IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Appearance model update based on online learning and soft-biometrics traits for people re-identification in multi-camera environments

ISSN 1751-9659
Received on 22nd March 2019
Revised 25th June 2019
Accepted on 1st July 2019
E-First on 17th September 2019
doi: 10.1049/iet-ipr.2019.0083
www.ietdl.org

Daniela Moctezuma[1,2] ✉, Eric S. Tellez[1,3], Sabino Miranda-Jiménez[1,3], Mario Graff[1,3]

[1]*Cátedras - CONACyT Consejo Nacional de Ciencia y Tecnología, Av. Insurgentes Sur 1582, Col. Crédito Constructor, Del. Benito Juárez, C.P. 03940, Ciudad de México., Mexico*

[2]*Laboratorio de GeoInteligencia - Centro de Investigación en Ciencias de Información Geoespacial, Circuito Tecnopolo Norte No. 117, Col. Tecnopolo Pocitos II, C.P., Aguascalientes, Ags 20313, Mexico*

[3]*Laboratorio de Analitica Computacional - Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Norte 117, Col. Tecnopolo Pocitos II, C.P. 20313, Aguascalientes, Ags, Mexico*

✉ *E-mail: dmoctezuma@centrogeo.edu.mx*

Abstract: Intelligent surveillance systems in multi-camera environments pose a hard-open problem for computer vision. The way the people look changes inside and also among cameras, so people re-identification task can be largely improved collecting data about people already identified and take advantage of it as time advances in surveillance video. Furthermore, a camera change or a slight change in the objective traits may require the complete re-formulation of the appearance models. In this paper, we propose several heuristics for updating the appearance model in a multi-camera surveillance environment. Through these heuristics, the subject's appearance model is updated across different time and environmental conditions. The update process is carried out primarily in three different aspects: 1) based on time lapses, 2) based on the change of camera, and 3) based on the automatic selection of the most representative samples selected through decision functions of the classifier. The proposed system focuses on video surveillance environments, that is, the objective is to identify an individual across the set of cameras in the surveillance area, the comparison considers only those people that share time and space. We used four public benchmarks to test our claims; the results confirm the importance of continuous appearance model's updating.

1 Introduction

An intelligent video surveillance system attempts to understand and describe behaviours in some environment. The aim behind this topic is to provide computer vision algorithms that can assist traditional video surveillance operators, especially in environments with a large number of cameras. Human attention is surpassed as the number of cameras, and the number of objectives, increases. Whenever the number of cameras is large, the amount of images produced is difficult to handle. Moreover, the computational resources needed for some tasks such as people re-identification and detection, event recognition, tracking, or trajectory analysis, pose additional tremendous challenges to these surveillance tasks. People re-identification tackles the problem of recognising a target (a person) across several cameras in a surveillance environment. For this task, the techniques focus on the representation of people using both biometric and soft-biometric traits [1]. The representativeness of samples is critical in the people re-identification task; i.e. it is essential to model people's appearance adequately, according to the environment's conditions and time. Under these circumstances, online and incremental learning is helpful. The related literature on incremental learning is vast, see for instance [1–3], where it is shown that the model's representativeness of any data is critical in any pattern-recognition problem. In the case of people re-identification, the human appearance model obtained in one camera is usually different from those from other cameras, because of variations in view angle, illumination, body poses, clothing, background clutter, and occlusion. As a consequence, the performance of the identification system is degraded when original appearance features are inadequate or non-representative for the substantial intra-class variations of the input samples [4]. The kinds of features to model each people appearance are also essential. Many people re-identification systems are based on biometrics traits; nevertheless, in surveillance environments, these features are useless due to the low quality and the distance acquisition conditions [5]; in contrast, the soft-biometrics features are better suited to these circumstances. However, they are designed to provide clues about the individuals with lack of distinctiveness and permanence [6]; among these characteristics, we found traits related to clothing, aspects of the human body, or gender. The use of soft-biometric traits can improve the performance of the system.

In this paper, we propose several methods to update people's appearance model. Our models use a set of soft-biometrics features; this model can change with time using several strategies to provide online learning capabilities. Our learning machinery is based on popular machine-learning techniques such as linear support vector machine (SVM) classifier with stochastic gradient descent (GD) learning. To evaluate the performance of the proposed approaches, four public data sets were considered: PETS 2006, PETS 2009, CAVIAR4REID, and SAIVT-SoftBio. The experimental analysis shows how competitive our online learning strategies are under different conditions and scenarios. Furthermore, we present and discuss a comprehensive comparison of our work with nine-related state-of-the-art alternatives.

The rest of this paper is organised as follows. Section 2 reviews the related work and Section 3 describes our benchmarks used in our experimental section. Section 4 is dedicated to describe our approaches for model updating based on online learning. The experimental validation of our approach is made in Section 5; finally, Section 6 is devoted to summarise and conclude our findings.

2 Related work

In the past few years, several studies on person re-identification based on appearance models have been published; some of them use online learning approaches to perform the recognition step. For instance, Lu *et al.* [1] propose an online human recognition system with an incremental classifier based on SVM, which updates only the required aspects. The authors extract features related to the colour and the texture of three different human body parts (i.e.

