A Context-Aware Method for View-Point Invariant Long-Term Re-identification

Athira Nambiar(✉) and Alexandre Bernardino

Institute for Systems and Robotics, Instituto Superior Técnico,
Av. Rovisco Pais, 1, 1049-001 Lisbon, Portugal
{anambiar,alex}@isr.tecnico.ulisboa.pt

Abstract. In this work, we propose a novel context-aware framework towards long-term person re-identification. In contrast to the classical context-unaware architecture, in this method we exploit contextual features that can be identified reliably and guide the re-identification process in a much faster and accurate manner. The system is designed for the long-term Re-ID in walking scenarios, so persons are characterized by soft-biometric features (*i.e.*, anthropometric and gait) acquired using a Kinect^{TM} v.2 sensor. Context is associated to the posture of the person with respect to the camera, since the quality of the data acquired from the used sensor significantly depends on this variable. Within each context, only the most relevant features are selected with the help of feature selection techniques, and custom individual classifiers are trained. Afterwards, a *context-aware ensemble fusion strategy* which we term as 'Context specific score-level fusion', merges the results of individual classifiers. In typical 'in-the-wild' scenarios the samples of a person may not appear in all contexts of interest. To tackle this problem we propose a cross-context analysis where features are mapped between contexts and allow the transfer of the identification characteristics of a person between different contexts. We demonstrate in this work the experimental verification of the performance of the proposed context-aware system against the classical context-unaware system. We include in the results the analysis of switching context conditions within a video sequence through a pilot study of circular path movement. All the analysis accentuate the impact of contexts in simplifying the searching process by bestowing promising results.

1 Introduction

We present a context-aware ensemble fusion framework based on soft-biometric features, for long term person re-identification (Re-ID) in-the-wild[1] surveillance scenarios. In particular, a biometric enabled person Re-ID system, leveraging two kinds of soft biometric features *i.e.*, anthropometric and gait features, is proposed. Since biometric feature extraction is strongly influenced by the viewpoint, we associate context to the viewing direction, and choose the best features

[1] 'in-the-wild' refers to the unconstrained settings.

A. P. Cláudio et al. (Eds.): VISIGRAPP 2017, CCIS 983, pp. 329–351, 2019.
https://doi.org/10.1007/978-3-030-12209-6_16

for each viewpoint (context). This is an extended version of the study the authors conducted in our previous work [1]. Building upon the same, we extrapolate the idea of '*view-point context*' analysis to the case where persons samples are non uniformly distributed in the gallery set by proposing a cross-context analysis. We also include the analysis of the challenging case of switching of contexts when walking paths change direction with respect to the camera, a condition not previously addressed in the literature.

In our approach, we use Kinect^{TM} sensor as the indoor Re-ID data acquisition device. Albeit some similar Kinect based Re-ID systems have been reported in the literature leveraging soft-biometric cues, [2–4] they are view-point dependent *i.e.*, data acquisition and the algorithm verification were carried out in a single direction (view-point) with respect to the camera. Such settings do not clearly represent general scenarios, where people walk in different directions. Hence, in order to assess the impact of view-point on Re-ID performance, as well as to use view-points as the contexts, we couldn't depend upon any of such existing datasets.

To tackle this issue, we collected a new set of data, where people were asked to walk in various directions (left lateral, left diagonal, frontal, right diagonal and right lateral) in front of Kinect. Along with this article, we also release the dataset named "KS20 VisLab Multi-View Kinect skeleton dataset", publicly available for research purposes[2]. We consider that some landmarks in an indoor space, (e.g., door entry/exit, lift location, printing and coffee machines etc.) determine the primary walking directions rather than random walking patterns. We term such predefined directional view-points as 'contexts', in this study. We hypothesize that this knowledge of strategic directions and the assignment of contexts are of great interest within the scope of Re-ID, since the camera positioning and gallery preparation. Benefit a lot from them in a realistic Re-ID surveillance scenario. In addition to that, not all the features are equally relevant in all contexts, because 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. Hence, the selection of the relevant features according to the context is also yet another interesting problem. Based on these two hypotheses, we redefine the classical Re-ID strategy by means of a novel *'context-aware ensemble fusion Re-ID framework'*, where we explicitly evaluate a context-specific feature matching criteria in Re-ID, and verify its experimental validity in a realistic scenario.

After studying the impact of context based Re-ID with baseline assumptions (equal gallery samples in all contexts as done in [1]), we further extend this study onto more realistic cases, where the gallery samples within the contexts vary. Such instances of data deficiency in some view-points are frequent in 'in-the-Wild' Re-ID scenarios. In many cases of practical interest, the number of samples per person will vary in different view-points. For example in a long

[2] More details on KS20 VisLab Multi-View Kinect skeleton dataset is available in the laboratory website http://vislab.isr.ist.utl.pt/vislab_multiview_ks20/.

corridor, the data acquisition will capture more samples of the person, whereas in an entry/exit point, this number may be smaller. Thus, a test sample from corridor sequence will have a large number gallery samples to match against, whereas for a test sample from the door sequence, the number of gallery samples will be very few. Hence, in order to cope with this issue and to find a solution by exploiting also the samples from other contexts, we propose a methodology called *'cross-context'*, wherein a learned feature mapping among various contexts can improve the results.

We include in this work the analysis of the challenging case of *'switching of contexts'* where the direction of person's walk keeps varying in the video sequence. We analyse a circular path movement as an instance of such a context-switching scenario. Detailed analysis of all these topics will be explained in the forthcoming sections. The major contributions of the paper are enumerated as follows:

- Public release of a new dataset with 20 people walking in 5 different directions acquired from Kinect^{TM} v.2, suitable for long-term pose-invariant Re-ID, named "KS20 VisLab Multi-View Kinect skeleton dataset".
- Proposal of a 'Context-aware ensemble fusion Re-ID framework' where different context specific classifiers are trained via adaptive selection of the potentially relevant features in each context.
- Proposal of 'Cross-context analysis', in order to cope with data deficient cases in the gallery contexts, and to improve the Re-ID performance via feature mapping.
- A pilot study of the 'context-switching' test case, by experimenting people walking along circular path (changing contexts) and conducting Re-ID trained with KS20 dataset.

The rest of the paper is organized as follows. The related works are described in Sect. 2. The proposed methodology is explained in Sect. 3, *i.e.* feature extraction method, Context-aware ensemble fusion framework and Cross-context analysis via feature mapping. In Sect. 4, our dataset and experimental results are discussed in detail. Finally, the summary of the paper and some future plans are enumerated in Sect. 5.

