testing on trajectories $\{2, 3\}$ of scenario 'fn', and trajectories $\{0, 1\}$ of the remaining scenarios ('fq', 'fs', 'fb'). In (Hofmann et al., 2014), the database is split into three standard partitions: 100 subjects for training and building models, 50 subjects for validation, and 155 subjects for testing. Since on the test partition of the regular scenarios we obtain almost perfect identification results[1], we focus on the *elapsed-time* scenarios. In such experiment, 32 subjects (16 for training and 16 for testing) are recorded in two different sessions (January and April), therefore, they wear different kind of clothes. This situation is extremely challenging due to severe changes in clothing and a very reduced amount of gallery samples – we will see later that published state-of-the-art results are quite low for this setup.

Identification results. We study in Fig. 4 the effect of using different parameters during descriptor computation for person identification on CASIA-B. In Fig. 4.a, we compare a selection of the four most representative combinations of values for dictionary size K (see Sec. 3.1), and the final descriptor size based on PCA. Focusing on scenario 'cl', we can see that a larger final dimensionality (256 vs 128) improve the accuracy. In contrast, a larger dictionary size does not yield better results, probably due to overfitting. Similar experiments on TUM-GAID dataset (elapsed-time scenario) have showed that good values for K and final descriptor size are 256 and 64, respectively. Fig. 4.b shows a comparison of the effect of the percentage of selected tracklets, where the first represented accuracy (left) corresponds to keeping just the 10% of the tracklets, and the last accuracy (right) corresponds to not removing any tracklet. For this experiment, the final PFM descriptor has been compressed to 128 dimensions with PCA. On the one hand, we can see that scenario 'nm' is almost not affected by the reduction of tracklets, showing a stationary behavior. On the other hand, both 'bg' and 'cl' are affected by the use of sparse tracklets. In the former case, the more tracklets used the better is the accuracy. However, in the latter case, a maximum is reached when using only a 30% of the tracklets. That means that a sparse representation of the motion is preferred when trying to recognize gait of people who wear winter clothing, disguising their identity.

The results of the experiment combining diverse strategies are summarized in Fig. 5. Left and middle plots correspond to results on CASIA-B and CASIA-C, respectively, whereas right plot corresponds to TUM-GAID. Within each plot, bars are grouped per database scenarios. The particular configuration of each gait descriptor is indicated in the title of each plot: K is the dictionary size used for FV, *sparse* indicates the percentage of tracklets used and *dims* refers to the final PFM size after dimensionality reduction. For CASIA, the figure contains results for the configuration PFM-256 and sparse 50%, and for TUM-GAID, PFM-64 and sparse 80%. Since, in general, it is easier to analyse the behaviour of the proposed improvements on the most difficult scenarios of each dataset, we will focus our discussion on them. For the case of CASIA-B, if we focus on scenario 'cl', we can see that starting

[1] On the test partition of TUM-GAID, we obtain: $N = 99.7\%, B = 99\%, S = 99\%$; with parameters: 150-dims DCS, $K = 600$ and PFM-256.

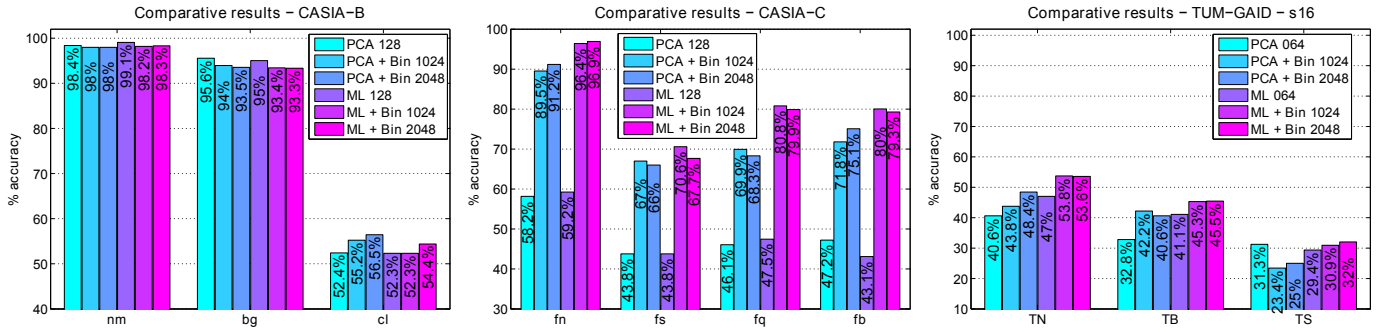

Fig. 6. NN-based subject identification: comparative results on CASIA-B, CASIA-C and TUM-GAID. A nearest-neighbor approach is used in this experiment: the identity of the nearest prototype is assigned. Percentage of correct identification is represented. Bars are grouped per scenario. (left) Results on CASIA-B. (middle) Results on CASIA-C. (right) Results on TUM-GAID (elapsed-time). (Best viewed in digital format)

