Context-Aware Person Re-identification in the Wild via fusion of Gait and Anthropometric features

Athira Nambiar[1], Alexandre Bernardino[1], Jacinto C. Nascimento[1] and Ana Fred[2]
[1] Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal
[2] Telecommunications Institute, Instituto Superior Técnico, Lisbon, Portugal

***Abstract*— In this work, we present a context-aware ensemble fusion framework based on soft-biometric features, for long term person re-identification (Re-ID) in wild surveillance scenarios. The characteristics of a person that best correlate to its identity depend strongly on the view point. For instance, a person with a short stride gait is better perceived from a lateral view, whereas a person with a large chest is more distinct from a frontal view. Thus we associate context to the viewing direction of walking people in a surveillance scenario and choose the best features for each case. Using the MS Kinect™ sensor v.2, we collect data from walking subjects and extract associated anthropometric and gait features. Each context is analysed with a *Feature selection* technique (Sequential Forward Selection) so that only the most relevant features for the context are retained. Then, individual context-specific classifiers are trained leveraging those selected features. Finally, we propose a context-aware ensemble fusion strategy, which we term as *'Context-specific score-level fusion'*, based on the adaptive weighted sum of the results of individual classifiers. The proposed context-aware Re-ID framework demonstrate significant performance improvement both in terms of speed (up to 4.5 times faster) and accuracy (up to 17% rank-1 Re-ID rate) compared to the Context-unaware systems. From the study, we show that gait features are better for lateral views and anthropometric features are better for frontal views, confirming the results of previous studies.**

I. INTRODUCTION

In this work, we discuss the application of soft-biometrics in long term Person re-identification (Re-ID) using cameras in arbitrary view-points. We study the influence of the view-point in Re-ID performance and propose a methodology to exploit the *'view-point context'* to improve the overall performance. In particular, we propose a biometric enabled person re-identification system, using two kinds of soft biometric features *i.e.*, anthropometric and gait features. The best features for each context are selected for training context specific classifiers. Then, during run time, a context-specific fusion method provides the person Re-ID score.

Some works have employed Kinect based person Re-ID approaches leveraging soft-biometric cues [1], [2], [3]. Nevertheless, they employed view-point dependent methods *i.e.*, data was collected and algorithms were tested with a single walking direction with respect to the camera, which does not represent 'in the Wild' scenario where people walk in various directions. On the contrary, in this work, we collect people walking freely in an indoor office like scenario. Depending upon the strategic points inside a building (entry/exit points, and coffee machine/printer locations etc.), it was observed that the probability of people walking indoor could be explicitly represented in various directional view-points, which we term as '*Contexts*', rather than random walking paths. In addition to that, the potential features extracted by the sensor also have indicated clear distinction, according to different contexts. Based on these postulates, we redefine the classical Re-ID strategy by means of a novel 'context-aware person re-identification method', where we explicitly evaluate a context-specific feature matching criteria in Re-ID. In this regard, the major contributions of the paper are as follows:

- Feature selection via Sequential Forward Selection algorithm, to adaptively select the potentially relevant features in each context.
- Proposal of a '*Context-aware ensemble fusion framework*', wherein individual classifiers are trained specific to each context, and the Re-ID performance is analysed via our proposed '*Context-specific score level fusion*' strategy.

The rest of the paper is organized as follows. The related works are described in Section II. The proposed methodology is explained in Section III, *i.e.*, the dataset used, feature extraction method, and Context-aware ensemble fusion framework. In Section IV, the experiments conducted and the results obtained are discussed in detail. Finally, the summary of the paper and some future plans are enumerated in Section V.

II. RELATED WORK

The arrival of Kinect™ RGBD sensor gave rise to unprecedented advancements in the biometric and computer vision community, to devise many sophisticated techniques allowing view point invariance. Many Re-ID works utilizing Kinect data have been reported in the literature. By exploiting soft-biometric cues in contrast to the primarily appearance cues (colour or texture), they promote long term person Re-ID. In one of the earlier works *viz.*, [1], a specific signature built from a composition of several soft biometric (*e.g.*, skeleton and surface based features) cues extracted from the depth data, was computed for each subject. Then, Re-ID was accomplished by matching these signatures against the test subjects from the gallery set. Similarly, person re-identification from soft biometric cues was also addressed in another work [4], where skeleton descriptors (by computing several limb lengths and ratios) and shape traits (using point cloud shape) were used in order to re-identify people. In [2]

978-1-5090-4023-0/17/$31.00

both anthropometric features (*e.g.*, height, leg length, etc) and dynamic parameter related to gait (*e.g.*, knees movement, head oscillation) were used. Also, in [3] a methodology to extract anthropometric and gait features was addressed showing the results of applying different machine learning algorithms on subject Re-ID tasks. However, in those approaches, the acquisitions were conducted in a constrained manner *i.e.*, in a particular view-point. In this work, we build upon the state-of-the-art works but in a less constrained conditions, by explicitly imposing view-point changes in the dataset and by exploiting relevant features in each of those view-points (contexts).

Many definitions of context were encountered in the literature, depending on its field of application. According to the dictionary, context is defined as "*the surroundings, circumstances, environment, background or settings that determine, specify, or clarify the meaning of an event or other occurrence*"[5]. In our work, we define context as the view-point setting, under which features are computed. The application of context has been reported in diverse fields, for instance, in customer behaviour applications [6], where the context is viewed as the intent of a purchase (*e.g.* context of a gift). In [7], an application for Re-ID of the subject from instant messaging in a web surfing navigation is proposed. The context is the special characteristics of chatting text (*e.g.* content, token, syntax and structural based features). In [8] context was used for online customer re-identification, where the intent was to investigate whether customer behavior models of the context (in which a transaction takes place), can increase client re-identification performance. The contextual information is interpreted as the time of day when or the location where digital data was created. Few works, however, addressed the concept of context within the re-identification setting as we propose in this paper. In particular, [9] proposed a Re-ID paradigm which leveraged heterogeneous contextual information together with facial features. In particular, they used clothing, activity, human attributes, gait and people co-occurrence as various contexts, and then integrated all of those context features using a generic entity resolution framework called RelDC. Some other recent Re-ID works utilized context as a strategy for refining the classical Re-ID results via re-ranking technique [10], [11]. In those works, in addition to the content information of the subjects, they also paid attention to the context information (k-common nearest neighbors) to fine tune the Re-ID results. From our literature review, it was comprehended that context is a new tool whose effectiveness in Re-ID applications is yet to be completely explored.