2 Related Works

Many Kinect based Re-ID works were reported in the literature in the last few years. The major advantage of such proposals were the incorporation of soft-biometric cues by exploiting the depth info and skeleton joints. This enabled the Re-ID paradigm to extend towards long term scenarios, from the traditional short-term scenarios which leverage primarily appearance cues (colour or texture).

One of the earlier works [2], proposed a specific signature to be built for each subject from a composition of several soft biometric cues (*e.g.*, skeleton and surface based features) extracted from the depth data. Then, Re-ID was

accomplished by matching these signatures against the gallery samples. Kinect based person re-identification from soft biometric cues was also addressed in another work by [5], leveraging skeleton descriptors (by computing several limb lengths and ratios) and shape traits (using point cloud shape). In some other recent works in [3] and [4], both static anthropometric features and dynamic gait features were employed towards Re-ID tasks. Nevertheless, in all of those methodologies, data acquisitions were conducted in a constrained manner *i.e.*, in a particular view-point. In this work, we build upon the aforementioned state-of-the-art works but in less constrained conditions by imposing an 'in-the-wild' indoor Re-ID scenario with various viewpoints and by exploiting relevant features in each of those view-points (contexts).

Many definitions of context were encountered in the literature, depending on the field of application. The dictionary definition of context is "*the surroundings, circumstances, environment, background or settings that determine, specify, or clarify the meaning of an event or other occurrence*" [6]. In our work, we deem context as the view-point setting, under which features are computed. The concept and application of context were reported in various fields, for instance, in customer behaviour applications [7], where the context was viewed as the intent of a purchase (*e.g.* context of a gift). In [8], subject re-identification has been conducted exploiting instant messaging in a web surfing navigation. The context used in that work was the special characteristics of chatting text (*e.g.* content, syntax and structural based features). In [9] context (time of the day/location where digital data created) was used for online customer re-identification towards customer behaviour model analysis. The concept of context in terms of predictable and unpredictable image quality characteristics was presented in the traffic monitoring research area in [10]. In [11], both the scene context (environment of the subject at global and local levels) and group context information (activity interaction of subject with group members) were exploited towards activity recognition.

In the person Re-identification paradigm, few works addressed the concept of context. The work of [12] 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 [13,14]. In those works, in addition to the content information of the subjects, they also leveraged context information (k-common nearest neighbors) to fine tune the Re-ID results. From our literature review, context is found to be a new tool whose effectiveness in Re-ID applications is yet to be completely explored.

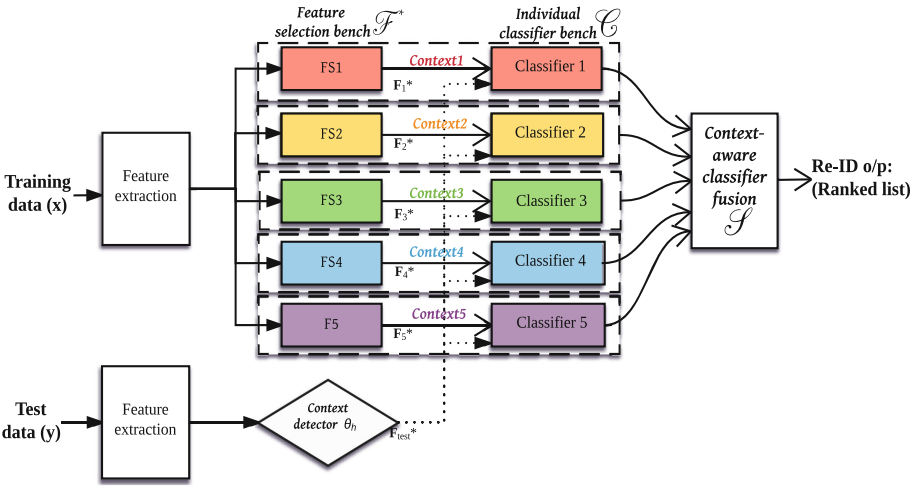


Fig. 1. Context-aware ensemble fusion system consist of a feature extraction module, feature selection context bench, individual classifier bench, a context detector module and a classifier fusion module. The individual classifiers for each context are trained using individual feature subspace ensembles $\mathbf{F_j^*}$, 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.

3 Methodology

In this section, we explain the proposed methodology *i.e.*, various stages of data analysis including feature extraction, context-aware ensemble fusion framework and cross-context mapping.

3.1 Feature Extraction

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 1 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 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 relative 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 1, the dimension of features derived are shown within parenthesis.

Table 1. 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*)

3.2 Context-Aware Ensemble Fusion

One key contribution of our work is the proposal of context-aware ensemble fusion Re-ID framework. As mentioned earlier, we experiment the impact of the different data features along different contexts *i.e.*, view-points, and then employ a context-based fusion method to obtain the final Re-ID result. We refer the work on feature subspace ensembles [15] to be a motivation to the authors to come up with a homogeneous ensemble fusion strategy. That work presented an approach to execute multiple parallel feature selection stages leveraging different training conditions, so as to obtain the best features, by using majority voting of the feature ensembles.

Our proposed framework is shown in Fig. 1. After the feature extraction is carried out, four further modules constitute the system: (i) *Feature selection Context bench* (ii) *Individual classifier bench*, (iii) *Context detector module* and (iv) *Context-aware classifier fusion module*.

Feature Selection Context Bench. Refering to Fig. 1, we illustrate our method with five context view-points as $Context1, ..., ContextN$, with $N = 5$. After the features are extracted from the training data within the feature extraction module, the feature descriptors are customized for individual contexts. This usage of 'right-data in the right context' is one of the main advantage of our framework in contrast to the classical approaches. This enables the data to be split and stored in an organized manner (according to context) within the gallery.

Each context module is internally built of a feature selection bench and an individual classifier bench. The former module analyses the feature vectors

entered into each context, by means of a feature selection scheme and retain only the most discriminative and relevant features. Specifically, we use the popular Sequential Forward Selection (SFS) algorithm [16] as an instance of Feature Selection (FS). It works iteratively by adding features to an initial subset, seeking to improve the Re-ID measurement. 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 $\mathbf{F} = (f_1, \cdots, f_d)$. FS analyses this d-dimensional space in order to identify the potentially relevant features $f_i \subset \mathbf{F}$, and discard the rest according to some feature subspace evaluation criteria $\boldsymbol{J}$ and ultimately derive $\mathbf{FS_j^*}$, which is the set of most relevant features for context j. Thus, the outputs of the Feature selection context bench consists of an ensemble of feature subspace *i.e.*, the features selected for each particular context $\mathcal{F}^* = (\mathbf{FS_1^*}, \cdots, \mathbf{FS_5^*})$.

