Tracklet and Signature Representation for Multi-shot Person Re-Identification.

Salwa Baabou, Furqan M Khan, François Bremond, Awatef Ben Frad, Mohamed Amine Farah, Abdennaceur Kachouri

Noname manuscript No.
(will be inserted by the editor)

Tracklet Pre-Processing For a performant/operational Multi-Shot Person Re-Identification System using Part Appearance Mixture Approach

Salwa BAABOU[1,2] · **Furqan M.KHAN**[3] · **François BREMOND**[3] · **Awatef BEN FRADJ**[2] · **Mohamed Amine FARAH**[2] · **Abdennaceur KACHOURI**[2]

Received: date / Accepted: date

Abstract Recognizing persons in a video surveillance scene in the real world is attractive and is now showing an increasing interest. The objective of person Re-Identification (Re-ID) system consists in recognizing a person by assigning the same identifier to all instances of a particular individual captured in a series of images or videos in a distributed network of cameras, even after the occurrence of significant gaps over time or space. The Re-ID system is divided into two main stages: *i)* extracting feature representations to construct a person's appearance signature and *ii)* establishing the correspondence/matching by learning similarity metrics or ranking functions. However, extraction of significant/meaningful features plays an essential role in person Re-ID. Thus, to extract robust features and to improve the accuracy of Re-ID, people should be well detected and tracked. This paper provides a detailed representation of the tracklets, *i.e* person trajectories, in another words we can say that the main contribution of this work is how to pre-process tracklets in order to pro-

Salwa BAABOU
e-mail: baabousalwa@gmail.com
Furqan M.KHAN
e-mail: furqan.khan@inria.fr
François BREMOND
e-mail: francois.bremond@inria.fr
Awatef BEN FRADJ
e-mail: benfradj_awatef@yahoo.fr
Mohamed Amine FARAH
e-mail: med.farah@yahoo.fr
Abdennaceur KACHOURI
e-mail: abdennaceur.kachouri@enis.rnu.tn

[1] National Engineering School of Gabes, University of Gabes, Tunisia
[2] Laboratory of Electronics and Information Technology (LETI), National Engineering School of Sfax, University of Sfax, B.P.W.3038 Sfax, Tunisia
[3] INRIA Sophia Antipolis-Mediterranee, 2004 Route des Lucioles-BP93 06902 Sophia Antipolis, France

vide a performant and operational person Re-ID system and then to compute the appearance descriptors, called *signatures* and represent it based on the approach of Part Appearance Mixture (PAM) in the context of multi-shot person Re-ID. An evaluation of the quality of this signature representation is also described in order to essentially solve the problems of high variance in a person's appearance, occlusions, illumination changes and person's orientation/pose. To deal with variance in a person's appearance, we represent it as a set of multi-modal feature distributions modeled by Gaussian Mixture Model (GMM). Experiments and results on two public datasets and on our own dataset show good performance.

Keywords Person Re-Identification (Re-ID) system · person detection · person tracking · tracklet pre-processing · Part Appearance Mixture (PAM) · signature representation

1 Introduction

Video surveillance applications are becoming an important aspect of every day life. Moreover, surveillance cameras usually cover large disconnected areas. Thus, it is important to associate people as they cross from different Field-of-view (FOV) of a camera to another. This association across camera and time is the Re-Identification (Re-ID) problem which we are dealing with in this work. Re-ID is a very active research area and has become an essential task in any automated video surveillance system. In fact, Person Re-Identification (Re-ID) aims to match individuals appearing across non-overlapping camera networks at distinct times and locations. However, the main difficulty in Re-ID is the variability of person's appearance across different cameras, which makes the recognition of individuals more and more challenging. Many recent work assume that persons are perfectly detected and tracked but this scenario is far from being true. so the step of person detection and tracking should be taken into consideration. However, person detection and multiple person tracking are difficult problems with their own hurdles. Significant amount of work has gone into the problem of person detection over the years as well as Multiple Object Tracking (MOT) within a single and multiple camera's FOV which has also been widely researched, but sustained tracking under varying observation environments remains an open problem. All those factors affect the overall performance of a surveillance system.

Fig.1 shows the diagram of person Re-ID system. A typical Re-ID system may have an image (single shot) or a video (multi-shot) as input for feature extraction and signature generation. It starts with automatic person detection and tracking. Then, the step of feature extraction and descriptor generation in order to learn a person's visual signature or model and then compare the two models to get either a match or a non-match.