III. PROPOSED METHOD

A. Database

In order to employ Re-ID in a realistic 'in-the-wild' scenario, it is quite essential to have a challenging unconstrained dataset, comprised of sequences of people walking in different directions. Since such a Kinect^{TM} based dataset (with different viewangles) towards gait based Re-ID was unavailable, we acquired our own dataset using a mobile platform, in the host laboratory. The Kinect^{TM} device is able to track movements from users by using a skeleton mapping algorithm, and is able to provide the 3D information related to the movements of body joints. The position of camera as well as the walking directions of subjects were deliberately altered in order to ensure a typical surveillance scenario. Multiple walking sequences of 20 subjects in five different directions *i.e.*, Left lateral (LL at $\sim 0°$), Left diagonal (LD at $\sim 30°$), Frontal (F at $\sim 90°$), Right diagonal (RD at $\sim 130°$) and Right lateral (RL at $\sim 180°$) were collected. Altogether we have 300 video sequences comprising 20 subjects (3 video sequences per person in a particular context) in the aforementioned directions. Different walking directions and sample video frames extracted from our dataset, are shown in Fig. 1.

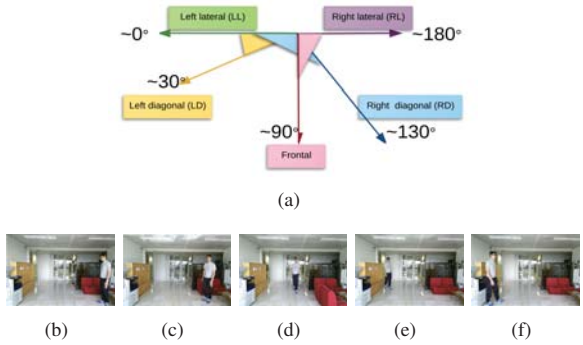


Fig. 1. Data acquisition: (a) Subject walking directions in front of the camera system (Direction angles are defined with respect to the image plane.) (b-e) Sample frames from our data acquisition, in five different directions- left lateral ($\sim 0°$), left diagonal ($\sim 30°$), frontal ($\sim 90°$), right diagonal ($\sim 130°$) and right lateral ($\sim 180°$) respectively.

B. Feature extraction

The real-time skeleton models tracked via Kinect^{TM} are composed of 25 body joints. The foremost step was pre-processing, to remove the noise contents in the data. By empirically analysing the evolution of lower body angles over time, we cleared the unwanted jerks in the signals especially, at the boundaries of the Kinect range. The detailed explanation of pre-processing and feature extraction phases were reported in the prior work by the authors in [12]. Then, based on those cleaned signals, the functional units of gait *viz.*, gait cycles, were estimated. A gait cycle comprises of sequence of events/movements during locomotion since one foot contacts the ground until the same foot again contacts the ground. Hence, based on the cleaned data, the periodicity of the feet movement is estimated to define gait cycle and various features were extracted within this gait period.

Two kinds of features were extracted: *(i)* Anthropometric features *i.e.*, the static physical features defining the body measurements and *(ii)* Gait features *i.e.*, dynamic features defining the kinematics in walking. See Table I for the list of features we used. Under the anthropometric feature set, body measurements defining the holistic body proportions of the subject such as height, arm length, upper torso length, lower torso length, upper to lower ratio, chest size, hip

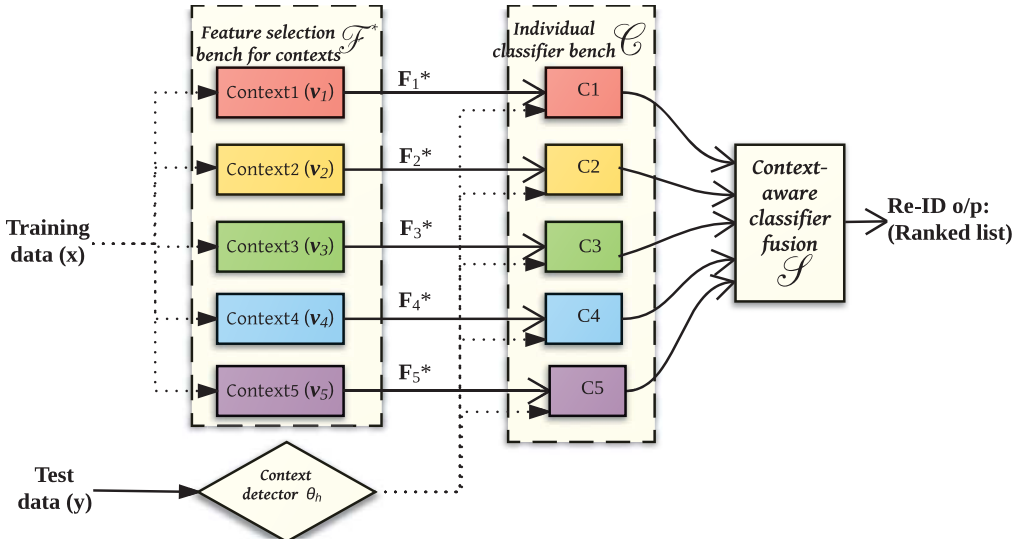


Fig. 2. Context-aware ensemble fusion system: It internally consists of a feature selection context bench, an individual classifier bench, a classifier fusion module and a context detector module. The individual classifiers for each context are trained using individual feature subspace ensembles $\mathbf{F}_\mathbf{j}^*$, obtained for each context. When the test data enters, context detector identifies the context and activates the corresponding ensemble classifiers. Then, the context-aware classifier fusion strategy finally combines the results of those ensemble classifiers to produce the global result.