Specifically, the Sequential Forward Selection (SFS) algorithm works as following: It begins from an empty feature set $\mathbf{FS_{t=0}^*}$. At each step, $\mathbf{FS_{t+1}^*}$ all possible super-spaces containing the most relevant feature subspace in the previous step, $\mathbf{FS_t^*}$, and one from the remaining features $f_i \in \mathbf{F} \backslash \mathbf{FS_t^*}$ 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 any super-space formed at a given step $\mathbf{FS_{t+1}^*}$ does not improve $\boldsymbol{J}$, the search stops and the subspace $\mathbf{FS_t^*}$ is considered as the best feature subset. At last, the outputs of the Feature selection context bench consist of an ensemble of feature subspaces *i.e.*, the features selected for each particular context $\mathcal{F}^* = (\mathbf{FS_1^*}, \cdots, \mathbf{FS_5^*})$. For the implementation of the algorithm, the authors used SFS package[3] [17]. We used 1NN classifier with an Euclidean neighborhood metric in the SFS scheme.

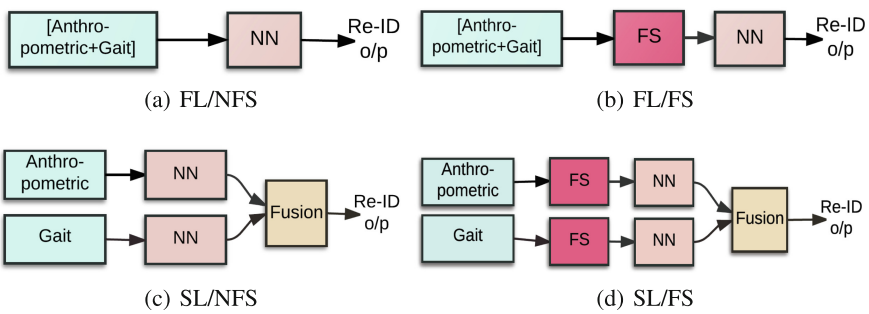


Fig. 2. Various fusion-feature selection schemes employed in this work. Top and bottom rows represents feature-level and score-level fusion strategies respectively. Feature selection (FS) is not used in case studies (a) FL/NFS and (c) SL/NFS, whereas (b) FL/FS and (d) SL/FS shows the inclusion of FS module.

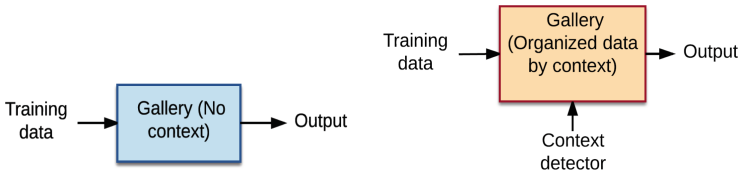
[3] http://users.spa.aalto.fi/jpohjala/featureselection/.

Individual Classifier Bench. After the feature selection phase, the selected most relevant features were leveraged to train each individual classifier. These sets of potential features consist of both anthropometric and gait features. In order to understand how to fuse all these various features during the training of classifiers, we use two popular traditional approaches *viz.*, feature level fusion and score level fusion. Also, the impact of feature selection module is verified at this stage by enabling and disabling the FS bench. Hence, four fusion-feature selection case studies are carried out: (i) Feature-level fusion without FS (FL/NFS), (ii) Score-level fusion without FS (SL/NFS), (iii) Feature-level fusion with FS (FL/FS) and (iv) Score-level fusion with FS (SL/FS). The schematic representations of all these cases are depicted in Fig. 2.

In feature level fusion (see Fig. 2 top row), 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 identity of an individual. Instead, in score level fusion (see Fig. 2 bottom row), 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 [18], 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.

In all the case studies mentioned here, a leave one out evaluation strategy is performed within each context, with a Nearest neighbour (NN) classifier using euclidean distance metric. The experimental results obtained are explained in Sect. 4.2, 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 $\mathcal{C} = [\textbf{Classifier1}, \cdots, \textbf{Classifier5}]$.

Context Detector. Context detector is the module where the context (viewpoint) of the test sample is estimated. This module is one of the most distinguishing components of our proposed context-aware Re-ID system, in comparison to the classical context-unaware Re-ID system. The holistic overview of both the classical vs. proposed systems is depicted in Fig. 3. In the classical approaches, no notion of *Context* is enabled, so that all the data in the gallery has to be used in the person matching procedure. Instead, in our proposed context-aware system, data are organised according to context. In addition to that, the *Context-detector* module determines the context of the probe sample and thus redirects the system to the corresponding gallery context in order to facilitate a faster and more accurate person matching.



(a) Classical (b) Ours

Fig. 3. Classical (context-unaware) vs. proposed context-aware systems.

The design of the context detector module was carried out by analysing any torso joint over a gait cycle[4]. Then, the direction of walking is estimated by analysing the direction of the joint vector. Suppose $\boldsymbol{h_{begin}}$ and $\boldsymbol{h_{end}}$ denote the position of the joint in the first frame and last frame respectively. Then the directional vector among these frames $\boldsymbol{h} = <h_x, h_y, h_z>$ is obtained as follows:

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

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

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

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.

Context-Aware Classifier Fusion. Classifier fusion module performs a context-specific adaptive fusion of the results obtained at the outputs of individual classifiers $\mathcal{C} = [\textbf{Classifier1}, \cdots, \textbf{Classifier5}]$, based on the result from context detector. We propose a score-level fusion based on context termed as '***Context-specific score level fusion***', which can be considered 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 [18]. In a similar way, in our proposal, we endorse adaptive weights to the scores from different classifiers according to its context, in order to increase the influence of more reliable context. We employed linear interpolation technique as an instance of the adaptive weighting scheme.

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.

[4] We used 'SpineShoulder' *i.e.*, the base of the neck refering to joint number 20 of Kinect™ v.2 (https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx) as the torso joint towards context detection, since it remains more or less stable while walking.

Suppose the context lies in between two neighbour pre-defined context views *i.e.*, $\mathbf{v_i}$ 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 $\mathcal{S}$ is computed as a weighted sum of those scores as follows:

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

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 a single context is activated, η of the nearest context turns to be 1, and all the others will be 0. Various case studies on this concept are analysed in detail, in the experimental Sect. 4.3.

3.3 Cross-Context Analysis

After proposing our 'Context-aware ensemble fusion framework', we also propose a special case of the scenario, where the number of subject samples varies among different contexts. This can be considered as an extended case of the baseline contextual analysis, where, in addition to the training of individual contexts, we also train combination of contexts. In detail, in the baseline scenario, we assume equal number of samples per person per view-point. This always enables the context-aware system for a particular test sequence, to search for the matching gallery sample in the very same context. However, in practical scenarios, the gallery samples differ among various contexts. In order to overcome such situations, we propose 'Cross-context analysis'. Here, even if the system lacks gallery samples of the test person in the very same context, it can search at other contexts as well, where the number of gallery samples are higher than the same context. This is realized via a feature mapping technique. Feature mapping learns the set of relevant features in a particular gallery context, given the same/different context as the test context. Based on this idea, we analyse which are the features of interest in Context B, given the test sample in Context A. Feature mapping among 5 contexts results in 25 various FS sets. A pictorical representation of the proposal is shown in Fig. 4.