A subject may be fully or partially occluded by other subjects or carrying items that lead to errors in matching between tracklets. Furthermore, some works in person Re-ID used body-parts methods (such as SDALF, MPMC)[10]

to solve the issue of signature alignment but this problem is still difficult and not efficient as these methods require real detections and many annotations. We can also cite the low image quality as another problem in person Re-ID where the captured images of a person may suffer from low resolution, noise or blur due to limited imaging quality of surveillance cameras. All these issues may affect the performance of person Re-ID which is still not robust enough to guarantee high accuracy in practice.

In that aspect, the main question is now: have we reached the point where we can rely on a completely autonomous, performant/efficient and operational surveillance system for detecting, tracking and re-identifying people across multiple cameras with non-overlapping FOV ?

To this end/To sum up, the contribution of this paper is: focusing on different components of the surveillance system (precisely detection and tracking) in order to perfectly identifying persons. In other words, we can say that our own work consists in the pre-processing of the tracklets (output of the detection and tracking step) to make them good and efficient for computing the signature and then represent it for multi-shot Re-ID based on Part Apperance Mixture PAM approach[9]. This may cater the high variance in a person's appearance and discriminate between persons with similar appearances. A Mahalanobis based distance is defined to compute similarity between two signatures. We finally evaluate the performance of the full Re-ID system.

The paper is organized as follows: The Re-ID process which contains person detection and tracking is described in the following section. Section III is the core of the paper: it introduces the pre-processing of tracklets and its representation based on the Part Appearance Mixture Approach (PAM) by presenting the signature representation and computing the similarity between these latter using metric learning algorithms. Finally, we evaluate the performance of the whole Re-ID system as well as the quality of our signature representation based on the realized experiments and results before concluding.

2 Overview of Our Approach

The advances in computer vision, as well as machine learning techniques in the recent years, have ameliorated this expedition towards smart surveillance at a fast pace and as a result, a plethora of algorithms for the automatic analysis of the video sequences have been proposed. They include, for instance, person detection, person tracking, activity monitoring, and person Re-Identification. Some survey papers such as [1, 2, 3, 4, 5, 6] have presented them in detail. It's in this context that in our proposed approach, we will use advanced computer vision approaches and algorithms to detect, track and re-identify persons and so that get a performant/operational/autonomous and efficient Re-ID system. Based on the step of tracklets pre-processing, we will evaluate the performance of the whole Re-ID system by computing and evaluating the quality of signature representations. Fig.2 illustrates an overview of the Re-ID process, containing the different steps that we will follow and explain it later. It starts

with automatic person detection. In recent years, most of the existing person Re-Identification works have ignored this step and assume perfect pedestrian detection. However, perfect detection is impossible in real scenarios and misalignment can seriously affect the person Re-ID performance. Therefore, this factor should be carefully studied in future studies.

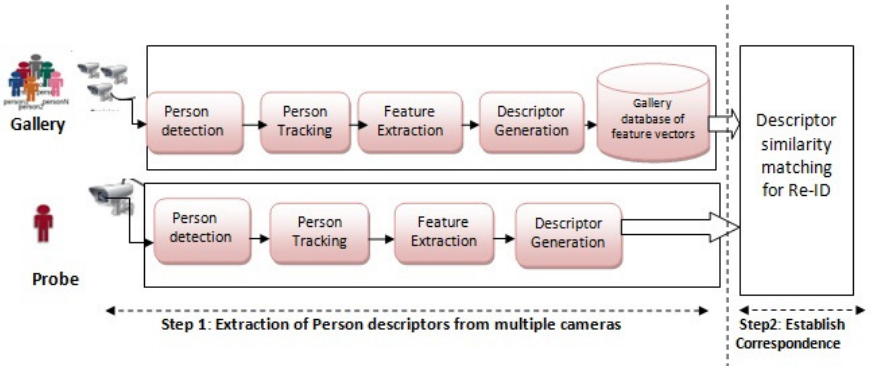
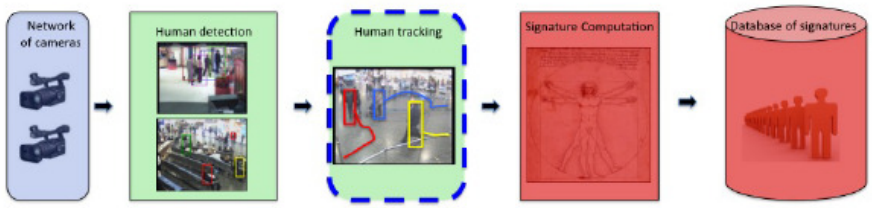