from the baseline value 85.9%, by applying the different proposals, we reach an accuracy value of 98.4%. What means, that on the challenging scenario where people change 'normal' clothing by winter clothing (i.e. thickening the body and covering large part of the legs) the final system is able to better handle the most discriminative gait features to identify the individuals. Note that one of the biggest improvements is provided by the use of sparse tracklets. For the case of CASIA-C, if we focus on scenario 'fs', we can see that a recognition accuracy of 99% can be obtained starting from the baseline value 94.8%. Note that the only clear improvement is brought by the use of RM for rescoring the classification outputs. In contrast, using sparse tracklets (i.e. 50%) decreases the recognition performance for the evaluated values, what means that the identification of people who have reduced their walking speed requires the use of dense tracklets. Note that the other scenarios are not negatively affected when using only half of the tracklets, indicating that a sparse representation can be safely used in those cases. For the elapsed-time scenarios of TUM-GAID, we can see that RootDCS generally brings improvement for 'TB' and 'TS', whereas using sparse tracklets (i.e. 80%) does not boost the accuracy as discussed previously. However, both metric-learning and RM rescoring systematically improve the identification accuracy.

To put our results in context, (Bashir et al., 2010) report on CASIA-B an average 74.1% of accuracy over the three scenarios (i.e. 'nm' 100%, 'bg' 78.3%, 'cl' 44%). In our case, we obtain an average of 99.5% (i.e. 'nm' 100%, 'bg' 100%, 'cl' 98.4%) with the configuration 'RootDCS+Sparse+ML+RM' (Fig. 5), clearly surpassing previous results. For the case of CASIA-C, (Guan and Li, 2013) obtain an average accuracy of 98.9% over the four scenarios, whereas we obtain 99.8% with the configuration 'RootDCS+Sparse+ML+RM'. In (Whytock et al., 2014), the average accuracy reported on the elapsed-time scenarios of TUM-GAID is 49% (i.e. 'TN' 65.6%, 'TB' 31.3%, 'TS' 50%), in contrast to the 77.1% that we obtain with the configuration 'RootDCS+Sparse+ML+RM', establishing a new state-of-the-art result on this dataset.

Instead of using an ensemble of binary SVMs, Fig. 6 shows the recognition accuracy of a subject identification system where a NN approach is used to assign an identity to the probe samples given the gallery. Euclidean distance is used to compare real-valued vectors, whereas Hamming distance is used for binarized vectors. In this experiment, 10 random splits (obtained from the original

training sets) are used to learn both PCA and ML projection matrices. Therefore, the reported recognition accuracy is an average on the 10 results obtained on the whole test set with each model. Subfigures corresponding to CASIA B and C contain results for gait descriptors compressed to 128 dimensions, in contrast to left subfigure (i.e. TUM-GAID) where 64 dimensions are used. In all cases, RootDCS on dense tracklets is used. We can see that results obtained by ML-based descriptors are always better than PCA-based for the 'normal' scenarios, where the projection matrix is learnt. On the other hand, the binarization strategy boosts the recognition accuracy in CASIA-C, using just 1024 bits – using double number of bits (i.e. 2048) does not always increase the accuracy. Such improvement also happens in the case of TUM-GAID for 1024 and 2048 bits. If we compare the results shown in Figs. 5 and 6, we can observe that the results obtained with NN on binarized data for 'normal' scenarios of CASIA-B and CASIA-C are comparable to the ones obtained with SVMs. However, NN-based identification does not require a training step, in contrast to the training needed by SVM (i.e. training as much SVM classifiers as target identities), making easier to include a new target identity to the system. Therefore, the choice of either SVM or NN will be determined by the need of high recognition accuracy in extremely uncontrolled scenarios or the easiness for updating the model gallery in fairly controlled scenarios.

4.3. Verification task

Given a pair of gait sequences, the goal of *verification* is to decide whether both sequences belong to the same subject or not. As done in Sec. 4.2, we learn the model on the training trajectories of the 'normal' scenarios and test on the others.

For evaluation purposes, the target training set is split into two equally sized subsets. One will be used for training the metric-based model, and the other subset, for validation of the model during training. So we can cross-validate the learning parameters. This procedure is repeated ten times, where for each repetition, a random set of pairs is generated. We report average values on the ten repetitions.

We define the following evaluation protocol. Given the test set, we define random pairs of positive pairs (same subject) and negative pairs (different subjects), where one component of the evaluated pair comes from the set of prototypes (i.e. gallery set) and the other component comes from the probe set. Positive pairs will be given a $+1$ label, and negative pairs will be given a -1 label. The expected behavior of the evaluated method is to assign a high score to positive pairs and a low score to negative pairs. In terms of retrieval, positive pairs should be ranked higher than negative pairs. The amount of negative pairs included for evaluation is ten times the amount of positive pairs generated. Since we are defining retrieval-like experiments, we use the area under the precision-recall curve (AUC) as metric for evaluating the different experimental setups.