In order to better understand the proposed concept, we conducted mainly three case studies: (i) 5 cross-context case known as *Full cover gallery* (ii) 4 cross-context case known as *Sparse cover gallery* and (iii) 1 cross-context case known as *Single cover gallery.*

- **5 cross-context (Full cover gallery)** is the case where all contexts of all subjects are represented in the gallery. Or in other words, we have the probe person in all the five context galleries.
- **4 cross-context (Sparse cover gallery)** is the case, where each subject is represented in many contexts but not exactly the one of the test. In other words, we remove the test person from the same context and thus only the matching person data samples available in the gallery are from other 4 different contexts.

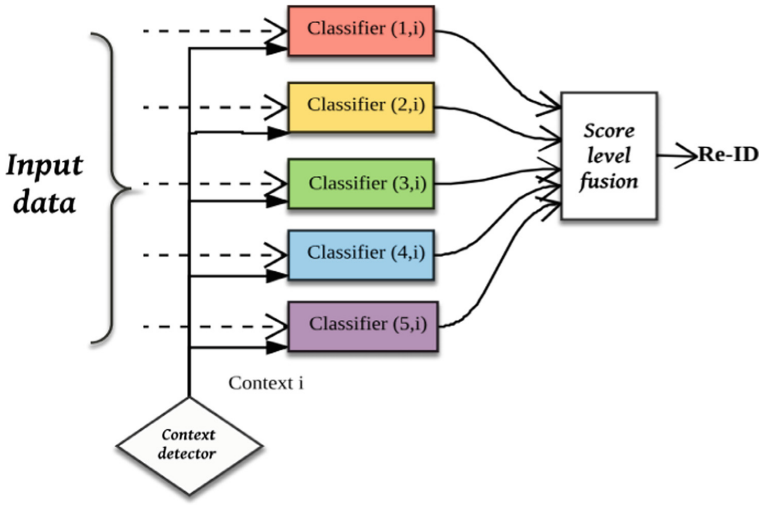


Fig. 4. A schematic overview of cross-context analysis. Individual cross-context classifiers are trained based on the learned feature mapping between the probe context i, and the gallery context of interest. Then, based on score level fusion overall Re-ID result is given as output.

– **1 cross-context (Single cover gallery)** is the case where each subject appears in a single context in the gallery, different from the probe *i.e.*, we remove the test person samples from all the contexts except a random context (other than the probe context).

4 Experiments and Results

In this section, we describe the various experiments conducted as a part of this study, and the results and related observations are explained in detail. First of all, we present our new dataset which we collected in connection with this study, named 'KS20 VisLab Multi-View Kinect skeleton dataset'[5]. We make it publicly available to the community for extending this line of works. Further, we present the performance analysis of our context-aware system and its extension towards cross-context analysis. Four major experiments were carried out in this regard. *(A) Training of the individual context-specific classifiers*, where each classifier model is learned based on respective context; *(B) Contextual analysis*, where the Re-ID system takes into account the context information of the scenario and thus significantly reduces the search space, *(C) Cross-context analysis*, where the issue of sample deficiency in the same context is tackled also by leveraging different

[5] KS20 VisLab Multi-View Kinect skeleton dataset: http://vislab.isr.ist.utl.pt/vislab_multiview_ks20/. Access to the Vislab Multi-view KS20 dataset is available upon request. Contact the corresponding author if you are interested in this dataset.

contexts via feature mapping technique, *(D) Switching of contexts*, wherein a circular path walking test scenario is analysed to verify the Re-ID performance of our proposed system.

4.1 Dataset

In order to employ Re-ID in a realistic 'in-the-wild' scenario, it is essential to have a challenging unconstrained dataset, comprised of the sequences of people walking in different directions. Since a Kinect^{TM} based dataset with different viewangles was unavailable, we acquired our own dataset, in the host laboratory, named 'KS20 Vislab Multi-view Kinect Skeleton dataset'. It is a set of multi-view Kinect skeleton (KS) data sequences collected from 20 walking subjects using Kinect V.2., in the context of long-term person re-identification using biometrics. Multiple walking sequences along 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 skeleton image sequences comprising 20 subjects (3 video sequences per person in a particular viewpoint) in the aforementioned directions.

Regarding the data acquisition, the Kinect sensor was kept at a height of an average human (See Fig. 6(a) for the data acquisition system). This simulates a normal video surveillance environment as well as changes in the position of camera over time, as in a long term 'in-the-wild' person Re-ID scenario. The position of camera as well as the walking directions of subjects were deliberately altered in order to ensure a typical surveillance scenario. 20 people within the age group 23–45, including 4 ladies and 16 men participated in the data collection. The statistical details of the people *i.e.*, age, gender, height and weight are highlighted in the Table 2. All of them were asked to walk in their natural gait, in front of the camera three times each along each direction. No markers were provided to determine the path, instead only approximate direction was instructed. The visualization of the existing five contexts in our dataset is given in Fig. 5 by plotting how 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 cluster means $\mu = [1.67, 35.63, 92.83, 130.70, 180.17]^\top$ degrees with standard deviations $\sigma = [3.64, 4.90, 3.29, 5.34, 3.99]^\top$ degrees. Different walking directions and sample video frames extracted from our dataset, are shown in Fig. 6.

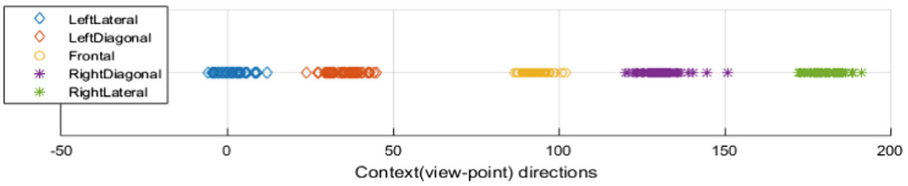


Fig. 5. Distribution of the contexts in the dataset (note: directions are in degrees.).

Table 2. Characteristics of people involved in KS20 Vislab Multiview Kinect Skeleton dataset. This table contains the statistics of 18 people, since two people were reluctant to provide the details.