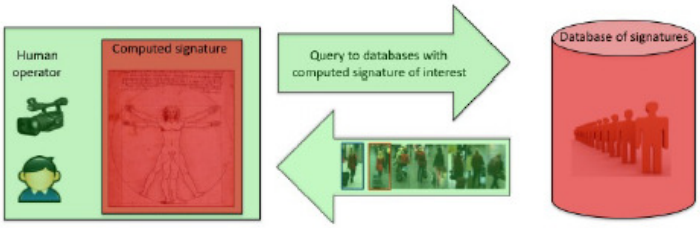
Fig. 1 Person Re-ID diagram

In order to build a strong visual signature of people appearances, persons have to be accurately detected and tracked, so the step of person tracking should be also taken into consideration. However, person detection and multiple person tracking are difficult problems with their own hurdles. Significant amount of work has gone into the problem of person detection over the years as well as Multiple Object Tracking (MOT)[18, 19] within a single camera's Field-Of-View (FOV) as well as multiple cameras which has also been widely researched but it remains an open problem. For feature extraction and descriptor generation, the most commonly used features are color, shape, position, texture, and soft-biometry. The adopted feature is determined by different factors. On one side, the signature should be unique and discriminative enough which can lead to the selection of biometry or soft-biometry features. On the other side, camera resolution, computational load and other implementation issues can prevent or limit their usage and more generic features are required. It is worth noting, that the Re-ID system; in which our work is focusing; as appearing in the relevant literature, turns out to be divided as we said into two distinctive steps (see Fig.1): i) extracting distinctive visual features to represent the human appearance and ii) establishing correspondence by learning or discovering an optimal metric that can maximize the distance between samples from different classes whilst minimizing the distance between those belonging to the same class.

All these steps will be described in details in the following section.



(a) the processing chain of an automatic human signature generation



(b) an interaction of a human operator with a database of signatures

Fig. 2 Overview of the proposed Re-ID process

2.1 Detection

Person detection is the process of detecting and localizing each subject in the images, represented via bounding boxes which is by itself an intensive research field. There are many problems that need to be solved in subject detection:

- Detecting subjects with different sizes
- Detecting subjects with different scales
- Detecting subjects in different locations
- Detecting different subjects in the scenes vs. background
- The detection process should be done in real time, so that it can be used in real applications such as autonomous cars, surveillance systems, etc.

Subject detection can be considered also as classification process; First intuitive idea is to deal with detection as classification of all possible bounding boxes in the image, and classify them as different subjects. To take this way, we need a sliding window with certain step to span the whole image. In addition, we need the windows to be with different sizes, and scales. At the end, we have the bounding box of the subject, and a score (confidence) of the classification. Theoretically, if we have a very fast classifier, this can work. But in reality, the sliding window is slow, we need too many windows to guarantee that all possible regions are tested. To solve the problem of sliding window, instead of looking at all possible positions, we can have a smarter system that can find some interesting regions, and tell the classifier where to look. Example of region proposals are selective search, and Edge-boxes.

There are many subject detectors that we can cite: R-CNN, fast/er R-CNN [11] which are two-stage detectors; first, they propose regions, then they apply classification and bounding box regression. The modern detectors deal with the whole detection process as bounding box regression, so they are much faster which are YOLO[12] and SSD[13].

In this paper, we will use the SSD detector[9]. Fig.3 shows the architecture of the SSD detector.

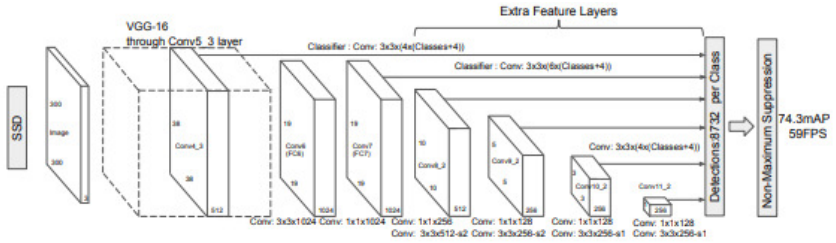


Fig. 3 Overview of the SSD detector architecture[13]

Fig.4 shows a visualization of a sample from CHU Nice dataset of the detection results. In fact, the SSD detector differs from other single shot de-