In the following set of experiments, we aim at evaluating the effect of dimensionality reduction of the FM descriptors for the task of verification. We compare two techniques for dimensionality reduction: PCA and ML (Sec.3.4).

For a given pair of subjects with projected descriptors $\mathbf{t}$ (test individual) and $\mathbf{p}$ (prototype), their similarity score s is computed

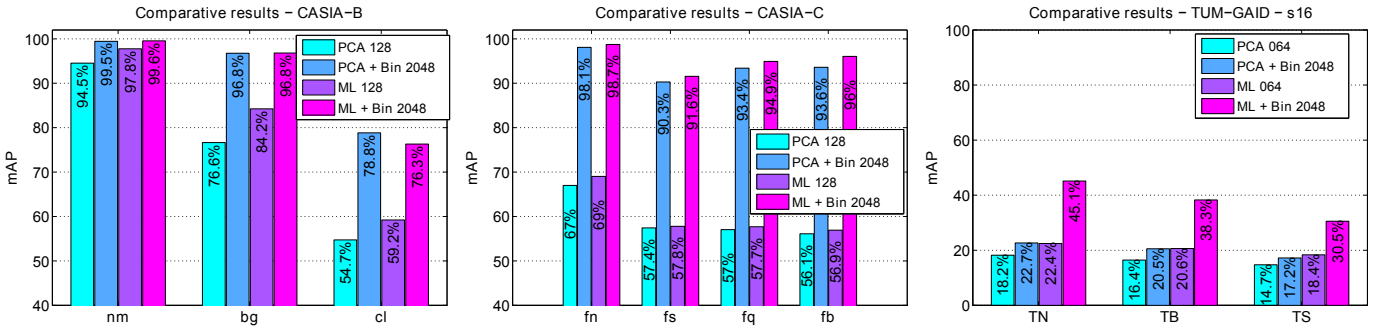

Fig. 7. Subject verification: comparative results on CASIA-B, CASIA-C and TUM-GAID. Mean average precision (mAP) is represented. Bars are grouped per scenario. (left) Results on CASIA-B. (middle) Results on CASIA-C. (right) Results on TUM-GAID (elapsed-time). (Best viewed in digital format)

as: $s(\mathbf{t}, \mathbf{p}) = b - \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{p}_i)^2$, where $b > 0$ is a bias value estimated on the training set, and N is the dimensionality of the vectors. In case we convert real-valued vectors to binary vectors, for a given pair of subjects with binary descriptors $\mathbf{t}$ (test) and $\mathbf{p}$ (prototype), their similarity score s is computed as: $s(\mathbf{t}, \mathbf{p}) = N - \sum_{i=1}^{N} \mathrm{XOR}(\mathbf{t}_i, \mathbf{p}_i)$, where N is the dimensionality of the vectors and $\mathrm{XOR}(\cdot, \cdot)$ applies the binary XOR operation.

For these experiments, we will use PFM descriptors (on dense RootDCS features) that will be compressed to 128 dimensions for CASIA and 64 dimensions for TUM-GAID, and binarized to 2048 bits.

Verification results. Fig. 7 shows that the metric learning (ML) approach described in Sec. 3.4 clearly improves on all the cases tested on CASIA-B, whereas the improvement on the scenarios of CASIA-C and TUM-GAID is moderated. Focusing on the binarization stage, we can see that using binary descriptors brings a consistent improvement across the different datasets and scenarios, what highlights the potentiality of using this strategy for the given task. Note that the binarized version of ML-based descriptors obtain a better mAP than the PCA-based ones in all the tested cases, indicating the utility of using a metric learning approach in this context. The improvement obtained by the binarization stage in our data is supported by the observations of (Gordo et al., 2013) for binarizing document image descriptors. Firstly, the PCA- and ML-compression applied to the descriptors helps to uncorrelate the dimensions; then, the use of the matrix obtained from the QR decomposition rotates the data which, in turn, spreads the energy across the dimensions, obtaining a better representation of the samples and, therefore, a larger identification/verification performance.