Fig. 1 *Small sample of the CAVIAR4REID data set*

head, torso, and legs) to represent each frame. The CASIA Gait data set containing 20 different people are used in its experimental validation. In [7], a temporal model adaptation scheme is proposed through a similarity learning method. The similarity is learnt through a stochastic derivation of alternative directions methods for multipliers [8]. The method is validated experimentally with VIPeR, PRID450S, and Market1501 benchmarks. An appearance-based person re-identification approach based on a human detection algorithm with a discriminatively trained human-part model is introduced in [9]. The idea is to identify, separately, the human front, side, and back; also, it computes the proportion of the head and body as a trait. So, different features are extracted from human body parts, and these features are compared with a predefined template to find out the similarity value among the body parts and the template. Another work is presented in [10], where an image pool for each person's identity is created. This image pool is created by a set of images produced by a tracking method; the method considers the diversity and representativeness of each image according to the angle and lighting variations.

Furthermore, several rules to update this image pool are proposed; these rules are based on the diversity factor to generate a ranking of them. Deep learning is also presented in people re-identification task such as in the case of [11]. Here, a deep-ranking model with feature learning is proposed. In particular, the authors create a convolutional neural network model to generate a set of local and global features; the model also performs the fusion of these features. The experiments were done using several data sets and, in some cases, the results show a slight improvement in comparison with some of the related works.

Fang *et al.* [12] propose to combine three different types of features to tackle the problem. The kinds of features include the combination of two space colour models, red, green, and blue (RGB) and hue, saturation, and value (HSV), and second, the first- and second-order magnitude gradients are calculated using the first ones. The authors also use histograms, mean, and co-occurrence matrix. The concatenation of all these characteristics defines the multi-statistics on hash feature map descriptor. A method, which employs several descriptors such as local maximal occurrence (LOMO), Weighted Histogram of Overlapping Stripes (WHOS), and Gaussian Of Gaussian (GOG), are presented in [13]. These characteristics are used to fit a learning model. The authors use VIPeR and CAVIAR4REID for benchmarking purposes.

In [14], an efficient algorithm (called early active learning) is proposed for selecting a subset of samples considered as more informative for training. For sample representation, the LOMO feature was used. Four public data sets (VIPeR, PRID, i-LID, and CAVIAR4REID) were considered for experimental validation. Another online update scheme for the people re-identification task is proposed in [15]. Here, the evaluation protocol is handled more realistically; that is, people are compared only with those sharing camera and time. The algorithm keeps it simple and straightforward. Metric Keep it Simple and Straightforward Metric (KISSME) is used as the metric learning method. The UCR, NLPR, RAiD, and SAIVT data sets were used for the validation.

To solve the variance issue in images from multiple cameras, an energy-based loss function is presented in [16] to tackle the people re-identification. This function takes into account two aspects: it favours short distances, which indicate a high level of similarity between instances of people's appearances, and penalise large distances which reflect low similarity, as well as the number of overlaps between the local neighbours of two compared people and local neighbours of other people. The evaluation was done with three public databases: ETHZ, the CAVIAR4REID, and Person Re-ID 2011. In [17], a method to analyse a set of different kinds of features is introduced. The proposed features introduce semantic information, and these features and a novel discriminative model are proposed to learn the attributes correlation. The experimental comparison includes four standard benchmarks, i.e. PKU-REID, SAIVT-SoftBio, iLIDS-VID, and PRID. With the purpose of transfer, the appearance variations in [18] proposed a model using cumulative weighted brightness transfer function employing a set of images instead of image-per-image transfer. For the experiments, the VIPeR, CAVIAR4REID, PRID2011, and SAIVT-SoftBio data sets were considered. A method, which exploits the body's visual clues, is presented in [19]. This descriptor supposes that the person is in an upright pose to capture the chromatic content, the spatial arrangement of colour in the region and the presence of recurrent local motifs with high entropy in the regions. García *et al.* [20] consider the pose and the orientation of the camera to obtain people's representation. The proposed method extracts multiple frames of the same person with different orientations. Furthermore, it learns a pairwise feature dissimilarity function in several sub-spaces; from each sub-space, a classifier is trained according to all the variations of pose and orientations corresponding to each camera view.

Several researchers tackle people's re-identification problem with metrics learning methods. The metric learning is conducted once per individual, but it is not updated according to time or space, see, for instance, [11, 21, 22]. The interested reader on the traditional people's re-identification problem is referred to [23–26], which survey the problem.

The work presented in this paper combines the model appearance with an online learning approach to outperform the identification rate in the people re-identification problem. The following sections detail our contribution.

3 Databases

The research community of the intelligent video surveillance area considers the experimental evaluation as a critical step. Therefore, we found a wide variety of benchmarking data sets. In this paper, we select four standard data sets in the area: CAVIAR4REID, PETS 2006, PETS 2009, and SAIVT-SoftBio. The CAVIAR4REID [27] data set contemplates a real video surveillance environment with complex images with low resolution, light changes, occlusions, and variations in pose and angle. CAVIAR4REID contains images from 72 different people and two non-overlapping cameras. In Fig. 1, several images from CAVIAR4REID can be seen.

The PETS 2006 data set [28, 29] has been designed for activity recognition; nevertheless, it has been employed in both people detection and recognition tasks. In particular, we use the sequences named S2-T3-C and S4-T5-A-C (for camera 1, camera 3, and camera 4, which are overlapping cameras). From PETS 2006, we select manually nine different people who stand in the three considered views. The PETS 2009 data set [30, 31] comprises image sequences containing crowd scenarios in an outdoor environment. We used named sequences S2-L1-Time, that is, camera 1, camera 5, camera 7, and camera 8, and all overlapping cameras. Similarly, to PETS 2006, from PETS 2009, we selected nine different people who stand in the four cameras used. Fig. 2 shows several images from these two data sets.

