Pyramidal Fisher Motion for Multiview Gait Recognition

Francisco M. Castro
IMIBIC – Dep. of Computing
and Numerical Analysis
University of Cordoba
Cordoba, Spain
Email: i92capaf@uco.es

Manuel J. Marín-Jiménez
IMIBIC – Dep. of Computing
and Numerical Analysis
University of Cordoba
Cordoba, Spain
Email: mjmarin@uco.es

Rafael Medina-Carnicer
IMIBIC – Dep. of Computing
and Numerical Analysis
University of Cordoba
Cordoba, Spain
Email: rmedina@uco.es

***Abstract*—The goal of this paper is to identify individuals by analyzing their gait. Instead of using binary silhouettes as input data (as done in many previous works) we propose and evaluate the use of motion descriptors based on densely sampled short-term trajectories. We take advantage of state-of-the-art people detectors to define custom spatial configurations of the descriptors around the target person. Thus, obtaining a pyramidal representation of the gait motion. The local motion features (described by the Divergence-Curl-Shear descriptor [1]) extracted on the different spatial areas of the person are combined into a single high-level gait descriptor by using the Fisher Vector encoding [2]. The proposed approach, coined *Pyramidal Fisher Motion*, is experimentally validated on the recent 'AVA Multiview Gait' dataset [3]. The results show that this new approach achieves promising results in the problem of gait recognition.**

I. Introduction

The term *gait* refers to the way each person walks. Actually, humans are good recognizing people at a distance thanks to their gait [4], what provides a good (non invasive) way to identify people without requiring their cooperation, in contrast to other biometric approaches as iris or fingerprint analysis. One potential application of gait recognition is video surveillance, where it is crucial to identify dangerous people without their cooperation. Although great effort has been put into this problem in recent years [5], it is still far from solved.

Popular approaches for gait recognition require the computation of the binary silhouettes of people [6], usually, by applying some background segmentation technique. However, this is a clear limitation in presence of dynamic backgrounds and/or non static cameras, where noisy segmentations are obtained. To deal with these limitations, we propose the use of descriptors based on the local motion of points. These kind of descriptors have become recently popular in the field of human action recognition [7]. The main idea is to build local motion descriptors from densely sampled points. Then, these local descriptors are aggregated into higher level descriptors by using histogram-based techniques (e.g. Bag of Words [8]).

Therefore, our research question is: *could we identify people by using only local motion features as represented in Fig. 1?* We represent in Fig. 1 the local trajectories of image points belonging to four different people. Our goal is to use each set of local trajectories to build a high-level descriptor that allows to identify individuals. In this paper we introduce a new gait descriptor, coined *Pyramidal Fisher Vector*, that combines the potential of recent human action recognition descriptors with the rich representation provided by Fisher Vectors encoding [2]. A thorough experimental evaluation is carried out on the recent 'AVA Multiview Gait' dataset showing that our proposal contributes to the challenging problem of gait recognition by using a modern computational approach.

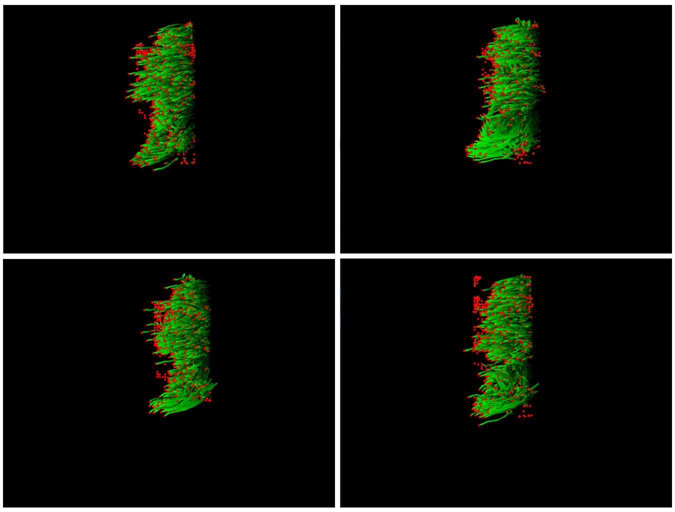

Fig. 1. **Who are they?** The goal of this work is to identify people by using their gait. We build the Pyramidal Fisher Motion descriptor from trajectories of points. We represent here the gait motion of four different subjects.