TABLE I

LIST OF ANTHROPOMETRIC AND GAIT FEATURES USED IN OUR EXPERIMENTS. L& R CORRESPOND TO 'LEFT AND RIGHT' AND X& Y CORRESPOND TO 'ALONG X AND Y AXES'. THE NUMBERS OF FEATURES DERIVED ARE SHOWN WITHIN PARENTHESIS.

Anthropometric features	**Gait features**	
Height-(*1*)	Hip angle(L&R)-(*4*)	Hip position(L&R)(x& y)-(*8*)
Arm length-(*1*)	Knee angle(L& R)-(*4*)	Knee position(L&R)(x& y)-(*8*)
Upper torso-(*1*)	Foot distance-(*2*)	Ankle position(L&R)(x& y)-(*8*)
Lower torso-(*1*)	Knee distance-(*2*)	Hand position(L&R)(x& y)-(*8*)
Upper-lower ratio-(*1*)	Hand distance-(*2*)	Shoulder position(L&R)(x& y)-(*8*)
Chestsize-(*1*)	Elbow distance-(*2*)	Stride-(*1*)
Hipsize-(*1*)	Head position(x& y)-(*4*)	Stride length-(*1*)
	Spine position(x& y)-(*4*)	Speed-(*1*)

size were collected. Similarly, under the gait features, the behavioural features deriving from the continuous monitoring of joints during the gait were collected. In particular, mean and standard deviation of the various measurements during a gait cycle were collected *i.e.*, (i) the angles at various body joints; (ii) the distance between various right-left limbs and; (iii) the position of body joints. Also three scalar features related to walking, *viz.*, stride length, stride time and the speed of walking, are computed within the gait features. Hence, the feature set contains a total of 7 anthropometric features and 67 gait features. In Table I, the numbers of features derived are shown in parenthesis.

C. Context-aware ensemble fusion

One of the most significant contributions of this work is a novel context-aware ensemble fusion strategy. First, we present an evaluation of the impact of the various data features in various contexts *i.e.*, view-points, and then employ a context-based fusion method to obtain the final Re-ID result. We accredit the work on Feature subspace ensembles [13] which acted as a motivation to the authors to come up with an analogous ensemble fusion strategy. That work presented an approach to run multiple parallel Feature selection stages with different training conditions, in order to obtain the best features, by using majority voting of the feature ensembles.

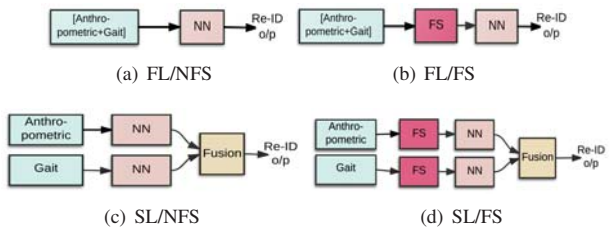
Our proposed framework is shown in Fig. 2. It is composed of four modules: (i) *Feature selection Context bench* (ii) *Individual classifier bench*, (iii) *Context detector module* and (iv) *Context-aware classifier fusion module*.

1) ***Feature selection Context bench****:* Our data for evaluation consists of the feature vectors extracted at various view-points, as mentioned earlier. We denote those five context view-points as $\mathbf{v_1}, ..., \mathbf{v_N}$, with $N = 5$, corresponding to LL, LD, F, RD and RL directions. We analyse the data in each context individually by leveraging a Feature Selection (FS) scheme in order to retain only the most discriminative and relevant features.

In particular, we employed Sequential Forward Selection (SFS) algorithm [14] as an instance of FS, as it is well known and widely used in practice. It works iteratively by adding features to an initial subset, seeking to improve a given measure, by selecting more features at each iteration. Suppose, $\mathbf{x} = \{x_1, \cdots, x_n\}$ denotes a set of n samples represented in a d-dimensional space, each with a d-dimensional feature set $\mathbf{F} = [f_1, \cdots, f_d] \in \mathbb{R}^{1\times d}$. FS analyses this d-dimensional space in order to identify which features $f_i \subset \mathbf{F}$ are potentially relevant, and which can be discarded according to some feature subspace evaluation criteria $\boldsymbol{J}$ and ultimately derive $\mathbf{F}_\mathbf{j}^*$, containing the most relevant features.

Specifically, the Sequential Forward Selection (SFS) algorithm works as follows: It starts from an empty feature set $\mathbf{F^*_{t=0}}$. At each step $\mathbf{F^*_{t+1}}$ all possible super-spaces containing the most relevant feature subspace in the previous step, $\mathbf{F^*_t}$, and one from the remaining features $f_i \in \mathbf{F} \backslash \mathbf{F^*_t}$ are formed and evaluated by $\boldsymbol{J}$. This iterative search will proceed until a stopping criteria is met, for which we considered the degradation of $\boldsymbol{J}$ *i.e.*, if none of the super-spaces formed at a given step $\mathbf{F^*_{t+1}}$ improves $\boldsymbol{J}$, the search stops and the subspace $\mathbf{F^*_t}$ is considered as the best feature subset. Finally, the outputs of the Feature selection context bench consists of an ensemble of feature subspace *i.e.*, the features selected for each particular context $\mathscr{F}^* = \left[\mathbf{F^*_1}, \cdots, \mathbf{F^*_5}\right]$. For the implementation of the algorithm, the authors used SFS package[1][15]. We used 1NN classifier with an Euclidean neighborhood metric in the SFS scheme.