The most recent data set considered in this work is SAIVT-SoftBio [32]. This data set consists of several image sequences at an indoor environment. The images proceed from eight

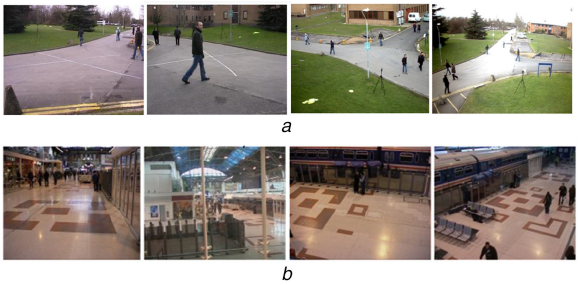

Fig. 2 *Images of*
(a) PETS 2009, ***(b)*** PETS 2006

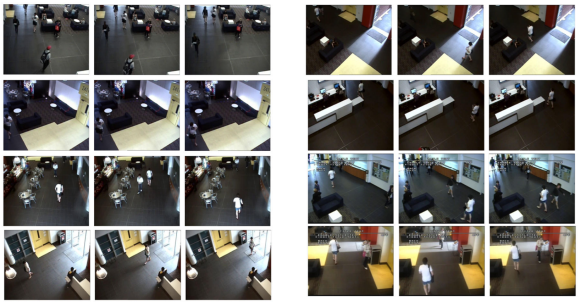

Fig. 3 *Images from SAIVT-SoftBio data set*

Table 1 General description of the extracted features

Category	Description
RGB colour	mean of red colour
RGB colour	mean of green colour
RGB colour	mean of blue colour
RGB colour	mean of three RGB
RGB colour	SD of three RGB
RGB colour	brightness
grey scale	mean
grey scale	SD
geometry	eccentricity
grey-scale histogram	entropy
grey-scale histogram	dispersion
grey-scale histogram	mean
grey-scale histogram	SD
grey-scale histogram	energy
grey-scale histogram	kurtosis
HSV colour	mean
HSV colour	SD
statistics of the co-occurrence matrix	energy
statistics of the co-occurrence matrix	maximum probability
statistics of Matrix of Co-Occurrence (MCO)	entropy
statistics of MCO	inertia
statistics of MCO	homogeneity
LBP	simple LBP

overlapping cameras and include 25 different people. Fig. 3 shows images from its eight different indoor cameras. Considering these four data sets, we test our approaches with 115 people, and with a variety of multi-cameras conditions such as the number and overlapping of the condition of the cameras.

4 Appearance model updating

Generation and updating of the people's appearance model are essential parts of the people's re-identification. In this work, people detection is made following the approach presented in [33]. A set of appearance-based soft-biometric characteristics were extracted for each detected person. Since these traits are highly related to people's clothing, once a person is detected, we compute characteristics based grey-scale statistics and histograms, co-occurrence matrix and local binary patterns (LBPs), and also features related to RGB and HSV colour spaces, and geometry.

As this work, it is focused on the updating process of an appearance model; all the features were computed following the work presented in [34]. These features were selected to work under sub-optimal resolution and lighting conditions, as it is the rule in a real security camera environment. In summary, we use 23 features, normalised between 0 and 1; these features are summarised in Table 1. Here, SD means standard deviation and Matrix of Co-Occurrence (MCO) means co-occurrence matrix. From these features, a vector is generated to identify each people in the scene or for a more detailed analysis, see [34]. Next, with this set of soft-biometric features, a model is constructed; that is, the people appearance model is generated and updated over several conditions.

In this paper, we introduce three different strategies for updating people's model. The first one is based on time lapses, the second one, on camera's changes, and the last one is based on an automatic selection of the one considered as the most representative of samples. We use a linear SVM with stochastic GD (SGD) training [35] as the classifier. In particular, we employ Python's Scikit-learn implementation [36]; this library implements mini-batch learning that helps us to carry out a sort of online learning; hence, we specify through these heuristics when this partial fit is done.

The SVM has a well-studied potential to deal with online and incremental learning [2, 3]. The SGD is a stochastic approximation of the GD iterative optimisation method [37, 38]. A significant difference between GD and SGD is that SGD needs more iterations than GD, but each iteration needs lesser net computational resources. The SGD provides a solution for the underlying SVM optimisation problem, which minimises a cost function, that is, given the training set $I = (x_i, y_i)$, $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$ in the case of binary classification problems. The first step is the initialisation of $W^0 = 0 \in \mathbb{R}^d$; then for each iteration, $1 \ldots T$ this weight vector $\boldsymbol{W}$ must be updated. With the SGD method, in each iteration, a random sample (x_i, y_i) for the training set $\boldsymbol{I}$ is chosen. Then, this subset (x_i, y_i) is used as the full data set to calculate and update the new values of weight vector $\boldsymbol{W}$ through the SGD optimisation approach [37, 38].

Leaving aside the learning procedure, we concentrate on the heuristics to determine the time when the re-trained process must be done. In the following sections, we focus on the details of the model's updating strategies.

4.1 Update based on time lapses

In multi-camera scenarios, we found that people's aspect changes dramatically from one camera to another due to changes in lighting, perspective, distance from the camera, the zooming capacity, or quality, in general. These variabilities make the identification of people across cameras difficult. With the purpose of minimising this problem, it is necessary to update the model calculated for re-identification.