This paper is organized as follows. After presenting the related work, we describe our proposed framework for gait recognition in Sec. II. Sec. III is devoted to the experimental results. And, finally, the conclusions and future work are presented in Sec. IV.

A. Related work

Many research papers have been published in recent years tackling the problem of human gait recognition. For example, in [5] we can find a survey on this problem summarizing some of the most popular approaches. Some of them use explicit geometrical models of human bodies, whereas others use only image features. A sequence of binary silhouettes of the body is adopted in many works as input data. In this sense, the most popular silhouette-based gait descriptor is the called Gait

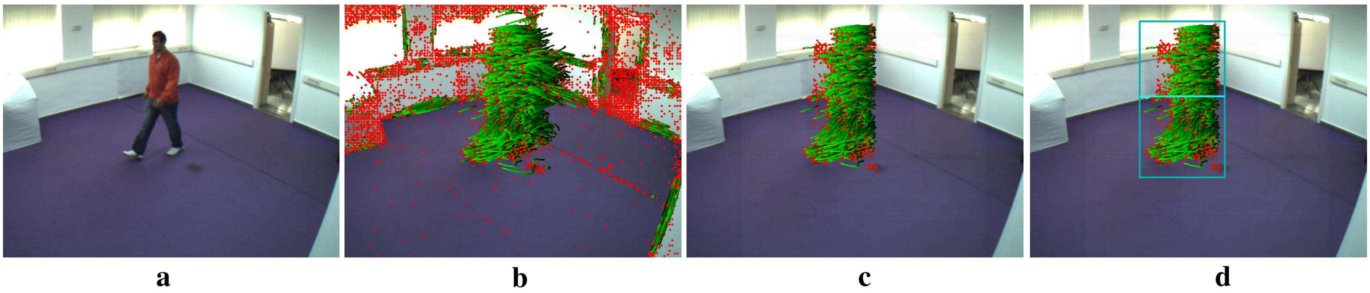

Fig. 2. **Pipeline for gait recognition.** a) The input is a sequence of video frames. b) Densely sampled points are tracked. c) People detection helps to remove trajectories not related to gait. d) A spatial grid is defined on the person bounding-box, so features are spatially grouped to compute a descriptor per cell. Then, those descriptors are concatenated into a single descriptor.

Enery Image (GEI) [9]. The key idea is to compute a temporal averaging of the binary silhouette of the target subject. Liu et al. [10], to improve the gait recognition performance, propose the computation of HOG descriptors from popular gait descriptors as the GEI and the Chrono-Gait Image (CGI). In [11], the authors try to find the minimum number of gait cycles needed to carry out a successful recognition by using the GEI descriptor. Martin-Felez and Xiang [6], using GEI as the basic gait descriptor, propose a new ranking model for gait recognition. This new formulation of the problem allows to leverage training data from different datasets, thus, improving the recognition performance. In [12], Akae et al. propose a temporal super resolution approach to deal with low frame-rate videos for gait recognition. They achieve impressive results by using binary silhouettes of people at a rate of 1-fps.

On the other hand, human action recognition (HAR) is related to gait recognition in the sense that the former also focuses on human motion, but tries to categorize such motion into categories of actions as *walking, jumping, boxing*, etc. In HAR, the work of Wang et al. [7] is a key reference. They introduce the use of short-term trajectories of densely sampled points for describing human actions, obtaining state-of-the-art results in the HAR problem. The dense trajectories are described with the Motion Boundary Histogram. Then, they describe the video sequence by using the Bag of Words (BOW) model [8]. Finally, they use a non-linear SVM with χ^2-kernel for classification. In parallel, Perronnin and Dance [13] introduced a new way of histogram-based encoding for sets of local descriptors for image categorization: the Fisher Vector (FV) encoding. In FV, instead of just counting the number of occurrences of a visual word (i.e. quantized local descriptor) as in BOW, the concatenation of gradient vectors of a Gaussian Mixture is used. Thus, obtaining a larger but richer representation of the image.