	Mean	Standard deviation	Minimum	Maximum
Age	31.72	6.08	23	45
Height (cm)	174.78	8.17	160	185
Weight (kg)	72.11	11.60	51	95

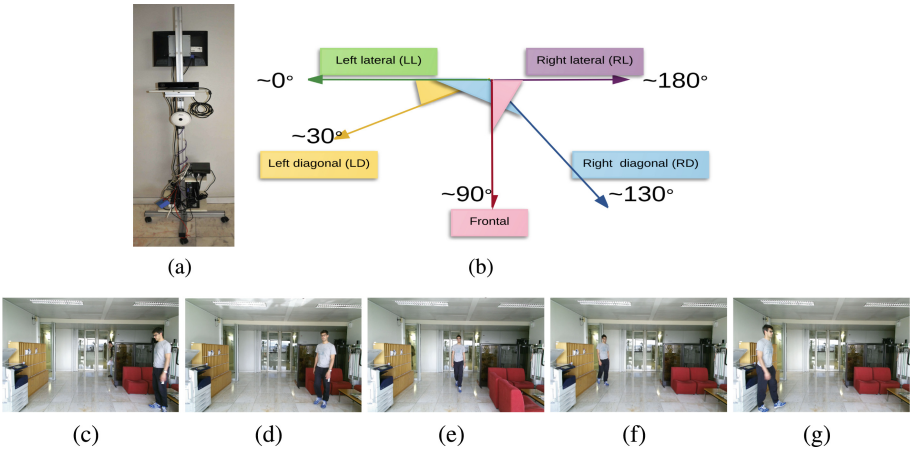


Fig. 6. Data acquisition: (a) subject walking directions in front of the camera system (direction angles are defined with respect to the image plane.) (b) Data acquisition set up (c–f) sample frames from our data acquisition, in five different directions- left lateral (∼0°), left diagonal (∼30°), frontal (∼90°), right diagonal (∼130°) and right lateral (∼180°) respectively.

Kinect™ sensor device is composed of a set of sensors, which is accompanied with a Software Development Kit (SDK), and can track movements from users by using a skeleton mapping algorithm that provides the 3D information related to the movements of body joints[6]. We acquired all the three available data *i.e.* skeleton, colour and depth. The skeleton data contains the position and orientation of 25 joints of the human body and was captured at the full frame rate of the sensor @ 30 fps. Colour and depth information are employed for appearance based features, which generally require single frame, and hence was captured at 1 fps. However, these were not used in the current work[7].

[6] For body joint types and enumeration, refer to the link: https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx.

[7] In the publicly available dataset also, only the skeleton data is provided. Nevertheless, color and depth information can be made available on demand.

Prior to the analysis, we had to pre-process the data to remove the noise contents in the data. We had discussed the preprocessing and feature extraction phases in detail in the prior work [19]. Usually, the primary effect of noise are jerks/abnormalities in the skeleton data, during the sequences. In addition to that, the skeleton is not detected in some frames, especially in the boundary of the kinect range. In order to tackle such situations, we use a semi-automatic approach to select the best frames to retain in the video sequence. By empirically analysing the evolution of lower body angles over time, we cleared the unwanted jerks in the signals. In particular, by observing the measurements of hip angle over the sequences, we noticed that the jerks made these angles increase abnormally, which results in drastic variations in the corresponding signals (see Fig. 7(a)). In order to clean/remove such unwanted frames, we assign some thresholds upon the angular values, and thus only the valid data signals are being selected (see Fig. 7(b)). Afterwards, based on those cleaned signals, the functional units of gait *viz.*, gait cycles, were estimated. A gait cycle is comprised of sequence of events/movements during locomotion from the point 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 cycles (see Fig. 7(c)) and various features were extracted within this gait period.

4.2 Training of the Individual Context-Specific Classifiers

This experiment is quite analogous to the one the authors conducted in the previous work [1], where we analysed the performance of best features among 74 features *i.e.*, feature subset selected via feature selection. Albeit we carried out similar analysis in the aforementioned paper, herein we have used some different features *i.e.*, relative joint positions instead of absolute joint positions. Based on all these 74 features, we conduct an extensive analysis of various fusion-Feature selection schemes, as mentioned in Sect. 3.2: (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 resulting Re-ID performance as well as the corresponding cumulative matching rank scores (showing overall CMC rank-1) are shown in Fig. 8 and Table 3 respectively.

Results highlight that: (i) Feature selection (FS) outperforms the cases without FS (NFS). (ii) Score-level fusion performs better than the feature level fusion in Re-ID. (iii) SL/FS is found to be 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.

Thus, we choose SL/FS as the feature selection scheme for the remaining of this work. This context-aware feature selection criteria resulted in the selection of best features in respective contexts as shown in Table 4. From those customized features, we can observe that some global discriminative anthropometric features such as height, arm length, chest size are highly relevant in almost all the contexts. However, certain features clearly show its affinity towards certain contexts, for e.g., vertical movements of joints associated to gait features

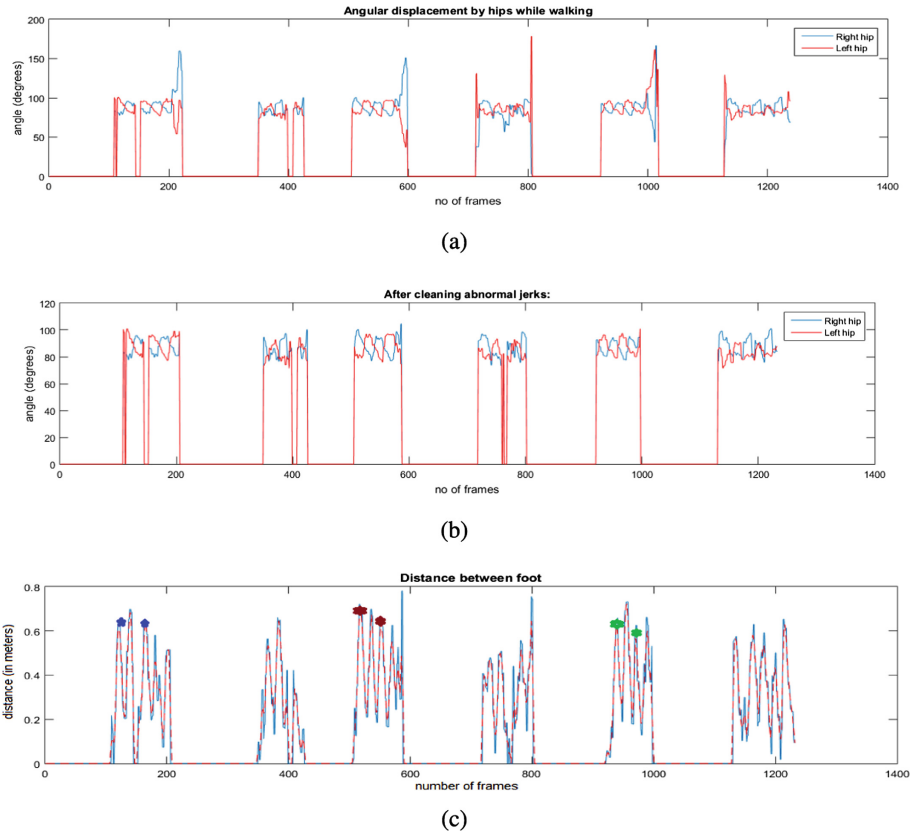


Fig. 7. (a) The abnormal transients at the ends of each sequence are due to the jerks of skeleton occurring at its respective frames; (b) after obtaining the cleaned frames, by filtering the abnormal frames; (c) gait cycle estimation. Three consecutive peaks (two adjacent markers) within a sequence, represent a gait cycle.