The first proposed approach is based on the construction of an initial appearance model, and then, after a time lapse, the model is updated regardless of any change detected. Hereafter, this method will be called 'update based on time lapses' (UBTL). In this method, when a specified period has elapsed, the partial fit is done, that is, the SVM linear classifier with SGD learning is re-trained integrating new samples to the model. The specified time t is varied from 5, 10 to 15, that is, every t seconds the re-train is performed.

Algorithm 1 (see Fig. 4) describes the process of this updated approach. The approach receives as input the images sequence *Images*, the parameter t which specifies when the update must be done, and the parameter *iModel,* which is the value for the generation of the initial model. The first images are used to generate an initial appearance model; the parameter that controls the number of images for this starting model is *iModel*. For instance, we set the value of *iModel* to 10; please note that setting

Name: UBTL
Input: Images, t, iModel
Produces: The updated classifier

```
1:  tCount ← 0
2:  for i ← 1 to totalTime do
3:      tCount ← tCount + 1
4:      if i == iModel then
5:          classifier.Train(Images[0 : iModel])
6:      else if i > iModel then
7:          if tCount == t then
8:              classifier.PartialFit(Images[i])
9:              tCount ← 0
10:         end if
11:     end if
12: end for
```

Fig. 4 *Algorithm 1: Update based on Time Lapses.*

Name: UBCC
Input: Images, iModel, Cameras
Produces: The updated classifier

```
1:  for i ← 1 to TotalTime do
2:      iCount ← 0
3:      for c ← 1 to Cameras do
4:          iCount ← iCount + 1
5:          if iCount == iModel then
6:              iCount ← 0
7:              if Classifier exists then
8:                  Classifier.PartialFit(Images[i : iModel])
9:              else
10:                 Classifier.Train(Images[i : iModel])
11:             end if
12:         end if
13:     end for
14: end for
```

Fig. 5 *Algorithm 2: update based on change of camera*

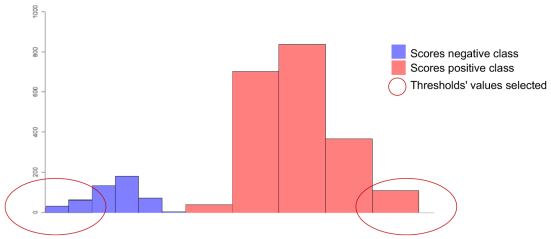

Fig. 6 *Histograms of positive and negative scores*

iModel to a low value allows the use of the rest of the images for updating and testing. After this initial period, the updating process is done through the partial fit method of our classifier. That is, all the individuals in the image, at the specified time, are used as new samples. All identities are included in the classifier. The simplicity and determinism are the main advantages of this method.

4.2 Update based on change of camera

As mentioned before, people's appearance changes across different cameras. Therefore, our second proposal is the model's updating based on camera's change. This strategy starts with an initial appearance model from one camera, and when the re-identified subject crosses from one camera to another, the appearance model is updated. This method will be named update based on change of camera (UBCC).

Algorithm 2 (see Fig. 5) details on how to proceed with camera changes; as before, the first *iModel* images are used to generate an initial model. The classifier is updated whenever the suspect is identified using a different camera. The core idea is to feed the classifier with images from the current cameras. For each camera, the first *iModel* images are used to re-train the classifier and the remaining to testing. Note that in this method, the more cameras are crossed by the suspect, the more samples are considered in the appearance model.

4.3 Update based on the most representative samples

This method is designed to take advantage of the most representative samples. Here, an initial model is generated with the *iModel* first images. In the next stages, most representative samples are selected with a quality threshold for the positive samples and another quality threshold for the negative samples.

To calculate these scores, we apply the classification process to a subset of the data set. This procedure gives us a set of decision functions values from all images considered. With these values, we plotted a histogram to visualise the distribution of both positive and negative decision functions values of the whole subset of images. Later, we choose threshold values by quartiles using the histogram's values; the final thresholds were established with the idea of selecting those values located in the limits of the histogram. Fig. 6 illustrates the above; here, we can see the distribution of both negative and positive classes, and the location of the extreme values. Note that positive-score values are defined by those where the target person is identified correctly, and the negative-score values are defined by values obtained from people not being the target.

Using this methodology in this subset, we found that positive scores have values between 0 and 6, and negative scores have values between −10 and 0. We consider the extreme values, so we fix the positive threshold to 1.5 and the negative threshold to −2.5. Therefore, the updating process is started whenever any of these thresholds are crossed.

Algorithm 3 (see Fig. 7) describes the behaviour of this approach. As well as prior methods, an initial appearance model is generated with the first images (specified by *iModel*). The parameters posTh and negTh are the positive and negative thresholds, respectively. Later, for each image, a score is obtained for each detected person using the prior appearance model, when the score of a sample is higher or lower than the considered quality thresholds, the appearance model is updated through the re-training of the classifier. Owing to the high number of calculations, this method is the most time-consuming among our approaches.

In all the cases, the parameters of the algorithms are the set of images, t is the time associated with each image, and the *iModel* value, which controls the bootstrapping of the system. In particular, the last value is provided by the operator of the surveillance system. On the other hand, all the evaluation metrics [see (1)–(4)], are calculated in the test phase.

5 Experiments and analysis

Our experimental evaluation consists of the use of four popular public benchmarks, PETS 2006, PETS 2009, CAVIAR4REID, and SAIVT-SoftBio databases. These data sets were created on both overlapping and non-overlapping surveillance environments. From these databases, a total of 115 different individuals were extracted, with about 274 images per person; the experiments were carried out with more than 31,000 images.