Borrowing ideas from the HAR and the image categorization communities, we propose in this paper a new approach for gait recognition that combines low-level motion descriptors, extracted from short-term point trajectories, with a multi-level gait encoding based on Fisher Vectors: the *Pyramidal Fisher Motion* (PFM) gait descriptor. We have discovered at submission time of this paper a very recent publication that shares some of our ideas, the work of Gong et al. [14]. It is similar to ours in the sense that they propose a method that uses dense local spatio-temporal features and a Fisher-based representation rearranged as tensors. However, there are some significant differences: *(i)* instead of using all the local features available in the sequence, we use a person detector to focus only on the ones related to the target subject; *(ii)* the information provided by the person detector enables a richer representation by including coarse geometrical information through a spatial grid defined on the person bounding-box; and, *(iii)* instead of dealing with a single camera viewpoint, we integrate in our system several camera viewpoints.

II. Proposed framework

In this section we present our proposed framework to address the problem of gait recognition. Fig. 2 summarizes the pipeline of our approach. We start by computing local motion descriptors from tracklets of densely sampled points on the whole scene (Fig. 2.b – Sec. II-A). Since, we do not assume a static background, we run a person detector to remove the point trajectories that are not related to people (Fig. 2.c –Sec. II-B). In addition, we spatially divide the person regions to aggregate the local motion descriptors into mid-level descriptors (Fig. 2.d –Sec. II-C). Finally, a discriminative classifier is used to identify the subjects (Sec. II-D).

A. *Motion-based features*

The first step of our pipeline is to compute densely sampled trajectories. Those trajectories are computed by following the approach of Wang et al. [7]. Firstly, dense optical flow $F = (u_t, v_t)$ is computed [15] on a dense grid (i.e. step size of 5 pixels and over 8 scales). Then, each point $p_t = (x_t, y_t)$ at frame t is tracked to the next frame by median filtering as follows:

$$p_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * F)|_{(\bar{x_t}, \bar{y_t})} \quad (1)$$

where M is the kernel of median filtering and $(\bar{x_t}, \bar{y_t})$ is the rounded position of p_t. To minimize drifting effect, the tracking is limited to L frames. We use $L = 15$ as in [1]. As a postprocessing step, noisy and uninformative trajectories (e.g. excessively short or showing sudden large displacements) are removed.

Once the local trajectories are computed, they are described with the Divergence-Curl-Shear (DCS) descriptor proposed by

Jain et al. [1], which is computed as follows:

$$\begin{cases} \mathrm{div}(p_t) = \dfrac{\partial u(p_t)}{\partial x} + \dfrac{\partial v(p_t)}{\partial y} \\ \mathrm{curl}(p_t) = \dfrac{-\partial u(p_t)}{\partial y} + \dfrac{\partial v(p_t)}{\partial x} \\ \mathrm{hyp}_1(p_t) = \dfrac{\partial u(p_t)}{\partial x} - \dfrac{\partial v(p_t)}{\partial y} \\ \mathrm{hyp}_2(p_t) = \dfrac{\partial u(p_t)}{\partial y} + \dfrac{\partial v(p_t)}{\partial x} \end{cases} \tag{2}$$

As described in [1], the divergence is related to axial motion, expansion and scaling effects, whereas the curl is related to rotation in the image plane. From the hyperbolic terms ($\mathrm{hyp}_1, \mathrm{hyp}_2$), we can compute the magnitude of the shear as:

$$\mathrm{shear}(p_t) = \sqrt{\mathrm{hyp}_1^2(p_t) + \mathrm{hyp}_2^2(p_t)} \tag{3}$$

B. *People detection and tracking*

We follow a tracking-by-detection strategy as in [16]: we detect full bodies with the detection framework of Felzenszwalb et al. [17]; and, then, we apply the clique partitioning algorithm of Ferrari et al. [18] to group detections into tracks. Short tracks with low-scored detections are considered as false positives and are discarded for further processing. In addition, to remove false positives generated by static objects, we measure the displacement of the detection along the sequence. Thus, discarding those tracks showing a static behaviour.

The tracks finally kept are used to filter out the trajectories that are not related to people: we only keep the trajectories that pass through, at least, one bounding-box of any track. In this way, we can focus on the trajectories that should contain information about the gait.

C. *Pyramidal Fisher Motion*

Fisher Vector encoding. As described above, our low-level features are based on motion properties extracted from person-related local trajectories. In order to build a person-level gait descriptor, we need to summarize the local features. We propose here the use of Fisher Vectors (FV) encoding [2].