Table 3. Chart showing the Re-ID accuracy rates for five contexts at rank-1 CMC. The highest and second highest Re-ID rates observed are highlighted in **bold** and *italic* letters, respectively.

Context	FL/NFS	FL/FS	SL/NFS	SL/FS
Left lateral	68.33	**90.00**	83.33	*88.33*
Left diagonal	55.00	76.67	**81.67**	*78.33*
Frontal	81.67	91.67	*93.33*	**95.00**
Right diagonal	65.00	*81.67*	78.33	**85.00**
Right lateral	68.33	*86.67*	*86.67*	**88.33**
Average for all contexts	67.66	*85.34*	84.6	**86.99**

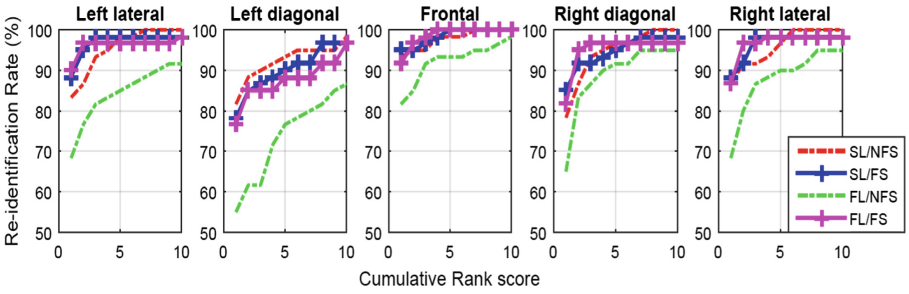


Fig. 8. The Re-ID performances of various fusion-FS schemes mentioned in Fig. 2 along five contexts *viz.*, left lateral (∼0°), left diagonal (∼30°), frontal (∼90°), right diagonal (∼130°) and right lateral (∼180°) respectively. Cumulative matching scores up to 10 subjects are shown.

(hipY$_{\mu,SD}$, handY$_{\mu,SD}$, ankleY$_{\mu,SD}$) are found to be selected in the lateral/diagonal contexts, whereas the limb distances (handDist$_{\mu,SD}$, elbowDist$_{\mu,SD}$) are found to be selected in the frontal context.

4.3 Contextual Analysis

This experimental analysis is to verify the overall performance of the proposed context-aware system against the baseline classical context-unaware systems. In the former *i.e.*, *Context-aware*, 1-context scenario and 2-contexts scenario are carried out. 1-context case is where the system will automatically select the nearest gallery context and search for the best match whereas the 2-context scenario is where the system will select the two neighbouring contexts and carry out a linear interpolation technique (via adaptive weighted sum), in order to re-identify the person. (see Sect. 3.2 for further details on Context-aware Re-ID paradigm). The latter case is the baseline scenario *i.e.*, *Context-unaware*, where we disable the context detector module, and hence no notion of the probe context is available to the system. We call this case as *'Pure'* baseline, since no notion of context has been considered even during the classifier training phase. Instead, feature selection has been done globally irrespective of any context, and the same features got selected globally thus making the FS in all the samples context-unaware. Then, the test sample is matched against all those gallery samples.

The results are presented in Table 5. Results clearly shows the outperformance of the *Context-aware* system against *Context-unaware* system. It is notable that context-aware methods (either by using a single or two contexts) bestow high performance level about 88%, whereas Context-unaware approaches 78%. Also, context-aware systems performed faster (6–11 s) compared to the context-unaware system (20 s), since the notion of Context helped the reduction of the search space and speed up the matching process. Hence, the knowledge of context is found to be vital in augmenting the performance of a Re-ID system in terms of both speed and accuracy.

Table 4. Context-specific features selected via SL/FS scheme, during the training of individual context classifiers.

Feature	LL	LD	F	RD	RL
height	✓	✓	✓	✓	✓
arm	✓	✓	✓	✓	
upper	✓			✓	
lower		✓		✓	✓
ULratio		✓		✓	
chestsize		✓	✓	✓	✓
hipsize	✓		✓	✓	
kneeAngle		✓		✓	
$\text{kneeDist}_{\mu,\text{SD}}$	✓	✓	✓	✓	
elbowDist_{μ}			✓	✓	
$\text{elbowDist}_{\text{SD}}$	✓	✓		✓	
headY_{μ}		✓		✓	
headY_{SD}	✓				
$\text{rhipY}_{\mu,SD}$	✓	✓		✓	
$\text{lhipY}_{\mu,SD}$	✓	✓			✓
lkneeY_{μ}					✓
$\text{ankleY}_{\mu,SD}$	✓			✓	
lhandY_{μ}			✓		
$\text{lhandY}_{\text{SD}}$	✓			✓	
$\text{rhandY}_{\mu,SD}$					✓
lshouldY_{μ}			✓		
$\text{handDist}_{\mu,\text{SD}}$			✓		
$\text{lshouldY}_{\text{SD}}$			✓		

Table 5. 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 (pure baseline)	1 context (binary weights)	2 contexts (adaptive weights)
Anthropometric	60.33%	68.67%	68.00%
Gait Re-ID	72.33%	80.67%	80.67%
Overall Re-ID	78.33%	88.00%	88.67%
Processing time	25.14 s	6.176 s	11.63 s

4.4 Cross-Context Analysis

Refering to Sect. 3.3, we also conduct an extension of the baseline context-aware framework, with the difference that the number of subject sample varies among different contexts. We carried out three case studies: (i) 5 cross-context case known as *Full cover gallery*, where the test sample is compared against all the remaining 299 data samples in all the 5 contexts, (ii) 4 cross-context case known as *Sparse cover gallery*, where each test will be compared against 297 data samples in the gallery, and (iii) 1 cross-context case known as *Single cover gallery*, where the test sample has to be matched against 288 samples.