Furthermore, the experiments were conducted regarding spatial and temporal restrictions, i.e. two individuals need to be compared if they occur in the same space (or camera) at the same time. The evaluation is oriented to simulate an operational environment; that is, a suspect must be compared with other people present in the same scene at the same time. These benchmarks help us to provide experimental evidence to support the importance of updating models in multi-camera surveillance methods and, in particular, to measure the performance of our strategies. Note that for some time lapse, we use all previously known images of an objective person as the training data and the remaining images in time as testing data. This scheme avoids the use of the same data for training and testing, but each approach has a different size of data and time; we

Name: UBMRS
Input: Images, iModel, posTh, negTh
Produces: The updated classifier

```
for i ← 1 to TotalTime do
    if i == iModel then
        classifier.Train (Images[0 : iModel])
    else if i > iModel then
        Score ← classifier.Predict(Images[i])
        {Score is the decision functions of classifier}
        if Score ≥ posTh or Score ≤ negTh then
            classifier.PartialFit(Images[i])
        end if
    end if
end for
```

Fig. 7 *Algorithm 3: update based on most representative samples*

Table 2 Baseline's performance

Database	Accuracy	F_1	Precision	Recall
PETS 2006	0.1425	0.1298	0.1089	0.1600
PETS 2009	0.2556	0.2361	0.1989	0.2750
CAVIAR4REID	0.0265	0.0147	0.0101	0.0277
SAIVT-SoftBio	0.0783	0.0671	0.0798	0.0588

Table 3 Performance obtained with the UBTL method

Database	Accuracy	F_1	Precision	Recall
$t = 5$				
PETS 2006	0.8999	0.8889	0.8929	0.8849
PETS 2009	0.7128	0.6435	0.6287	0.6592
CAVIAR4REID	0.2197	0.1630	0.1787	0.1499
SAIVT-SoftBio	0.4953	0.4124	0.3920	0.4350
$t = 10$				
PETS 2006	0.8506	0.8328	0.8390	0.8267
PETS 2009	0.6511	0.5768	0.6231	0.5368
CAVIAR4REID	0.1769	0.1317	0.1489	0.1179
SAIVT-SoftBio	0.4379	0.3551	0.3493	0.3611
$t = 15$				
PETS 2006	0.8353	0.8035	0.7821	0.8262
PETS 2009	0.6133	0.5164	0.6161	0.4909
CAVIAR4REID	0.1654	0.1185	0.1380	0.1039
SAIVT-SoftBio	0.3988	0.3189	0.3289	0.3094

Table 4 Performance obtained with the UBCC method

Database	Accuracy	F_1	Precision	Recall
*i*Model = 5				
PETS 2006	0.7534	0.5996	0.5870	0.6127
PETS 2009	0.7870	0.6562	0.6692	0.6437
CAVIAR4REID	0.5806	0.4277	0.4197	0.4361
SAIVT-SoftBio	0.6769	0.3640	0.3559	0.3724
*i*Model = 10				
PETS 2006	0.8023	0.6310	0.6268	0.6352
PETS 2009	0.8440	0.6998	0.6835	0.7169
CAVIAR4REID	0.6965	0.6014	0.6046	0.5982
SAIVT-SoftBio	0.6912	0.4256	0.4315	0.4199
*i*Model = 15				
PETS 2006	0.8343	0.6777	0.6662	0.6896
PETS 2009	0.8813	0.7485	0.7590	0.7382
CAVIAR4REID	0.8507	0.7162	0.7291	0.7038
SAIVT-SoftBio	0.6924	0.4334	0.4379	0.4290

present the average values of ten independent runs to reduce the variation among the stochastic behaviour of underlying methods.

For the evaluation, we use F_1 and accuracy scores. Equation (1) shows how the accuracy is calculated. Here, TP means for true positives, that is, it counts all correct predictions of the positive class. TN means for true negatives, i.e. it counts all correctly predicted negatives. FP counts the mismatch regarding positive predictions, while FN (false negatives) counts failures regarding predictions of the negative class. The F_1 recall and precision metrics are calculated as (2)–(4) indicate

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F_1 = 2\,\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

It is possible to know the interconnection of the cameras in any real-world surveillance system. Our approach takes into account this information considering the camera and recording time. This approach reduces the search space significantly since a person is compared only with those people who are in the same space and time.

Table 2 shows the results obtained without the update process, that is, the initially generated model using the *iModel* first images without any updating strategy. It can be seen that the results obtained in all databases are dramatically low with an average 0.1257 of accuracy and 0.1119 of F_1 measure. With the purpose to analyse the benefit of online learning in the people re-identification task, we considered these results as our baseline.

Table 3 shows the results obtained with the UBTL method. In this case, the experiments were carried out considering several values of t, and we tested with 5, 10, and 15. As it was expected, best results were achieved with lower values of t since the finely grained update; nevertheless, the differences with other values of t are small. Low t values produce the best performances, and this happens because we have more images for training, and hence, more chances of executing the update process. Although we expected a higher difference between different t results; nevertheless, the achieved results between the t values are similar.

We observe an improvement in both accuracy and F_1; that means, when the appearance model is updated, at each elapsed time, the performance is improved. Despite the improvement, the results have a high variance among data sets. For instance, in PETS 2006, it reaches an accuracy of 0.8999, but in CAVIAR4REID, it achieves an accuracy of 0.2197. These results are quite different, and this variation is related to the complexity of each database. PETS 2006 has the best image quality of the four data sets considered here; furthermore, the number of people re-identified in PETS 2006 with nine individuals versus CAVIAR4REID with 72 individuals; this contributes to the result differences.