*2) **Individual classifier bench**:* Since our training data consists of both anthropometric and gait features, we need to exploit both of them in training our each individual classifier. In this regard, we exploit various fusion techniques in order to combine anthropometric and gait features. Traditionally, there are many fusion strategies at various levels *viz.*, feature level fusion, score level fusion, rank level fusion or decision level fusion [16], of which we select both feature level fusion and score level fusion strategies in our work. In order to see the impact of various fusion strategies, we conduct two baseline fusion schemes without Feature selection: (i) Feature-level fusion without FS, represented as FL/NFS and (ii) Score-level fusion without FS, represented as SL/NFS. The schematic representations of the aforementioned are shown in Fig. 3 (a) FL/NFS and (c) SL/NFS respectively.



(a) FL/NFS (b) FL/FS

(c) SL/NFS (d) SL/FS

Fig. 3. Various Fusion-Feature selection schemes employed in this work. Top and bottom rows represents feature-level and score-level fusion strategies respectively. Feature selection (FS) is not used in case studies (a) FL/NFS and (c) SL/NFS, whereas (b) FL/FS and (d) SL/FS shows the inclusion of FS module.

In Feature level fusion (see Fig. 3 (a)), the biometric sets of the same individual are concatenated after an initial normalization (Min-max) scheme. This way, we concatenate our 7D anthropometric features and 67D gait features in order to make a 74D feature vector. Then, the concatenated feature vector is used in the classifier in order to represent the identify of an individual. Instead, in score level fusion (see Fig. 3 (c)), the fusion is carried out at the score level. The matching scores of each biometric sets are determined independently using two different classifiers and the matching scores at their outputs are fused in order to provide an aggregate score result. As explained in [16], normalized distance scores obtained at each individual classifiers can be fused using some combination rule such as sum, product, min, max or median. In our approach, we adopted sum rule as the classifier combination rule.

After the baseline cases, we further conduct our proposed FS-enabled fusion strategies as well. Here, the biometric sets are fed into a FS module prior to the classification stage so that, only the selective feature subspace $\mathbf{F^*_j}$ (as explained in Section III-C.1) will be used as the individual feature vector. In this regard, two more fusion schemes with Feature selection are carried out: (i) Feature-level fusion with FS, represented as FL/FS and (ii) Score-level fusion with FS, represented as SL/FS. The schematic representations of the aforementioned are shown in Fig. 3 (b) FL/FS and (d) SL/FS respectively.

Thus, as explained above, four different Fusion-FS schemes are conducted in order to assess the performance of each individual context classifiers within the classifier bench. In all of those case studies, a leave one out evaluation strategy is performed within each context, with a classifier specification of Nearest neighbour (NN) using euclidean distance metric. The experimental results obtained are explained in Section IV-A, and the best among all those fusion-FS scheme is further used as the *de facto* standard scheme in our framework. Based on this standard scheme, five different classifiers are trained corresponding to each context, which will form the Individual Classifier bench $\mathscr{C} = [\mathbf{C_1}, \cdots, \mathbf{C_5}]$.

*3) **Context detector**:* Context detector is the module where the context (view-point) of the test sample is estimated. The design of the context detector module was carried out by analysing the evolution of any static joint along the sequences over a gait cycle. We used 'SpineShoulder' *i.e.*, the base of the neck refering to joint number 20 of Kinect^{TM} v.2[2], since it remains more or less stable while walking. Then, the direction of walking was estimated by analysing the direction of the joint vector. Suppose $\vec{h_{begin}}$ and $\vec{h_{end}}$ denotes the position of the joint in the first frame and last frame respectively. Then the directional vector among these frames $\vec{h} =< h_x, h_y, h_z >$ can be obtained as follows:

$$\vec{h} = \vec{h_{end}} - \vec{h_{begin}}, \tag{1}$$

The y component h_y is only related to the vertical direction and hence is ignored. Then, the angular direction $\theta_{\vec{h}}$ made by $\vec{h}$ can be determined by measuring the inverse tangent of h_z/h_x.

$$\theta_{\vec{h}}(degrees) = \tan^{-1}(h_z/h_x) * 180/\pi \tag{2}$$

Whenever a test data $\mathbf{y} \in \mathbb{R}^{1\times d}$ enters into the system, its context is estimated using (1) and (2), and the corresponding ensemble classifiers are activated in order to proceed with context-aware classifier fusion.

[1]http://users.spa.aalto.fi/jpohjala/featureselection/

[2]https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx

*4) **Context-aware Classifier fusion**:* Based on the results from context detector module, this classifier fusion module performs a context-specific adaptive fusion of the results obtained at the outputs of individual classifiers $\mathscr{C} = [\mathbf{C_1}, \cdots, \mathbf{C_5}]$. In order to facilitate this, an extended version of score-level fusion based on context is proposed in this work, which we term as '***Context-specific score level fusion***'. This could be analysed homologous to the concept of user-specific score-level fusion in multibiometric systems, where user-specific weights were assigned to indicate importance of individual biometric matchers [16]. In a similar way, in our proposal, we endorse adaptive weights to scores from different classifiers according to its context, in order to increase the influence of more reliable context. In order to facilitate this adaptive weighting scheme, we employ linear interpolation technique.