Now, for each of the aforementioned cases, five matching techniques are performed: (a) No FS (b) Pure baseline (c) 1-nearest context (d) 2 neighboring contexts and (e) Cross-contexts. Method 'no FS' doesn't consider any Feature selection criteria, thus the matching will be the basic feature matching of 74D feature vectors in all the gallery contexts. The second method (Global FS) conducts feature selection globally upon the whole set of data. This is the pure baseline analysis mentioned in Table 5. Then, upon the selected feature set, it carries out the feature matching. Since both of these cases don't consider the notion of context, they are categorized under *Context-unaware* paradigms. The latter ones *i.e.* (c) 1-nearest context (d) 2 neighboring contexts and (e) Cross-contexts, execute feature selection and context-aware Re-ID. In both (c) and (d), baseline context-aware framework is considered whereas in (e), the cross-context technique is exploited.

Table 6. Chart showing the Re-ID accuracy rates of cross-context analysis. Full, sparse and single cover gallery cases with different feature selection schemes *i.e.*, no FS, Global FS and Customized FS are shown. The accuracy rates shown in each cell represents Rank-1 CMC rate (in percentage).

	Context-unaware		Context-aware		
	No FS	Global FS	1-context	2 contexts	Cross-context
(i) Full cover (5 contexts)	74.67	78.33	88.00	88.67	82.33
(ii) Sparse cover (4 contexts)	28.00	41.67	x	x	44.33
(iii) Single cover (1 context)	8.33	12.67	x	x	18.33

The results for the aforementioned cases are reported in Table 6. The primary observation made out of the results is that the context-aware cases always outperforms the context-unaware cases[8]. We can observe the improvement in Re-ID performance by incorporating feature selection scheme as well as context framework. It is notable that 'Pure baseline' (global FS) could improve the results compared to the 'no FS'. While exploiting contextual analysis, the best performance is reported (they are applicable only in the full-cover scenarios).

[8] 1-context and 2-contexts work only for the full cover scenario, and hence other sparse cover and single cover scenarios for the same are represented via crossmark, refering 'Not Applicable'.

However, Cross-context outperforms both 'no FS' and 'Pure baseline', in all the three gallery cover scenarios. Thus, it is confirmed that when the relevant features are selected according the context by learning the mapping of features among various contexts, it can improve the result of Re-ID. Since deficiency of samples in some viewpoints are a big challenge in realistic practical scenarios, such cross-context customized FS approach is of great interest.

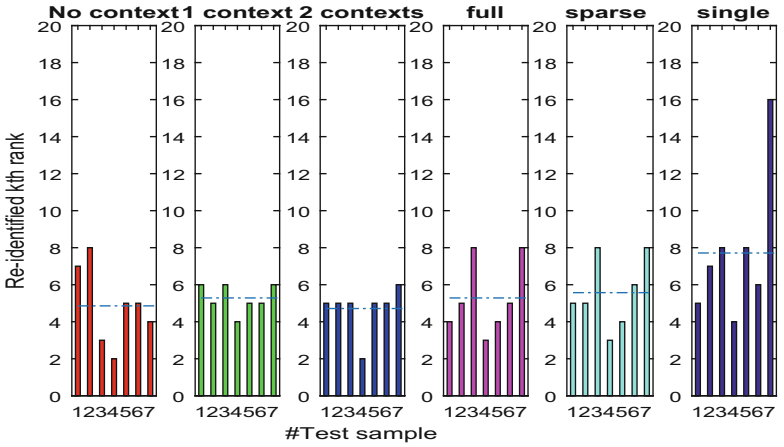
The best Re-ID performance among the 3 cases of gallery settings was observed in the Full-cover gallery context *i.e.*, 5 contexts. This may be due to the fact that there are more and better examples to match to the test sample. Sparse cover produces a bit worse results compared to the former since there is no availability in the very same context, instead it searches and finds the best matching in the four other different contexts. The worst case is where only a single cover gallery (other than the test context) is provided, where always the matching is poor in terms of the number of samples and quality of data, but still outperforms the context-unaware case.

4.5 Switching of Contexts

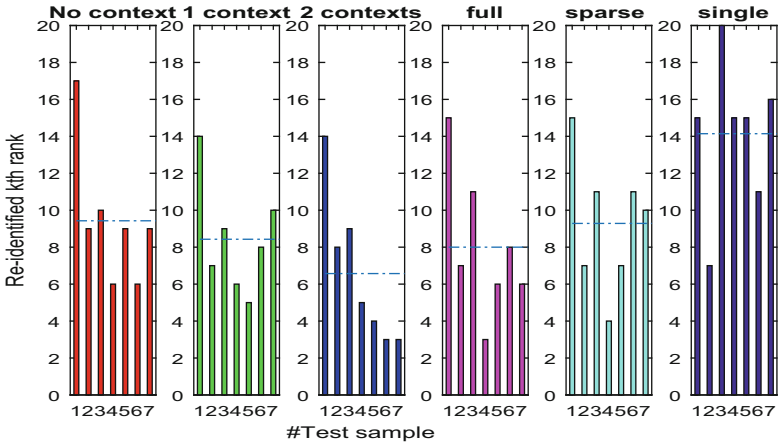
Yet another experiment we conducted as a part of this study was that of switching of contexts. This is a scenario where the person continuously changes his direction of walk, and hence the context (*i.e.*, view point) also continuously changes. We analyze this issue by considering a circular path walking. This is a pilot study in order to understand the feasibility of applying our system towards 'Context-switching' scenarios. Hence, we acquired new circle path data from two people (who belong to Vislab Multiview KS20 dataset), and try to match them against the KS20 dataset sequences which where collected almost one year before. Two advantages of such an acquisition were: (i) This makes a perfect long-term Re-ID validation system since collected with a gap of long duration and (ii) good for the analysis of varying context scenario.

In this experiment, we asked the people to walk in front of Kinect sensor in circular paths. Either three or four complete walking sequences were recorded. For the processing, we cleaned the data, and then segmented the data to separate gait cycles, as described in Sect. 4.1, with the assumption that, within a gait cycle, the person is walking in a linear path. Hence, a complete circular path contains five or six gait cycles. Note here that, since in our training of the data we used only the directions towards the camera, we ignore the gait cycles where the person is walking away from the camera. Thus, out of a single circular path walking, we extract either 2 or 3 gait cycles. Ultimately, we succeeded in making 7 gait cycles extracted out of whole sequences of walking.