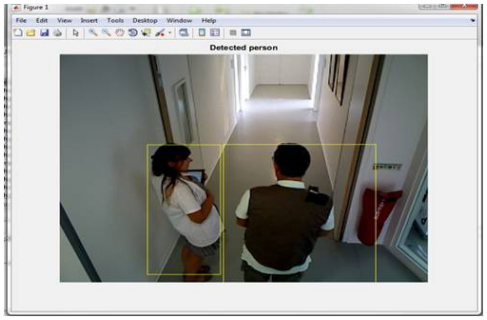


Fig. 4 A visualization of a sample from CHU Nice dataset of the Detection Results

tectors due to the usage of multiple layers that provide a finer accuracy on subjects with different scales. (Each deeper layer will see bigger subjects). It starts with a VGG pre-trained model. Then after the image is passed on the VGG network, some convolutional (conv) layers are added producing feature maps of sizes 19x19, 10x10, 5x5, 3x3, 1x1. These, together with the 38x38 feature map produced by VGG's conv 4-3, are the feature maps which will be used to predict bounding boxes. Using the same concept as anchor boxes in Faster R-CNN, and the idea of dividing the image to grid in YOLO, they

apply different boxes with different sizes and scales to different feature maps. For each default box on each cell, the network output are the following:

- A probability vector of length c, where c are the number of classes plus the background class that indicates no subject.
- A vector with 4 elements (x,y,width,height) representing the offset required move the default box position to the real subject.

2.2 Multi-Object Tracking

Depending on the time of data association process, tracking algorithms can be categorized into 2 types: short-term and long-term tracking. Short-term trackers [14], [15] associate object detections in current frame with the most matching object trajectories in the past. These methods are able to perform online processing based on frame-to-frame association and therefore, could be applied in real-time applications. In general, short-term trackers use bipartite matching methods for short-term data association where Hungarian algorithm is the most popular method. Although these methods are computationally inexpensive, object identification could fail due to inaccurate detections (false alarms) and only short-term occlusions can be handled. Long-term trackers [16], [17] can overcome the shortcomings of short-term trackers by extension of the bipartite matching into network flow. tracklets and edges are the similarity links between However, long-term tracking methods also have their obvious drawbacks, such as: their huge computational cost due to iterative association process to generate globally optimized tracks and their pre-requirement for entire object detection in a given video. Recently, some proposed trackers tried to combine both short-term and long-term tracking methods in a framework to perform online object tracking. The MOT methods in [18], [19] use a frame-to-frame association to generate tracklets followed by a tracklet association process with a time buffer latency. However, their performance is limited by their object features and tracklet representation. These methods utilize basic features (e.g. 2D information, color histogram or constant velocity) applied on whole body parts and use normal Gaussian distribution to describe the object. This way of representation could lose important information to discriminate objects and consequently, could fail to track objects in complex scene conditions (such as occlusion, low video resolution or insufficient lighting of environment). On the other hand, multiple-shot person re-identification methods [20], [21], [9] gained high performances in matching objects from different camera views. In order to match a given person in a camera to the closest person in a gallery in another camera, these re-identification methods use efficient features and object representations. These methods are adapted to solve problems that involve pose and camera view setting variation. Since person Re-identification usually deals with identification of a person from different camera views, it is expected that using Re-id representation becomes even more effective in single-view multi-object tracking problem.

Therefore, we propose a robust online multi-object tracking method named MTSTracker which extends object representation and methods proposed for Re-Identification domain to address problems in MOT. While the re-identification works in offline mode, MTSTracker works in online mode. This method uses a time-window buffer to extract tracklets and associates tracklets in each time-window by using Re-identification techniques. MTSTracker integrates a short-term and long-term trackers in a comprehensive framework. The short-term tracker generates object trajectories called tracklets. Object features are computed for full and body parts, then, each tracklet is represented by a set of multi-modal feature distribution modeled by GMMs. The long-term tracker associates tracklets after mis-detections or occlusions based on learning Mahalanobis distance between GMM components. In order to learn this metric, KISSME [24] algorithm is adopted to learn feature transformations between different scenes by directly learning transformation between probability distributions.