The FV, that can be seen as an extension of the Bag of Words (BOW) representation [8], 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.

Assuming that our local motion descriptors $\{x_t \in R^D, t = 1 \ldots T\}$ of a video V are generated independently by a GMM $p(x|\lambda)$ with parameters $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \ldots N\}$, we can represent V by the following gradient vector [13]:

$$G_\lambda(V) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log p(x_t|\lambda) \tag{4}$$

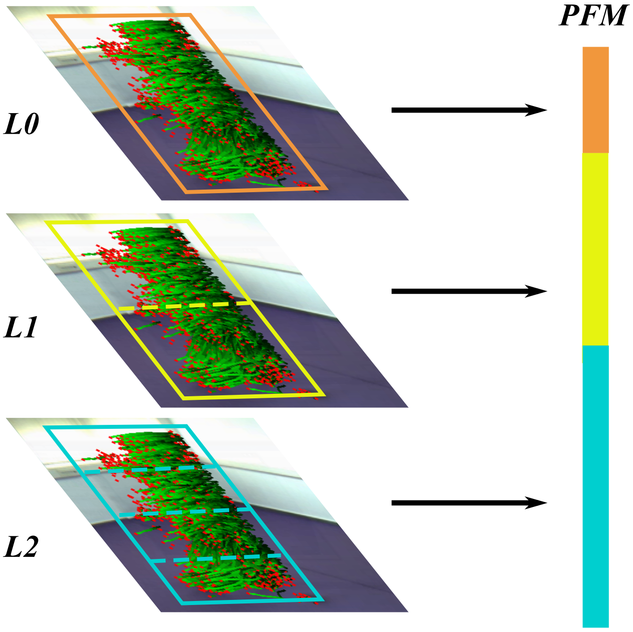

Fig. 3. **Pyramidal Fisher Motion descriptor.** Fisher Vector encoding is used at each spatial region on the Divergence-Curl-Shear descriptors computed on the dense trajectories enclosed in the person bounding-box. All Fisher Vectors are concatenated to obtain the final PFM gait descriptor.

Following the proposal of [2], to compare two videos V and W, a natural kernel on these gradients is the Fisher Kernel: $K(V, W) = G_\lambda(V)^T F_\lambda^{-1} G_\lambda(W)$, where F_λ is the Fisher Information Matrix [19].

As F_λ is symmetric and positive definite, it has a Cholesky decomposition $F_\lambda^{-1} = L_\lambda^T L_\lambda$, and $K(V, W)$ can be rewritten as a dot-product between normalized vectors Γ_λ with: $\Gamma_\lambda(V) = L_\lambda G_\lambda(V)$. Then, $\Gamma_\lambda(V)$ is known as the Fisher Vector of video V. As stated in [2], the capability of description of the FV can be improved by applying it a signed square-root followed by L2 normalization. So, we adopt this finding for our descriptor.

The dimensionality of FV is $2ND$, where N is the number of Gaussians in the GMM, and D is the dimensionality of the local motion descriptors x_t. For example, in our case, the dimensionality of the local motion descriptors is $D = 318$, if we use $N = 100$ Gaussians, then, the FV would have 63600 dimensions. In this paper, we will use the term *Fisher Motion* (FM) to refer to the FV computed on a video from low-level motion features.

Pyramidal representation. We borrow from [20] the idea of building a pyramidal representation of the gait motion. Since each bounding-box covers the whole body of a single person, we propose to spatially divide the BB into cells. Then, a Fisher vector is computed inside each cell of the spatio-temporal grid. We can build a pyramidal representation by combining different grid configurations. Then, the final feature vector, used to represent a time interval, is computed as the concatenation of the cell-level Fisher vectors from all the levels of the pyramid. This idea is represented in Fig.3, where each colored segment of the PFM descriptor comes from a different level of the pyramid.