On the basis of the idea that the best moment to capture the different views of people in a video surveillance area is when they move to another area; that means, to another camera view, we proposed the method based on the change of camera or UBCC. Table 4 shows the results of the UBCC method. The appearance model is generated with the first images in each camera view, where people are, that is, *iModel* has three different values 5, 10, and 15.

This strategy produces competitive results for all benchmarks; for instance, updating with the first 15 images from each camera (UBCC with *i*Model = 15), the accuracy values obtained were 0.8343, 0.8813, 0.8507, and 0.6924 for PETS 2006, PETS 2009, CAVIAR4REID, and SAIVT-SoftBio, respectively.

Table 5 lists the results of update based on the most representative sample (UBMRS). In this case, the initial model was generated with *iModel* equal to ten; that means, once the initial

Table 5 Results with the UBMRS method

Database	Accuracy	F_1	Precision	Recall
PETS 2006	0.7279	0.6744	0.6686	0.6802
PETS 2009	0.5442	0.4864	0.4799	0.4930
CAVIAR4REID	0.1526	0.1098	0.1087	0.1109
SAIVT-SoftBio	0.3048	0.2385	0.2420	0.2351

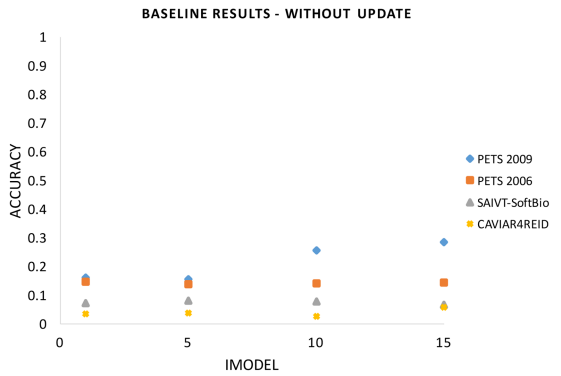

Fig. 8 *Baseline's performance, i.e. no updating policy*

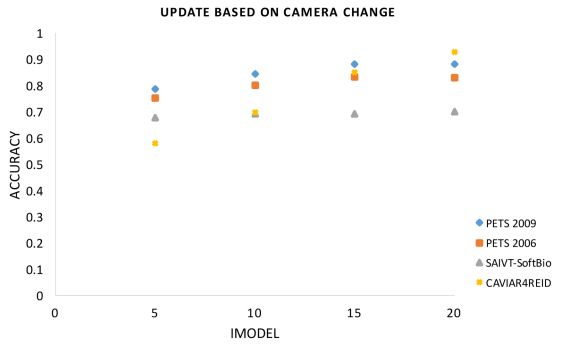

Fig. 9 *Performance of UBCC*

Table 6 Comparison with related work in terms of accuracy metric

Work	CAVIAR4REID	SAIVT-SoftBio
Fang *et al.* [12]	0.3144	—
Mirmahboub *et al.* [13]	0.4520	—
Liu *et al.* [14]	0.3875	—
Zhang *et al.* [16]	0.3300	—
Shiva Kumar *et al.* [15]	—	**0.8570**
Su *et al.* [17]	—	0.6887
Bhuiyan *et al.* [18]	0.2395	0.4540
Bazzani *et al.* [19]	0.2800	—
García *et al.* [20]	—	0.3350
ours	**0.8507**	0.6924

Bold values indicate are the best ones in each dataset.

model is generated, every time a sample reaches the established thresholds, the updating process is activated.

This method selects the best samples when it reaches a threshold, established to be bigger than or equal to 1.5 for the best positive item and smaller than or equal to −2.5 for the best negative. The performance of UBMRS is low; among the possible causes, we found low representativeness of the sample.

In summary, in PETS 2006, the best results were obtained with the UBTLs with an accuracy of 0.8999 and F_1 0.8889. With the approach of UBCC, the obtained results show that an appearance model built with the first images of a camera is enough to represent the variations in lighting, shadows, or perspective that could appear in the whole camera network. We found a significant improvement with our methods regarding results obtained with the selected baseline, a model without any update process. The best results with CAVIAR4REID are achieved with UBCC with an accuracy of 0.8507, using *i*Model = 15. This result shows the relevance of considering an update process in non-overlapping environments, which is the case of CAVIAR4REID.

In summary, Fig. 8 compares the accuracy obtained across benchmarks. It shows the performance of static models and several values of *iModel* (the *X*-axis). On the other hand, in Fig. 9, the results of all data sets with the UBCC approach are shown. From these two figures, the improvement in accuracy only for using an update process can be observed. Furthermore, several statements can be analysed; for instance, CAVIAR4REID produces lower results, at least when no model updating is applied; nevertheless, we reached a competitive result for this benchmark using UBCC. Our baseline is a model based on the same machine-learning method and representation without updating capabilities. Our benchmarking data sets achieve accuracy scores of 0.1425, 0.2556, 0.0265, and 0.0783, for PETS 2006, PETS 2009, CAVIAR4REID, and SAIVT-SoftBio, respectively. These scores improve to 0.8023, 0.8440, 0.6965, and 0.6912 using UBCC with *i*Model = 10; these scores mean improvements of 5.6, 3.3, and 2.6 times for PETS and CAVIAR4REID, and a small performance impact for SAIVT-SoftBio.