Consider a test sample $\mathbf{y}$, at an arbitrary view-point context $\mathbf{v_{test}}$, is entering into the system. The context is detected using the context-detector module. Suppose the context lies in between our pre-defined context views say, $\mathbf{v_i}$ and $\mathbf{v_j}$. The individual classifiers for both aforementioned contexts $\mathbf{C_i}$ and $\mathbf{C_j}$ are selected alongwith their matching scores $\mathbf{s_i}$ and $\mathbf{s_j}$ respectively. The context-specific score level fusion $\mathscr{S}$ is computed as weighted sum of those scores as follows:

$$\mathscr{S} = \eta * \mathbf{s_i} + (1-\eta) * \mathbf{s_j}, \tag{3}$$

where $\eta \in [0,1]$. The weight η is computed via linear interpolation of the two contexts *i.e.*, $\eta = |\mathbf{v_j} - \mathbf{v_{test}}|/|\mathbf{v_j} - \mathbf{v_i}|$. The special case where only single context is activated, η of the nearest context turns to be 1, and all the others will be 0. Regarding these concepts, we analyse different case studies in detail, in the experimental section Section IV-B.

IV. RESULTS AND DISCUSSION

The performance of the context-aware ensemble fusion strategy was evaluated on our own database collected from 20 people, mentioned in previous section (see Section III-A). Two major experiments were carried out: *(A) Training the individual context-specific classifier*, in which each individual classifier was learned specific to its context. Intermediate experiments leading to this standard scheme (such as various fusion-FS schemes for performance assessment, context-specific feature ensembles selected) are also detailed in this section. The second experiment is *(B) Context-Specific Score Level Fusion*, wherein the final Re-ID result was achieved via adaptive fusion of the ensemble classifiers. Under this part, the experiments on context detection and context-specific fusion strategy are detailed. In order to evaluate the performance of our Re-ID algorithms, we use the popular method of choice, cumulative matching characteristic (CMC) curve. As per [17], "CMC shows how often, on average, the correct person ID is included in the best K matches against the training set for each test image".

A. Training the individual context-specific classifiers

In this step, we assessed the individual context classifier performance leveraging the 7D anthropometric features and 67D gait features. The impact of various fusion and FS schemes were analysed in this stage, via the four extensive case studies explained in Section III-C.2.

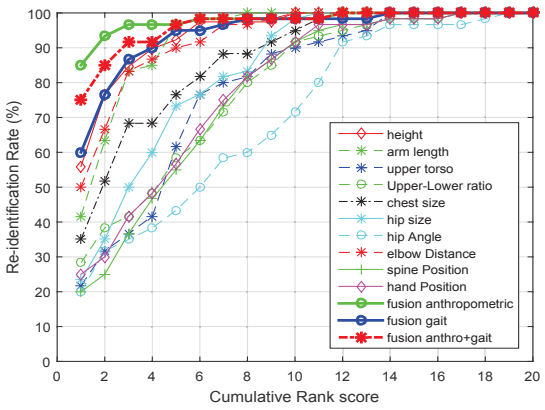


Fig. 4. Re-ID performances of individual as well as fused features in frontal context. Only a subset of individual features with classification rate $\geq$ 20% at Rank-1, are shown. The fusion results of anthropometric features (green with circle markers), gait features (blue with circle markers) and both anthropomeric and gait features (red with star markers) are shown via bold curves. For fusion of features, feature-level fusion strategy is adopted.

*1) **Why Feature selection is required?**:* Prior to the selection of feature subspace ensembles, initially we tried to analyse the Re-ID performance of each gait as well as anthropometric feature individually. This is performed with the NN classifier explained before but using a single feature *i.e.*, no fusion. Fig. 4 shows the individual performances of some of the best features[3] in person Re-ID in the frontal context. We can observe that certain features are quite relevant and discriminative (*e.g.*, height 55.89%, arm length 41.67%, elbow distance 50.00% and chest size 35.00%) compared to the others in re-identifying people. Another interesting result was that the feature level fusion of all features among a biometric set (gait or anthropometric) resulted in better performance. Or in other words, multimodal fusion outperformed the individual Re-ID results. Referring to the CMC curves in Fig. 4, we can observe that fusion of all anthropometric features resulted in 85.00% Re-ID rate at Rank-1 (bold green curve) and fusion of all gait features resulted in 60.00% Re-ID rate at Rank-1 (bold blue curve).

After the feature level fusion of anthropometric features and gait features separately, we further conducted multimodal fusion of the biometric features altogether *i.e.*, both anthropometric and gait features. Within this scheme, we first utilized feature-level fusion strategy *i.e.* as per Fig. 3 (a) FL/NFS. However, we could observe that the multimodal fusion at feature level resulted in lower Re-ID rate (75% illustrated by bold red Dash-dot curve in Fig. 4) lying in between the anthropometric and gait fusion results. This

[3]Among 74 features, only those features with individual Re-ID performance $\geq$ 20% are illustrated here.

under-performance could be ascribable to the large number of potentially misleading irrelevant/redundant features in the feature vector. To tackle this issue, we applied feature selection strategy by exploiting SFS algorithm as explained in Section III-C.1 and carried out its FS-enabled counterpart Fig. 3 (b) FL/FS.

*2) **Various Fusion-Feature selection schemes**:* After observing the lower performance of the multi-modal system without feature selection, we thereafter carried out an extensive analysis on different fusion-FS schemes as mentioned in Section III-C.2.Within this set of assessment studies, we carried out all the four fusion-FS schemes *i.e.*, (a) FL/NFS, (b) FL/FS, (c) SL/NFS and (d) SL/FS, leveraging both feature level/score level fusion and without/with FS. The performance results of all those case studies are illustrated in Fig. 5. The corresponding cumulative ranked list (showing anthropometric, gait and overall CMC rank-1) is also shown in Table II. These experimental results corroborate that:

- *Feature selection (FS) improves Re-ID accuracy, compared to without FS (NFS).*
- *Score-level fusion works better than the feature level fusion in Re-ID.*
- *Overall performance of SL/FS is the best among the group and thus is considered as the 'de-facto' in our context-aware ensemble fusion framework, at the individual classifier bench.*

TABLE II

CHART SHOWING THE RE-ID ACCURACY RATES FOR FIVE CONTEXTS. THE ACCURACY RATES SHOWN IN EACH CELL REPRESENT ANTHROPOMETRY BASED RE-ID, GAIT BASED RE-ID AND OVERALL RE-ID RESPECTIVELY, AT RANK-1 CMC. THE HIGHEST RE-ID RATE OBSERVED IS HIGHLIGHTED IN BOLD LETTERS.