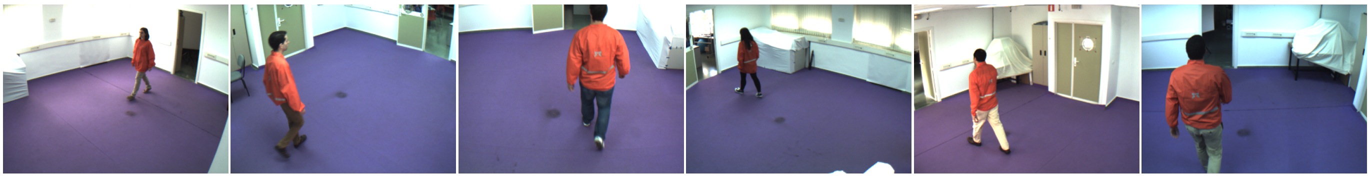

Fig. 4. **AVAMVG dataset.** Different people recorded from six camera viewpoints. The dataset contains both female and male subjects performing different trajectories through the indoor scenario. Note that cameras 3rd and 6th (from left to right) are prone to show people partially occluded.

D. Classification

The last stage of our pipeline is to train a discriminative classifier to distinguish between the different human gaits. Since, this is a multiclass problem, we train P binary linear Support Vector Machines (SVM) [21] (as many as different people) in a *one-vs-all* strategy. Although the χ^2 kernel [22] is a popular choice for BOW-based descriptors, a linear kernel is typically enough for FV, due to the rich feature representation that it provides.

E. Implementation details

For people detection, we use the code published by the authors of [17]. For computing the local motion features, we use the code published by the authors of [1]. The Fisher Vector encoding and the classification is carried out by using the code included in the library VLFeat [1].

III. Experimental Results

We carry out diverse experiments in order to validate our approach. With these experiments we try to answer the following questions: a) *is the combination of trajectory-based features with FV a valid approach for gait recognition?*; b) *can we learn different camera viewpoints in a single classifier?*; c) *can we improve the recognition rate by spatially dividing the human body region?*; d) *what is the effect of using PCA-based dimesionally reduction on the recognition performance?*; and, e) *can the proposed model generallize well on unrestricted walk trajectories?*

A. Dataset

We perform our experiments on the “AVA Multi-View Dataset for Gait Recognition” (AVAMVG) [3]. In AVAMVG 20 subjects perform 10 walking trajectories in an indoor environment. Each trajectory is recorded by 6 color cameras placed around a room that is crossed by the subjects during the performance. Fig. 4 shows the scenario from the six available camera viewpoints. Note that depending on the viewpoint and performed trajectory, people appear at diverse scales, even showing partially occluded body parts. In particular, the 3rd and 6th camera viewpoints represented in Fig. 4 are more likely to show partially visible bodies most of the time than the other four cameras. Therefore, in our experiments, and without loss of generality, we will use only four cameras (i.e. $\{1, 2, 4, 5\}$). Trajectories 1 to 3 follow a linear path, whereas the remaining seven trajectories are curved. The released videos have a resolution of 640×480 pixels. Each video has around 375 frames, where only approximately one third of the frames contains visible people.

B. Experimental setup

Since we have multiple viewpoints of each *instance* (i.e. pair subject–trajectory), we assign a single label to it by majority voting on the viewpoints. This approach helps to deal with labels wrongly assigned to individual viewpoints. Note that instead of training an independent classifier (see Sec. II-D) per camera viewpoint, we train a single classifier with samples obtained from different camera viewpoints, allowing the classifier to learn the relevant gait features of each subject from multiple viewpoints. In order to increase the amount of training samples, we generate their *mirror* sequences, thus, doubling the samples during learning.

We describe below the different experiments performed to give answer to the questions stated at the beginning of this section.

Experiment A: baseline. We use the popular Bag of Words approach (BOW) [8] as baseline, which is compared to our approach. For this experiment, we use trajectories 1, 2 and 3 (i.e. straight path). We use a leave-one-out strategy on the trajectories (i.e. two for training and one for test). We sample dictionary sizes in the interval $[500, 4000]$ for BOW [2], and in the interval $[50, 200]$ for PFM. Both BOW and PFMs have a single level with two rows and one column (i.e. concatenation of two descriptors: half upper-body and half lower-body).

Experiment B: half body features. Focusing on PFM, we compare four configurations of the PFM on trajectories 1, 2 and 3: a) no spatial partition of the body; b) using only the top half of the body; c) using only the bottom half of the body; and, d) using the concatenation of the top and bottom half of the body.