5.1 Comparison with other alternatives

Our main contribution is to analyse the performance of the re-identification task with and without an appearance model update process. The above makes complicated evaluations under the same conditions of all the compared methods due to spatial and temporal information required that many data sets lack. In spite of this and to analyse and position our results in the tested data sets, in this section, we present the overall results of people re-identification in two of the standard data sets used in this work: CAVIAR4REID and SAIVT-SoftBio. This comparison was made with the works using the data sets employed in this paper and with the accuracy metric reported. In Table 6, the results of nine-related works in terms of the accuracy metric with CAVIAR4REID and SAIVT-SoftBio data sets were shown.

The compared works are proposed in [12, 13, 15–20], all described in Section 2. Here, it can be seen that in the CAVIAR4REID data set, the best result was obtained with our proposed methodology, in contrast, with the SAIVT-SoftBio data set, the best result was obtained with the work presented in [15]; nevertheless, they only use images from four cameras, instead of the total of eight cameras included in this data set.

Finally, it is important to say that we tested the methods proposed to update the appearance model emulating the operation of an intelligent video surveillance system. In this sense, we must have the time and camera metadata of each image sequence used; in the case that we do not have this information, we manually labelled the image sequence to generate a logic camera–time relationship. Under these conditions, all the proposed methods were tested, and though the results are far from being perfect, we think that the improvement achieved using some of the proposed methodologies to update the people appearance model had proved their potential on improving the re-identification rate. Hence, the proposed methods have some limitations; for instance, it is easy to observe that when the number of people in the search list is large, the performance is decreased dramatically in both accuracy and F-measure metrics. For instance, in the results, we can observe that the performance is lower with the CAVIAR4REID and SAIVT-SoftBio data sets, which both have more than eight identities. Nevertheless, emulating the real operational way of any surveillance system, the methodology can be useful and serve as a good assistant to the system's operator.

6 Conclusions

This work introduces novel strategies for updating people's appearance model in intelligent surveillance systems in multi-camera environments. The aim is to update models to maximise the performance of people's re-identification task; for each detected person, we use a bag of soft-biometric features in any of our

methods. We propose three different strategies for updating people's models: (i) based on time lapses (UBTL), (ii) based on the camera's change (UBCC), and (iii) based on the selection of most representative samples (UBMRS). We evaluated our strategies experimentally using a set of popular benchmarks, all of them freely available to the community. On the basis of these benchmarks, our methods were tested in non-overlapping and overlapping scenarios.

Our experimental evaluation has shown the relevance of updating the appearance model in the people's re-identification task, and in particular, the use of online learning for this purpose has achieved promising results. In comparison with other related works, our approach achieved the best result with the CAVIAR4REID data set, and the second-best result with the SAIVT-SoftBio data set. On the basis of our results, it is needed to develop better strategies to take advantage of new images generated by surveillance systems online. As part of future research, we may focus on improving the adaptability of algorithms to different camera's conditions, and also adapt and update models based on particular characteristics of each person.

7 References

[1] Lu, Y., Boukharouba, K., Boonaert, J., *et al.*: 'Application of an incremental SVM algorithm for on-line human recognition from video surveillance using texture and color features', *Neurocomputing*, 2014, **126**, pp. 132–140

[2] Cauwenberghs, G., Poggio, T.: 'Incremental and decremental support vector machine learning'. Proc. 13th Int. Conf. Neural Information Processing Systems NIPS'00, Cambridge, MA, USA, 2000, pp. 388–394.

[3] Syed, N.A., Huan, S., Kah, L., *et al.*: 'Incremental learning with support vector machines', 1999

[4] Rattani, A., Freni, B., Marcialis, G.L., *et al.*: 'Template update methods in adaptive biometric systems: a critical review'. Int. Conf. Biometrics (ICB), Alghero, Italy, 2009, pp. 847–856

[5] Satta, R., Fumera, G., Roli, F., *et al.*: 'A multiple component matching framework for person re-identification'. ICIAP, Ravenna, Italy, 2011, pp. 140–149

[6] Jain, A.K., Dass, S.C., Nandakumar, K.: 'Soft-biometric traits for personal recognition systems'. Int. Conf. Biometric Authentication (ICBA), Hong Kong, 2004, pp. 731–738

[7] Martinel, N., Das, A., Micheloni, C., *et al.*: 'Temporal model adaptation for person re-identification', in Leibe, B, Matas, J, Sebe, N, Welling, M (Eds.): '*Computer vision – ECCV 2016*', (Springer International Publishing, Cham, 2016), pp. 858–877

[8] Boyd, S., Parikh, N., Chu, E., *et al.*: 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Found. Trends Mach. Learn.*, 2010, **3**, (1), pp. 1–122

[9] Wei, W., Ma, H., Zhang, H., *et al.*: 'Person re-identification based on adaptive feature selection', in Zu, Q, Vargas-Vera, M, Hu, B, (Eds.): '*Pervasive computing and the networked world*', (LNCS, **8351**, (Springer International Publishing, Phnom Penh, Cambodia, 2014), pp. 441–452

[10] Yuan, M., Yin, D., Ding, J., *et al.*: 'A multi-image joint re-ranking framework with updateable image pool for person re-identification', *J. Vis. Commun. Image Represent.*, 2019, **59**, pp. 527–536

[11] Wang, J., Zhou, S., Wang, J., *et al.*: 'Deep ranking model by large adaptive margin learning for person re-identification', *Pattern Recognit.*, 2018, **74**, pp. 241–252