Context	FL/NFS	FL/FS	SL/NFS	SL/FS
Left	61.67	61.67	**63.33**	58.33
Lateral	53.33	81.67	53.33	**83.33**
	63.33	85.00	78.33	**86.67**
Left	70.00	63.33	**75.00**	63.33
Diagonal	48.33	**61.67**	48.33	**61.67**
	55.00	68.33	71.67	**80.00**
Frontal	**85.00**	78.33	**85.00**	78.33
	60.00	**71.67**	58.33	70.00
	75.00	**93.33**	91.67	**93.33**
Right	**66.67**	**66.67**	**66.67**	**66.67**
Diagonal	46.67	**66.67**	46.67	63.33
	51.67	65.00	66.67	**78.33**
Right	40.00	45.00	40.00	**48.33**
Lateral	63.33	81.67	63.33	**85.00**
	70.00	81.67	80.00	**83.33**

*3) **Context-specific FS ensembles**:* Based on the results obtained, we attribute SL/FS as the de-facto strategy in our framework. The score level fusion of the selected features (both gait and anthropometric) were used to train individual classifiers for each context, at the classifier bench. Also, at this training phase, we also comprehended the relevant features for each context. For analysing the same, we conducted a holistic FS criteria with Cross-validation scheme in each context, which resulted in Table. III. This shows the context-specific features selected for each individual classifier, and using these results, the classifier bench is trained for future evaluation.

It is quite remarkable that the impact of globally discriminative anthropometric features such as height, arm length, chest size are highly relevant in almost all the contexts. However, some features clearly show its influence dependent on the context. For example, gait features presenting angular evolution (hipAngle) and distance showing various right-left limbs during the gait (knee distance, hand distance, elbow distance etc.) were selected in the frontal view. At the same time, many other gait features such as stride length, vertical position evolutions at various joints (headY_μ, kneeY_μ, spineY_μ,hipY_μ, handY_μ etc.) clearly exhibited the evidence of their influence in the lateral contexts. As a consequent result, lateral cases bestowed higher performance of gait features against anthropometric features in our experiments (see SL/FS results in Table II where LL and RL achieved 83.33% and 85.00% gait based Re-ID accuracies respectively against corresponding anthropometric based Re-ID accuracies 58.33% and 48.33%), in contrast to the frontal cases where anthropometric features showed better Re-ID performance (Anthropometry based Re-ID is 78.33% against gait based Re-ID of 70.00%). This results clearly corroborate the reason behind why usually gait analysis techniques are better manifested in lateral view rather than in front view.

TABLE III

CONTEXT-SPECIFIC FEATURES SELECTED VIA SL/FS SCHEME, DURING THE TRAINING OF INDIVIDUAL CONTEXT CLASSIFIERS. ONLY 28 FEATURE SUBSET OUT OF WHOLE 74 FEATURES WERE SELECTED.

Feature	LL	LD	F	RD	RL	Feature	LL	LD	F	RD	RL
height	✓	✓	✓	✓	✓	spineY_μ	✓				
arm	✓	✓	✓	✓		lhipY_μ	✓				
upper	✓			✓		lkneeY_μ	✓	✓		✓	✓
lower		✓		✓	✓	rkneeY_μ	✓	✓		✓	
ULratio		✓		✓		rankleY_μ				✓	
chestsize		✓	✓	✓	✓	lhandX_μ			✓		
hipsize	✓		✓	✓		lhandY_μ	✓				
hipAngle			✓			lhandY_{SD}				✓	
kneeDist$_{\mu,SD}$			✓			rhandY_μ				✓	✓
handDist$_{\mu,SD}$			✓			lshouldY_μ	✓				
elbowDist$_\mu$		✓	✓	✓		lshouldY_{SD}		✓			
elbowDist$_{SD}$				✓		rshouldY_μ					✓
headY_μ	✓	✓	✓		✓	rshouldY_{SD}			✓		
headY_{SD}			✓			strideLength	✓				✓

B. Context-Specific Score Level Fusion

After the training of each individual classifier leveraging the context-specific features selected as per in Table III, we conducted testing of our proposed method in pose-invariant scenario where test sample could be at any arbitrary context. In all of our test experiments, we employ a leave-one-out evaluation strategy, where we select one sample at a time and is compared against the rest of the samples in the gallery. This procedure is iterated throughout all the samples in the dataset.