Experiment C: dimensionality reduction. Since the dimensionallity of PFM is typically large, we evaluate in this experiment the impact of dimensionality reduction on the final recognition performance. We run Principal Component Analysis (PCA) both on the original low-level features (318 dimensions), and on the PFM vectors. We use the PFM descriptor, as in experiment A, on trajectories 1, 2 and 3.

Experiment D: training on straight paths and testing on curved paths. In this experiment, we use the PFM descriptor as in experiment A. We use trajectories 1 to 3 for training, and trajectories 4, 7 and 10 for testing. Note that in the latter sequences, the subjects perform curved trajectories, thus, changing their viewpoint (with regard to a given camera).

[1] VLFeat 0.9.17 is available at http://www.vlfeat.org/

[2] Larger dictionary sizes for BOW did not show any significative improvement. In contrast, the computational time increased enormously.

TABLE I. **Comparison of recognition results.** Each entry contains the percentage of correct recognition in the multiview setup and, in parenthesis, the recognition per single view. Each row corresponds to a different configuration of the gait descriptor. K is the GMM size used for FM. Best results are marked in bold. (See main text for further details.)

Experiment	*K*	*Trj=1+2*	*Trj=1+3*	*Trj=2+3*	*Avg*
BOW	4000	95 (78.8)	85 (62.5)	100 (84.4)	93.3 (75.2)
PFM-FB	150	100 (98.8)	100 (95)	100 (100)	100 (97.9)
PFM-H1	150	100 (95)	100 (87.5)	100 (97.5)	100 (93.3)
PFM-H2	150	100 (97.5)	95 (93.8)	100 (97.5)	98.3 (96.3)
PFM	150	100 (98.8)	100 (96.2)	100 (97.5)	100 (97.5)
PFM+PCAL50	150	100 (100)	100 (97.5)	100 (98.8)	100 (98.8)
PFM+PCAH256	100	100 (100)	100 (97.5)	100 (98.8)	100 (98.8)
PFM+PCAL100+PCAH256	150	100 (100)	100 (97.5)	100 (98.8)	100 (98.8)
PFM+PCAL50+PCAH256+pyr	100	100 (100)	100 (97.5)	100 (98.8)	**100 (98.8)**

C. Results

We present here the results of the experiments described above.

The results shown in Tab. I correspond to *experiments A, B and C* (see Sec. III-B) and have been obtained by training on two of the three straight trajectories ($\{1, 2, 3\}$) and testing on the remaining one (e.g. '*Trj=1+2*' indicates training on trajectories #1 and #2, then, testing on trajectory #3). Therefore, each model is trained with 160 samples (i.e. 20 subjects $\times$ 4 cameras $\times$ 2 trajectories) and tested on 80 samples. Each column '*Trj=X+Y*' contains the percentage of correct recognition per partition at instance level (i.e. combining the four viewpoints) and, in parenthesis, at video level; column '*Avg*' contains the average on the three partitions. Column K refers to the number of centroids used for quantizing the low-level features in each FM descriptor. Row 'BOW' corresponds to the baseline approach (see Sec. III-B). Row 'PFM-FB' corresponds to the PFM on the full body (no spatial partitions). Rows 'PFM-H1'and 'PFM-H2' correspond to PFM on the top half and on the bottom half of the body, respectively. Row 'PFM' corresponds to a single-level PFM obtained by the concatenation of the descriptors extracted from both the top and bottom half of the body. Row 'PFM+PCAL50' corresponds to our proposal but reducing with PCA the dimensionality of the low-level motion descriptors to 50 before building our PFM (i.e. final gait descriptor with $K = 150$ is 15000-dimensional). Row 'PFM+PCAH256' corresponds to our proposal but reducing with PCA the dimensionality of our final PFM descriptor to 256 dimensions before learning the classifiers (i.e. final gait descriptor is 256-dimensional). Row 'PFM+PCAL100+PCAH256' corresponds to our proposal but reducing both the dimensionality of the low-level descriptors and the final PFM descriptor (i.e. final gait descriptor is 256-dimensional). Row 'PFM+PCAL50+PCAH256+pyr' corresponds to a two-level pyramidal configuration where the first level has no spatial partitions and the second level is obtained by dividing the bounding box in two parts along the vertical axis, as done previously. In addition, PCA is applied to both the low-level descriptors and the final PFM vector.