[12] Fang, W., Hu, H.-M., Hu, Z., *et al.*: 'Perceptual hash-based feature description for person re-identification', *Neurocomputing*, 2018, **272**, pp. 520–531

[13] Mirmahboub, B., Mekhalfi, M.L., Murino, V.: 'Person re-identification by order-induced metric fusion', *Neurocomputing*, 2018, **275**, pp. 667–676

[14] Liu, W., Chang, X., Chen, L., *et al.*: '*Early active learning with pairwise constraint for person re-identification*', (Springer International Publishing, Cham, 2017)

[15] Shiva Kumar, K.A., Ramakrishnan, K.R., Rathna, G.N.: 'Inter-camera person tracking in non-overlapping networks: re-identification protocol and on-line update'. Proc. 11th Int. Conf. Distributed Smart Cameras ICDSC 2017, New York, NY, USA, 2017, pp. 55–62

[16] Zhang, G., Kato, J., Wang, Y., *et al.*: 'Adaptive metric learning in local distance comparison for people re-identification'. 2013 Second IAPR Asian Conf. Pattern Recognition (ACPR), Naha, Japan, November 2013, pp. 196–200

[17] Su, C., Zhang, S., Yang, F., *et al.*: 'Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping', *Pattern Recognit.*, 2017, **66**, pp. 4–15

[18] Bhuiyan, A., Perina, A., Murino, V.: 'Exploiting multiple detections for person re-identification', *J. Imaging*, 2018, **4**, (2), p. 28

[19] Bazzani, L., Cristani, M., Murino, V.: 'Symmetry-driven accumulation of local features for human characterization and re-identification', *Comput. Vis. Image Underst.*, 2013, **117**, (2), pp. 130–144

[20] García, J., Martinel, N., Gardel, A., *et al.*: 'Modeling feature distances by orientation driven classifiers for person re-identification', *J. Vis. Commun. Image Represent.*, 2016, **38**, pp. 115–129

[21] Dai, J., Zhang, Y., Lu, H., *et al.*: 'Cross-view semantic projection learning for person re-identification', *Pattern Recognit.*, 2018, **75**, pp. 63–76

[22] Liao, S., Hu, Y., Zhu, X., *et al.*: 'Person re-identification by local maximal occurrence representation and metric learning'. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, June 2015, pp. 2197–2206

[23] Bedagkar-Gala, A., Shah, S.K.: 'A survey of approaches and trends in person re-identification', *Image Vis. Comput.*, 2014, **32**, (4), pp. 270–286

[24] Vezzani, R., Baltieri, D., Cucchiara, R.: 'People reidentification in surveillance and forensics: a survey', *ACM Comput. Surv.*, 2013, **46**, (2), pp. 29:1–29:37

[25] Harle, R.: 'A survey of indoor inertial positioning systems for pedestrians', *Commun. Surv. Tutor. IEEE*, 2013, **15**, (3), pp. 1281–1293

[26] Neves, J., Narducci, F., Barra, S., *et al.*: 'Biometric recognition in surveillance scenarios: a survey', *Artif. Intell. Rev.*, 2016, **46**, (4), pp. 515–541

[27] Cheng, D.S., Cristani, M., Stoppa, M., *et al.*: 'Custom pictorial structures for re-identification'. Proc. British Machine Vision Conf., Dundee, Scotland, 2011, pp. 68.1–68.11. doi: http://dx.doi.org/10.5244/C.25.68

[28] PETS 'Performance evaluation of tracking and surveillance 2006 benchmark data', 2006. Available at http://www.cvg.rdg.ac.uk/PETS2006/data.html, January 20, 2019

[29] Ferryman, F., Thirde, D., Li, L.: 'Overview of the pets2006 challenge'. 2009 IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, New York, USA, 2006, pp. 47–50

[30] Ferryman, J., Shahrokni, A.: 'Pets2009: dataset and challenge'. 2009 12th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, Miami, Florida, December 2009, pp. 1–6

[31] Ellis, A., Shahrokni, A., Ferryman, J.: 'Overall evaluation of the pets 2009 results'. 11th IEEE Int. Workshop on PETS, Miami, Florida, 2009, pp. 117–124

[32] Bialkowski, A., Denman, S., Sridharan, S., *et al.*: 'A database for person re-identification in multi-camera surveillance networks'. 2012 Int. Conf. Digital Image Computing Techniques and Applications (DICTA), Miami, Florida, December 2012, pp. 1–8

[33] Conde, C., Moctezuma, D., de Diego, I.M., *et al.*: 'Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments', *Neurocomputing*, 2013, **100**, pp. 19–30

[34] Moctezuma, D., Conde, C., De Diego, I.M., *et al.*: 'Soft-biometrics evaluation for people re-identification in uncontrolled multi-camera environments', *EURASIP J. Image Video Process.*, 2015, **2015**, (1), p. 28

[35] Bottou, L.: 'Large-scale machine learning with stochastic gradient descent'. Proc. 19th Int. Conf. Computational Statistics (COMPSTAT'2010), Paris, France, August 2010, pp. 177–187.

[36] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.*: 'Scikit-learn: machine learning in python', *J. Mach. Learn. Res.*, 2011, **12**, pp. 2825–2830

[37] Robbins, H., Monro, S.: 'A stochastic approximation method', *Ann. Math. Stat.*, 1951, **22**, (3), pp. 400–407

[38] Kiefer, J., Wolfowitz, J.: 'Stochastic estimation of the maximum of a regression function', *Ann. Math. Stat.*, 1952, **23**, (3), pp. 462–466