*1) **Context detection**:* When a test sample at an arbitrary context enters into the system, the foremost stage is to detect the context of the test sample by enabling a Context detector module (see Section III-C.3). Then, based on the

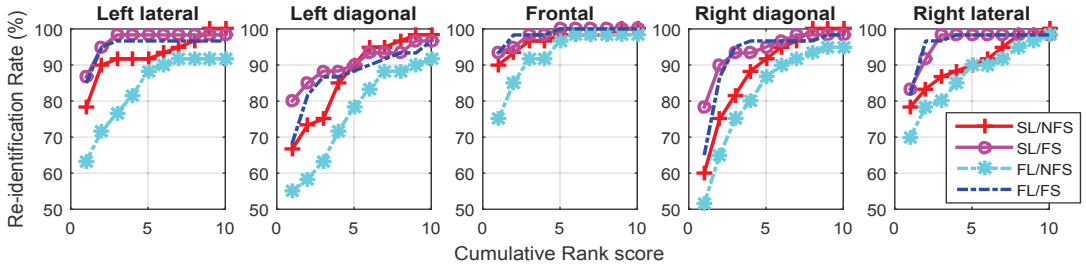


Fig. 5. The Re-ID performances of various Fusion-FS schemes mentioned in Fig. 3 along five contexts *viz.*, left lateral($\sim 0^\circ$), left diagonal($\sim 30^\circ$), frontal ($\sim 90^\circ$), right diagonal($\sim 130^\circ$) and right lateral($\sim 180^\circ$) respectively. Cumulative matching scores up to 10 subjects are shown.

detected context, corresponding context-specific classifiers are activated. In order to enable this, a prerequisite was to empirically verify the actual contextual view-points existing in our global dataset, and thus define *'contexts'* based on the gross view-points along a particular direction. So, in order to comprehend the existing contexts in our dataset, a prior context analysis was carried out on our global database, which resulted in context clusters as shown in Fig. 6. This empirical analysis enabled us to obtain better insight of the actual view-points spread within each contexts. Based on this study, we could observe that five contexts $\mathbf{v_1}, ..., \mathbf{v_5}$ are spread around their respective clustermeans $\mu = [1.67, 35.63, 92.83, 130.70, 180.17]^\top$, with standard deviations $\sigma = [3.64, 4.90, 3.29, 5.34, 3.99]^\top$.

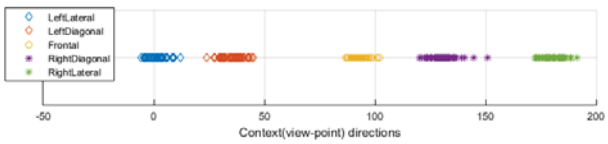


Fig. 6. Distribution of the contexts in the dataset.

*2) **Context-specific fusion strategy**:* In this fusion stage, information from different context specific classifiers are fused using the runtime estimate of the current context, given by the context detector. After the context detector determines the current context, the corresponding ensemble classifiers are activated. Under this context-aware paradigm, two schemes were proposed: (i) Using ***single context (binary weighted)***, in which only the closest context is selected based on the nearest cluster mean among all the clusters. Hence, only that specific context in the gallery is activated and the test is matched against all the rest of the 59 samples in that particular context and the ranked Re-ID list is obtained. The second context-aware scheme is (ii) using ***Two contexts (linear interpolated weights)***, wherein depending upon the probe context, two nearby contexts (between which the test probe lies) are activated. Then, the test sample is matched against the rest of the gallery samples in those two contexts, and respective matching scores are generated[4]. Then, depending on the distance of the test context with respect to those contexts, adaptive weights are assigned via linear interpolation technique, and Context-specific score level fusion strategy is applied to obtain the aggregate Re-ID result (see Section III-C.4).

In order to perform a comparison of our context-aware proposal, we also conducted baseline studies without the notion of context (Context-unaware). In these baseline scenarios, we disabled the context detector module, and hence no notion of the probe context is available to the system. Three baseline studies were performed. In the first case study, the test sample enters into the system and it matches against all the rest of the 299 gallery samples (from all the contexts), and computes the ranked Re-ID list based on the matching score. In this method, albeit the testing is performed as context-unaware, the features used per person are context-aware. In other words, the samples per person used in the test mode were trained a priori based on the context-specific feature selection. Hence, we term this scenario as ***'Pseudo baseline'***. To tackle this context dependency, we conducted the second case study called ***'Pure baseline'***, where we made the system context-unaware not only at the testing phase, but also at the training phase. In order to conduct this analysis, we retrained our system and applied global feature selection upon all the samples, independent of the context. Thus, the same features got selected globally, thus making the FS in all the samples context-unaware. Afterwards, testing was conducted as in the *'Pseudo baseline'* case, where the probe is matched against all the rest of the 299 gallery samples, and computes the ranked Re-ID list based on the matching score. The third context-unaware case study is with the assumption that the chance of the probe sample within the pre-defined contexts are equally likely *i.e.*, the same probability of occurring. Hence, ***equal weights*** of 0.2 is assigned to each contexts. The probe sample is tested against the gallery samples in each context and then weighted sum upon all the five individual classifier matching scores are performed to obtain the aggregate matching score and the consequent ranked list.

The results of all the five case studies mentioned above are shown in Table IV. It is quite remarkable to observe that, context-aware methods (either by using a single or two contexts) bestow high performance level $\sim$88%, whereas

[4] Since the number of gallery samples per context may vary, the matching scores per sample are of different size. Hence, we use matching scores per person by computing the best score (minimum distance score) per person.

TABLE IV

RESULTS OF CLASSIFIER FUSION SHOWING OUR PROPOSED CONTEXT-AWARE CLASSIFIER FUSION AGAINST CONTEXT-UNAWARE BASELINE CASE STUDIES. IN CONTEXT-AWARE CASES, CONTEXT DETECTOR MODULE IS ENABLED, WHEREAS IN THE CONTEXT-UNAWARE CASES, CONTEXT DETECTOR MODULE IS DISABLED

	Context-unaware			**Context-aware**	
	No context (Pseudo baseline)	**No context (Pure baseline)**	**All contexts (equal weights)**	**1 context (binary weights)**	**2 contexts (adaptive weights)**
Anthropometric	25.33%	60.33%	45.67%	68.67%	68.00%
Gait Re-ID	26.67%	70.33%	53.33%	84.67%	85.67%
Overall Re-ID	74.33%	79.33%	71.33%	88.67%	88.33%
Processing time	25.7sec.	21.64sec.	25.92sec.	5.59sec.	10.47sec.

all variants of Context-unaware cases miss good results ∼71%-79%. Also, since there is no notion of the context in Context-unaware cases, the probe sample has to be matched against all the rest 299 samples in the global dataset. At the same time, in context-aware cases, the information of the direction helps to reduce the size of the gallery set drastically by making it context-specific. Due to this reason, context-aware systems performed faster (∼5-10 sec.) compared to the context-unaware system (∼21-25 sec.). This highly accentuates the fact that, in unconstrained scenarios, the knowledge of context can augment the performance of a Re-ID system in terms of both speed and accuracy.