TABLE II. **Comparative of recognition results on curved trajectories.** Training on trajectories $1 + 2 + 3$. Each column indicates the tested trajectory and each row corresponds to a different configuration of the gait descriptor. K is the GMM size used for FM. Best results are marked in bold.

Experiment	*K*	*Test=04*	*Test=07*	*Test=10*
PFM	150	90 (75)	95 (91.3)	**95** (81.0)
PFM+PCAL100	150	90 (72.5)	95 (92.5)	80 (77.2)
PFM+PCAL100+PCAH256	150	90 (73.8)	95 (90)	85 (81.1)
PFM+PCAL50+PCAH256+pyr	100	**95 (80)**	90 (88.8)	85 (82.3)
PFM+PCAL50+PCAH256+pyr	150	90 (75)	95 (90)	90 (**87.3**)
PFM+PCAL100+PCAH256+pyr	150	85 (71.3)	**95 (92.5)**	85 (82.3)

The results shown in Tab. II correspond to *experiment D* (see Sec. III-B) and have been obtained by training on trajectories $\{1, 2, 3\}$ (all in the same set), and testing on trajectories $\{4, 7, 10\}$ (see corresponding columns). As done in the previous experiments, different configurations of PFM have been evaluated. Each entry of the table contains the percentage of correct recognition in the multiview setup and, in parenthesis, the recognition per video.

D. Discussion

The results presented in Tab. I indicate that the proposed pipeline is a valid approach for gait recognition, obtaining a 100% of correct recognition on the multiview setup. In addition, the FV-based formulation surpasses the BOW-based one, as stated by other authors in the problem of image categorization [2]. In addition, the large dimensionality of the PFM can be drastically reduced by applying PCA, without worsening the final performance. Actually, reducing the dimensions of the low-level motion descriptors to 100, and the final PFM to 256, allows to achieve a similar recognition rate but decreasing significantly the computational complexity ($\approx \times 370$ smaller with $K = 150$).

If we focus on the idea of spatially dividing the human body for computing different gait descriptors, the results show that the most discriminant features are localized on the lower-body (row 'PFM-H2'), what confirms our intuition (i.e. gait is mostly defined by the motion of the legs). In addition, although in a slight manner (see values in parenthesis), the upper-body features (row 'PFM-H1') contribute to the definition of the gait as well.

Focusing on Tab. II, we can observe that PFM generalizes fairly well, as derived from the results obtained when testing on curved trajectories. From the three tested trajectories, the number #04 resulted to be the hardest when trying to classify per individual cameras (i.e. values in parenthesis). However, the use of the majority voting strategy on the multiview setup clearly contributed to boost the recognition rate (e.g. from 80 to 95).

With regard to the use of more than one level in PFM, we can see in Tab. I that similar results are obtained with the single- and two-level configurations. However, in the two-levels case, the number of needed GMMs is lower (i.e. 100 vs 150) than with single-level, as well as the low-level features can be reduced to half size (i.e. 50 vs 100). In addition, in the experiment on the curved trajectories (Tab. II), we can find in many cases an improvement at video level (e.g. 75 to 80 in trajectory #04). Although we tried an additional third level in the pyramid, the recognition rate did not increase.

In addition to the reported experiments, we also experimented with splitting the curved video sequences along the temporal axis to try to find nearly linear trajectories that could be better classified. However, the results did not show any improvement.

In summary, we can conclude that the proposed PFM allows to identify subjects by their gait by using as basis local motion (i.e. short-term trajectories) and coarse structural information (i.e. spatial divisions on the person bounding-box). Moreover, PFM does not need either segmenting or aligning the gait cycle of each subject as done in previous works.

IV. Conclusions and Future Work

We have presented a new approach for recognizing human gait in video sequences. Our method builds a pyramidal representation of the human gait based on the combination of densely sampled local features and Fisher vectors: the *Pyramidal Fisher Motion*.

The results show that PFM allows to obtain a high recognition rate on a multicamera setup: the AVAMVG dataset. In particular, a perfect identification of the individuals is achieved when we combine information from different cameras and the subjects follow a straight path. In addition, our pipeline shows a good behaviour on unconstrained paths, as shown by the experimental results – the model is trained on subjects performing straight walking trajectories and tested on curved trajectories. With regard to the PFM configuration, we have observed that it is beneficial to decorrelate (by using PCA) both the low-level motion features and the final PFM descriptor in order to achieve high recognition results and, in turn, decreasing the computational burden at test time – the classification with a linear SVM is extremely fast on 256-dimensional vectors. Since we use a person detector to localize the subjects, the proposed system in not restricted to deal with scenarios with static backgrounds. Moreover, the motion features used in this paper can be easily adapted to non static cameras by removing the global affine motion as proposed recently by Jain et al. [1].