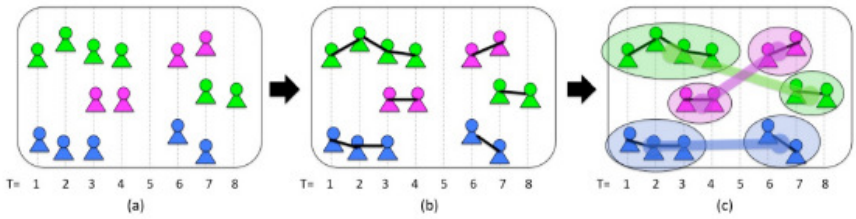


Fig. 5 General Overview of the tracking approach in T frames: (a) The raw detection results; (b) the tracking results; (c) the results of matching trajectories using online discriminative appearances

To sum up and inspired by Multi-object tracking approach in [23], the tracking process is then based on the method of a robust online multi-object tracking which combines a local and global tracker. In the local tracking step, we use the frame-to-frame association to generate the tracklets (*object trajectories*; which are represented by a set of multi-modal feature distributions modeled by the GMMs. In the global tracking step, the tracklet bipartite association method is used based on learning Mahalanobis metric between GMM components using KISSME[22] metric learning algorithm. The local tracker's objective is to find correct object trajectories in the past, while, the global tracker tries to find object associations between aggregated tracklets. In the first step, the tracklets are constructed by putting together frame-to-frame tracker's output. For a reliable tracklet, tracklet filtering is applied by splitting spatially disconnected or occluded tracks and filtering out noisy tracklets. In the second step, in every video segment Δt, the global tracker carries out data association and performs online tracklet matching. Association and matching process happen based on Mahalanobis metric among representations of tracklets stacked in two previous video segments ($2\Delta t$). Fig.5 and Fig.6 illustrate

an general overview of the tracking approach and the tracking framework, respectively.

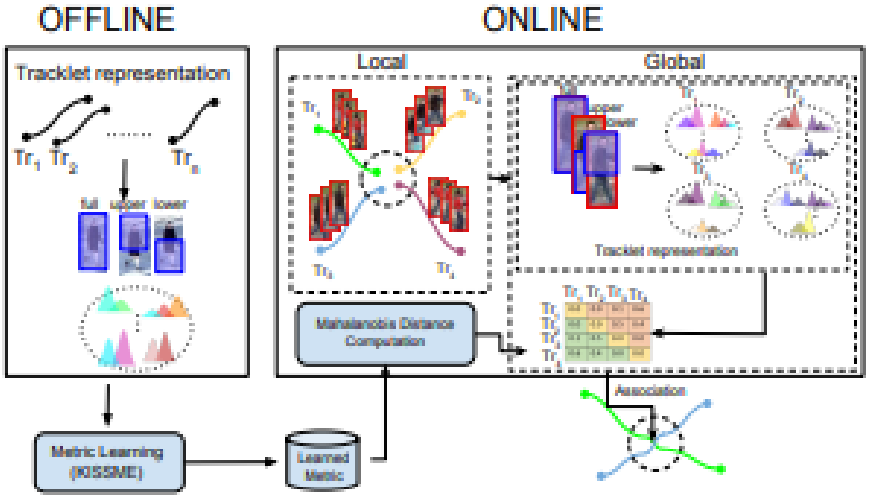


Fig. 6 Overview of the tracking framework [23]

2.2.1 Tracklet initiation and filtering

In the first step, the tracklets are constructed by putting together frame-to-frame tracker's output. For a reliable tracklet representation, tracklet filtering is applied by splitting spatially disconnected or occluded tracks and filtering out noisy tracklets. In the second step, in every video segment Δt, the global tracker carries out data association and performs data association and tracklet matching based on Mahalanobis metric. We define a tracklet Tr_i between two frames m and n as a sequence of tracked subject's bounding-boxes as follows:

$$Tr_i = \{N_i^m, N_i^{m+1}, ..., N_i^{n-1}, N_i^n\}(1)$$

Where N represents the subject bounding-box and i is the subject ID. A tracklet is defined as a connected sequence of tracked object's bounding boxes, which are the output of the global tracker. The initial tracklets should follow a filtering process. If an inconsistency in tracklet initialisation is observed, due to miss-detection, the tracklet get re-evaluated. If distance of two bounding-boxes in two consecutive frames was larger than a threshold, split operation will be followed. Otherwise, if two adjacent tracklets are occluded by each other, split operation is applied on overlapped subject detection. And also if a tracklet's length is smaller than a threshold, the tracklet is considered as noise and is eliminated. For every created tracklet, a relationship with other