V. CONCLUSIONS

In this work, a novel context-aware ensemble fusion framework has been proposed towards long term Re-ID in the wild. In order to develop this framework, we first analysed the individual as well as fused Re-ID results leveraging anthropometric and gait features. Based on the observation that albeit multimodal fusion improves the result, naïve integration of large number of potentially irrelevant features can cause degradation of results, we proposed a Feature selection (FS) technique by employing Sequential Feature selection (SFS) algorithm. In this regard, various fusion-FS strategies were analysed and the best among all (SL/FS) has been selected as the *de facto* standard in our framework.

Another contribution was the concept of context-specific classifiers. This was quite significant depending upon the property of the sensor that, specific features are well acquired in specific directions. Based on our FS scheme, we adaptively selected those features depending upon the directions which we term as *'contexts'*, and trained each individual classifiers based on the selected features for that particular context. During the run time, the direction of the probe sample was determined using a Context-detector module, and the corresponding neighboring context/contexts were activated. Afterwards, a context-aware classifier fusion was facilitated via our proposed *'Context-specific score level fusion'*, and the Re-ID was carried out. The experimental results showed that comparing to the Context-unaware systems, context-aware systems performed significantly faster (up to 4.5 times) and accurate (up to 17 percentage point better).

In the future works, we envisage to extrapolate this study by collecting more data in more random directions of walk (moving from a denser context clusters to scatter clusters), and to analyse how the linear interpolation strategy can enhance the results. Another idea is also to incorporate multiple contexts in the scenario, (*i.e.*, in addition to the view-point, also include distance to the camera, occurrence of face, person co-occurances etc.) in order to improve the re-identification performance.

VI. ACKNOWLEDGMENTS

This work was supported by the FCT projects [UID/EEA/50009/2013], AHA CMUP-ERI/HCI/0046/2013 and FCT doctoral grant [SFRH/BD/97258/2013].

REFERENCES

[1] I.B. Barbosa, M. Cristani, D.B. Alessio, L. Bazzani, and V. Murino, "Re-identification with RGB-D Sensors", *First International Workshop on Re-Identification*, 2012, pp 433-442.

[2] E. Gianaria, M. Grangetto, M. Lucenteforte and N. Balossino, "Human Classification Using Gait Features", *First International Workshop*, BIOMET, Sofia, Bulgaria, 2014, pp 16-27.

[3] V.O. Andersson and R.M. Araujo, "Person Identification Using Anthropometric and Gait Data from Kinect Sensor", *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[4] M. Munaro, A. Fossati, A. Basso, E. Menegatti and L.V. Gool, One-Shot Person Re-Identification with a Consumer Depth Camera, *Person Re-Identification*, 2014, pp 161-181.

[5] `https://en.wiktionary.org/wiki/context`

[6] C. Palmisano, A. Tuzhilin and M. Gorgoglione, Using context to improve predictive modeling of customers in personalization applications, *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, 2008, pp 1535-1549.

[7] Y. Ding, X. Meng, G. Chai and Y. Tang, User identification for instant messages, *In Neural information processing*, vol. 7064, 2011, pp 113-120.

[8] U. Panniello, S. Hill and M. Gorgoglione, Using context for online customer re-identification, *Expert Systems With Applications*, vol. 64, 2016, pp 500-511.

[9] L. Zhang, D.V. Kalashnikov, S. Mehrotra and R. Vaisenberg, Context-based person identification framework for smart video surveillance, *Machine Vision and Applications*, vol. 25, 2014, pp 1711-1725.

[10] Q. Leng, R. Hu, C. Liang, Y. Wang and J. Chen, Person re-identification with content and context re-ranking, *Multimedia Tools and Applications*, vol. 74, 2015, pp 6989-7014.

[11] J. Garcia, N. Martinel, C. Micheloni and A. Garde, "Person Re-Identification Ranking Optimisation by Discriminant Context Information Analysis", *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp 1305-1313.

[12] A. Nambiar, A. Bernardino, J.C. Nascimento and A. Fred, "Towards view-point invariant Person Re-identification via fusion of Anthropometric and Gait Features from Kinect measurements", *International Conference on Computer Vision Theory and Applications (VISAPP)*, Porto, 2017.

[13] H. Silva and A. Fred, "Feature Subspace Ensembles: A Parallel Classifier Combination Scheme Using Feature Selection", *International Workshop on Multiple Classifier Systems*, Prague, 2007, pp 261-270.

[14] A. W. Whitney, A direct method of nonparametric measurement selection, *IEEE Transactions on Computers*, vol. 20, 1971, pp. 1100-1103.

[15] J. Pohjalainen, O. Rsnen and S. Kadioglu, Feature Selection Methods and Their Combinations in High-Dimensional Classification of Speaker Likability, Intelligibility and Personality Traits, *Computer Speech and Language*, vol. 29, 2015, pp 145-171.

[16] A.A. Ross, K. Nandakumar, A. Jain, *Handbook of Multibiometrics*, International Series on Biometrics, Springer US, 2006.

[17] A. Nambiar, M. Taiana, D. Figueira, J. Nascimento and A. Bernardino, A Multi-camera video data set for research on High-Definition surveillance, *International Journal of Machine Intelligence and Sensory Signal Processing*, vol. 1, 2014, pp 267-286.