We show the results of Re-ID performance of switching contexts in 2 mode settings of the gallery samples: (i) Complete gallery and (ii) Incomplete gallery. In the former, gallery is provided with sufficient set of samples, and thus we analyse the Re-ID as we conducted Contextual analysis (Sect. 4.2). In the latter, we assume the practical scenario of deficiency of gallery samples, and thus we analyse Re-ID as we conducted Cross-contextual analysis (Sect. 4.3). The corresponding results are shown in Fig. 9. Each diagram shows the k-th rank



(a) Person#1



(b) Person#2

Fig. 9. A case study of "switching the context", carried out upon two persons are depicted above. In each row, first three diagrams result from contextual analysis (complete gallery samples) and the last three diagrams result from cross-contextual analysis (incomplete gallery samples). The mean k-Rank in each case is marked via dash-dot lines. The best results (lower mean k-Rank) were found in 2-Contexts case scenario, in both persons.

at which the person is correctly re-identified, hence, lower the rank, better the performance. We can observe in both cases that, the Contextual analysis outperforms the Cross-contextual analysis, which clearly accentuates the importance of having good enough number of samples in the gallery set. Within the contextual analysis, the best Re-ID performance is reported in 2-context case, exploiting linear interpolation technique(adaptive weighted sum). Person#1 and #2 are

respectively a woman and a man. Considering the relative population of women and men in the dataset (16 men and 4 ladies), better Re-ID was observed for Person#1 (lower k-th rank), implying that Re-ID of the lady candidate was much easily done compared to the man candidate.

5 Conclusions

In this work, a context-aware person re-identification system named 'Context-aware ensemble fusion Re-ID framework', and its extension towards Cross-context analysis have been discussed. As a part of the study, we acquired a new multi-view Kinect skeleton (KS) dataset, containing 300 data sequences collected from 20 subjects, using Kinect^{TM} v.2. We make the dataset publicly available to the research community as one of the contributions of this paper, under the name '*KS20 VisLab Multi-View Kinect skeleton dataset'*.

We conducted extensive study on the impact of various anthropometric and gait features upon person Re-ID. Since certain features have upperhand in specific view-points, we associate context to the viewing direction of walking people in a surveillance scenario and choose the best features for each case. Such a Context-aware proposal exploiting view-point as the contexts is one of the very first of that kind in the Re-ID literature. Building upon our previous works in the same area [1], we analysed various fusion schemes (Score level vs. Feature level) and feature selection (Sequential Forward Selection), we could observe the *Score level fusion with Feature Selection* schemes works the best among all of them and is selected as the de-facto standard for our framework. Other major contributions of the framework are context detection module and context-aware classifier fusion technique. The experimental results of the holistic Re-ID system performance shows that Context-aware system works faster (upto 4 times) and accurate (up to 10% point better) compared to the context-unaware system.

Some other major extension studies were also conducted in this work. First one was cross-context analysis, in order to overcome the practical limitation of gallery data deficiency in the same context. The proposed cross-contextual paradigm enables a feature mapping technique with which the best features could be learned among different contexts, and hence the probe can search and find the best matching even in different contexts. Results show that cross-context beats context-unaware cases. Among the context aware methods, the cross-context is the only applicable to cases of incomplete gallery, eventhough the 1 context and 2 context methods are the best in the full gallery cases. Another very interesting experiment was the context-switching, where the person keeps on changing the direction. In order to validate Re-ID in such scenarios, we exploited a circular path walking for 2 people as a pilot study, and tested againsted KS20 data gallery. Among various cases, 2-neighboring context (context-aware Re-ID) method performed the best. In the future works, we envisage to incorporate multiple contextual features (*i.e.*, view-point, distance to the camera, occurrence of face, person co-occurrence etc.), as well as to learn contexts automatically (e.g., data clustering).

References

1. Nambiar, A., Bernardino, A., Nascimento, J.C., Fred, A.: Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In: B-Wild Workshop at 12th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 973–980 (2017)
2. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with RGB-D sensors. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 433–442. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_43
3. Gianaria, E., Grangetto, M., Lucenteforte, M., Balossino, N.: Human classification using gait features. In: Cantoni, V., Dimov, D., Tistarelli, M. (eds.) International Workshop on Biometric Authentication, pp. 16–27. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13386-7_2
4. Andersson, V.O., de Araújo, R.M.: Person identification using anthropometric and gait data from kinect sensor. In: AAAI, pp. 425–431 (2015)
5. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Gong, S., Cristani, M., Yan, S., Loy, C.C. (eds.) Person Re-Identification. ACVPR, pp. 161–181. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4_8
6. Context definition. https://en.wiktionary.org/wiki/context
7. Palmisano, C., Tuzhilin, A., Gorgoglione, M.: Using context to improve predictive modeling of customers in personalization applications. IEEE Trans. Knowl. Data Eng. **20**, 1535–1549 (2008)
8. Ding, Y., Meng, X., Chai, G., Tang, Y.: User identification for instant messages. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011. LNCS, vol. 7064, pp. 113–120. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24965-5_13
9. Panniello, U., Hill, S., Gorgoglione, M.: Using context for online customer re-identification. Expert Syst. Appl. **64**, 500–511 (2016)
10. Alldieck, T., Bahnsen, C.H., Moeslund, T.B.: Context-aware fusion of RGB and thermal imagery for traffic monitoring. Sensors **16**, 1947 (2016)
11. Wei, L., Shah, S.K.: Human activity recognition using deep neural network with contextual information. In: International Conference on Computer Vision Theory and Applications, pp. 34–43 (2017)
12. Zhang, L., Kalashnikov, D.V., Mehrotra, S., Vaisenberg, R.: Context-based person identification framework for smart video surveillance. Mach. Vis. Appl. **25**, 1711–1725 (2014)
13. Leng, Q., Hu, R., Liang, C., Wang, Y., Chen, J.: Person re-identification with content and context re-ranking. Multimedia Tools Appl. **74**, 6989–7014 (2015)
14. Garcia, J., Martinel, N., Micheloni, C., Gardel, A.: Person re-identification ranking optimisation by discriminant context information analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1305–1313 (2015)
15. Silva, H., Fred, A.: Feature subspace ensembles: a parallel classifier combination scheme using feature selection. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 261–270. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72523-7_27
16. Whitney, A.W.: A direct method of nonparametric measurement selection. IEEE Trans. Comput. **C-20**(9), 1100–1103 (1971)
17. Pohjalainen, J., Räsänen, O., Kadioglu, S.: Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. Comput. Speech Lang. **29**, 145–171 (2015)

18. Ross, A.A., Nandakumar, K., Jain, A.: Handbook of Multibiometrics, vol. 6. Springer, Heidelberg (2006). https://doi.org/10.1007/0-387-33123-9
19. Nambiar, A., Bernardino, A., Nascimento, J.C., Fred, A.: Towards view-point invariant person re-identification via fusion of anthropometric and gait features from Kinect measurements, pp. 108–119 (2017)