4.4. Final remarks

Based on the previous results, we analyse here what are the techniques that should be chosen to improve the accuracy of a tracklet-based gait recognition system. Firstly, we will analyse each one in an isolated way (i.e. not combined with other of the proposals). Both *metric learning* and *RM rescoring* improve always the final identification accuracy, regardless of the scenario, and *binarization* enhances the gait representation for both NN-based identification and validation. In contrast, *RootDCS* and *sparse tracklets* boost the results only in some situations: the former does not present a clear pattern where it can be defined as the right

choice, but the latter is useful when trying to recognize people wearing long coats and bags. Then, if we had to choose the best combination of techniques, in most situations the choice of *metric learning* plus *rescoring* is a good decision, adding a *binarization* stage if we want to use a simple NN classifier. The use of *sparse tracklets* will depend on whether we knew in advance that we are dealing with subjects heavily disguised with clothing, where filtering out tracklets has shown to be beneficial. Since there is not a clear pattern where *RootDCS* should be applied, it should be tested for each particular system.

5. Conclusions

This paper has proposed and evaluated diverse strategies to improve tracklet-based gait recognition systems. The experimental results carried out on CASIA-B, CASIA-C and TUM-GAID have shown that: (a) RootDCS features are in general a good choice to build gait descriptors, although its use will depend on the particular system; (b) for specific situations, as people wearing long coats, using sparse tracklets improves the identification accuracy; (c) compressing gait descriptors of sparse tracklets with a discriminatively trained projection matrix increase the recognition performance; (d) binarization of gait descriptors not only reduces storage needs and computational complexity, but also improves the accuracy for NN-based identification and verification systems; and, (e) the use of a final rescoring stage (i.e. Rank Minimization) on the set of probe samples benefits the final recognition accuracy of the system.

Acknowledgments

This work has been partially funded by the Research Projects TIN2012-32952 (Spanish Ministry of Science and Technology) and TIC-1692 (Junta de Andalucía).

References

Al-Maadeed, S., Almotaeryi, R., Jiang, R., Bouridane, A., 2014. Robust human silhouette extraction with laplacian fitting. Patt. Recog. Letters 49, 69 – 76.

Arandjelovic, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2911–2918.

Bashir, K., Xiang, T., Gong, S., 2010. Gait recognition without subject cooperation. Pattern Recognition Letters 31, 2052 – 2060.

Castro, F.M., Marín-Jiménez, M., Medina-Carnicer, R., 2014. Pyramidal Fisher Motion for multiview gait recognition, in: Proceedings of the International Conference on Pattern Recognition, pp. 1692–1697.

Chakraborty, B., Holte, M., Moeslund, T., Gonzalez, J., 2012. Selective spatio-temporal interest points. Computer Vision and Image Understanding 116, 396 – 410.

Fihl, P., Moeslund, T., 2009. Invariant gait continuum based on the duty-factor. Signal, Image and Video Processing 3, 391–402.

Gordo, A., Perronnin, F., Valveny, E., 2013. Large-scale document image retrieval and classification with runlength histograms and binary embeddings. Pattern Recognition 46, 1898–1905.

Guan, Y., Li, C., 2013. A robust speed-invariant gait recognition system for walker and runner identification, in: Biometrics (ICB), 2013 International Conference on, pp. 1–8.

Han, J., Bhanu, B., 2006. Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 316–322.

Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G., 2014. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. Journal of Visual Communication and Image Representation 25, 195 – 206. Visual Understanding and Applications with RGB-D Cameras.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34, 334–352.

Iwashita, Y., Ogawara, K., Kurazume, R., 2014. Identification of people walking along curved trajectories. Pattern Recognition Letters 48, 60 – 69.

Jain, M., Jegou, H., Bouthemy, P., 2013. Better exploiting motion for better action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2555–2562.

Jégou, H., Furon, T., Fuchs, J.J., 2012. Anti-sparse coding for approximate nearest neighbor search, in: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2029–2032.

KaewTraKulPong, P., Bowden, R., 2002. An improved adaptive background mixture model for real-time tracking with shadow detection, in: Video-Based Surveillance Systems. Springer, pp. 135–144.

Osuna, E., Freund, R., Girosi, F., 1997. Support Vector Machines: training and applications. Technical Report AI-Memo 1602. MIT.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 143–156.

Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A., 2013. Fisher Vector Faces in the Wild, in: Proc. of the British Machine Vision Conference.

Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos, in: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1470–1477.

Tan, D., Huang, K., Yu, S., Tan, T., 2006. Efficient night gait recognition based on template matching, in: Proceedings of the International Conference on Pattern Recognition, pp. 1000–1003.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2011. Action Recognition by Dense Trajectories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176.

Wang, H., Schmid, C., 2013. Action recognition with improved trajectories, in: Proceedings of the International Conference on Computer Vision (ICCV), pp. 3551–3558.

Whytock, T., Belyaev, A., Robertson, N., 2014. Dynamic distance-based shape features for gait recognition. Journal of Mathematical Imaging and Vision 50, 314–326.

Ye, G., Liu, D., Jhuo, I.H., Chang, S.F., 2012. Robust late fusion with rank minimization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3021–3028.

Yu, S., Tan, D., Tan, T., 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: Proceedings of the International Conference on Pattern Recognition, pp. 441–444.