In conclusion, PFM enables a new way of tackling the problem of gait recognition on multiple viewpoint scenarios, removing the need of using people segmentation as mostly done so far.

As future work, we intend to evaluate the proposed method on additional multiview datasets that include both people carrying objects and *impostors* (i.e. people external to the learnt subjects). With regard to the latter issue, as we use the continous output of the SVM to decide the identity, we could eventually discard impostors by thresholding such value.

Acknowledgments

This work has been partially funded by the Research Projects TIN2012-32952 and BROCA, both financed by FEDER and the Spanish Ministry of Science and Technology. We also thank David López for his help with the setup of the AVAMVG dataset.

References

[1] M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better action recognition,” in *CVPR*, 2013, pp. 2555–2562.

[2] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *ECCV*, 2010.

[3] D. López-Fernández and F. J. Madrid-Cuevas. (2013) The AVA Multi-View Dataset for Gait Recognition (AVAMVG). [Online]. Available: http://www.uco.es/investiga/grupos/ava/node/41

[4] J. E. Cutting and L. T. Kozlowski, “Recognizing friends by their walk: Gait perception without familiarity cues,” *Bulletin of the psychonomic society*, vol. 9, no. 5, pp. 353–356, 1977.

[5] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 34, no. 3, pp. 334–352, 2004.

[6] R. Martín-Félez and T. Xiang, “Gait recognition by ranking,” in *ECCV*, 2012, pp. 328–341.

[7] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Action Recognition by Dense Trajectories,” in *CVPR*, 2011, pp. 3169–3176.

[8] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *ICCV*, vol. 2, 2003, pp. 1470–1477.

[9] J. Han and B. Bhanu, “Individual recognition using gait energy image,” *IEEE PAMI*, vol. 28, no. 2, pp. 316–322, 2006.

[10] Y. Liu, J. Zhang, C. Wang, and L. Wang, “Multiple HOG templates for gait recognition,” in *Proc. ICPR*. IEEE, 2012, pp. 2930–2933.

[11] R. Martin-Felez, J. Ortells, and R. Mollineda, “Exploring the effects of video length on gait recognition,” in *Proc. ICPR*, 2012, pp. 3411–3414.

[12] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, “Video from nearly still: an application to low frame-rate gait recognition,” in *CVPR*, 2012, pp. 1537–1543.

[13] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *CVPR*. IEEE, 2007, pp. 1–8.

[14] W. Gong, M. Sapienza, and F. Cuzzolin, “Fisher tensor decomposition for unconstrained gait recognition,” in *Proc. of ECML/PKDD – TML Workshop*, 2013.

[15] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Proc. of the 13th Scandinavian Conf. on Image Analysis*, ser. LNCS, vol. 2749, June-July 2003, pp. 363–370.

[16] M. Eichner, M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, “2D articulated human pose estimation and retrieval in (almost) unconstrained still images,” *IJCV*, vol. 99, no. 2, pp. 190–214, 2012.

[17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE PAMI*, vol. 32, no. 9, 2010.

[18] V. Ferrari, T. Tuytelaars, and L. Van Gool, “Real-time affine region tracking and coplanar grouping,” in *CVPR*, 2001.

[19] T. Jaakkola, D. Haussler *et al.*, “Exploiting generative models in discriminative classifiers,” in *NIPS*, 1999, pp. 487–493.

[20] M. Marín-Jiménez, N. Pérez de la Blanca, and M. Mendoza, “Human action recognition from simple feature pooling,” in *Pattern Analysis and Applications*, 2012.

[21] E. Osuna, R. Freund, and F. Girosi, “Support Vector Machines: training and applications.” MIT, Tech. Rep. AI-Memo 1602, March 1997.

[22] A. Vedaldi and A. Zisserman, “Efficient additive kernels via explicit feature maps,” *IEEE PAMI*, vol. 34, no. 3, pp. 480–492, 2012.