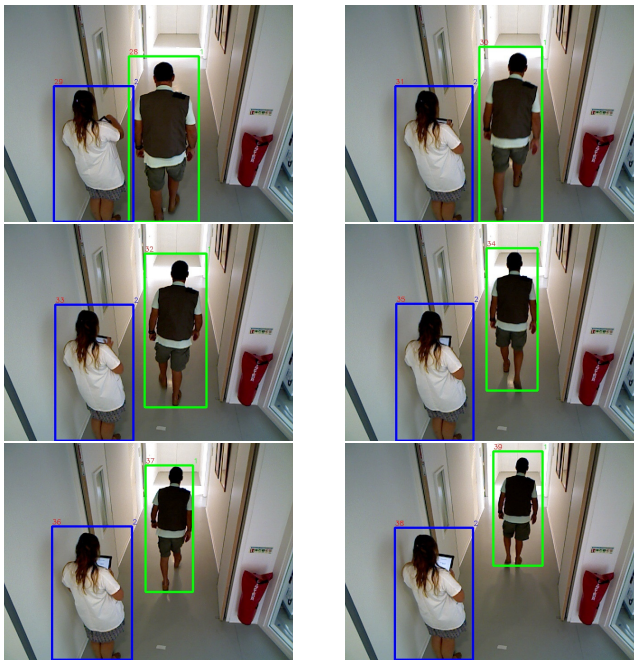


Fig. 7 A visualization of a sample of frames from CHU Nice dataset of the Tracking Results

Fig. 8 A sample of tracklets representation from CHU Nice Dataset

tracklets is defined. In fact, if the time intervals of two tracklets are overlapped than they can not associate with each other and are called neighbour. On the contrary, when a given tracklet has no time overlap and the distance between this tracklet and another one is close than we can say that this tracklet can associate with the given tracklet.

Fig.7 presents a visualization of a sample of consecutive frames from CHU Nice dataset of the tracking results.

3 Tracklet Representation based on Part Appearance Mixture (PAM) approach

Inspired by Part Appearance Mixture PAM approach in [9], and to cater the variance in a person's appearance, we model it as a multi-modal probability distribution of descriptors, using GMM to represent this appearance. Thus, the tracklets representation are modeled as a multi-channel appearance mixture (appearance model). The representation divides body into three parts: full, upper and lower. Each channel in the mixture model corresponds to a particular body part.

Due to different illumination conditions and arbitrary pose of a person with respect to camera view, the representation should stay invariable to these changes in order to have effective object tracking. By accepting variability as a natural property of appearance, we deal with it as a multi-modal probability distribution of features. To be invulnerable against occlusion, the appearance models are created independently, one for each part of the detection bounding box (full, upper and lower part of the bounding-box).Given a set of nodes (detection bounding-boxes) belonging to a tracklet Tr_i, its PAM signature representation Q is defined as a set of appearance models M_i^p :Q= {M$_q^p$ |p$\in$ $\{full, upper, lower\}$}; one for each part p of person q. Each appearance model in the set is a multivariate GMM distribution of low-level features of part p. Appearance models help to overcome occlusion, pose variation and illumination problems. To describe a subject, we use appearance features that are locally computed on spatial grid of object detection bounding-boxes; the features are computed efficiently to be shared between the parts (upper and lower body regions are defined as 60% of bounding-box of the person) including: HOG[24], LOMO[25], HSCD[26] and Color histogram features. While the framework exploits HOG feature as a shape based feature to overcome difficulties of pose variation, it benefits from other features to cope with illumination and appearance changes happening in long occlusions.

The similarity between two signatures is partially based on computing Mahalanobis distance between means of GMM components. People appearing in a video have different appearance and produce GMMs with variable number of components. Therefore, the number of components are not a priori determined and need to be retrieved. In order to infer the number of GMM components for each appearance model automatically, Akaike Information Criterion (AIC) model selection is used. Knowing fixed number of the components, the parameters of a GMM could be learned conveniently using Expectation-Maximization method.

Fig.8 shows a representation of some samples of tracklets of a person from CHU Nice dataset.

3.1 Similarity metric for Tracklet Representations

Similarity between two tracklets G_i and G_j is defined as the sum of similarities between the corresponding appearance models. Given the distance between two appearance mixtures d(M_1,M_2), we can convert this distance into similarity using Gaussian similarity kernel:

$$Sim(G_i, G_j) = \sum_{p \in \mathrm{P}} \exp(-\frac{\overline{d(M_p^q, M_p^g)} - \gamma_{p,g}}{\frac{1}{3}(\beta_{p,g} - \gamma_{p,g})}) \tag{2}$$

where P = $\{full, upper, lower\}$, $\overline{d(M_p^p, M_p^g)}$ is max normalized distance between a query person q and a gallery person g of part p. $\beta_{p,g}$ and $\gamma_{p,g}$ are the maximum and minimum normalized distances, respectively, between person g in gallery and any other person q in query set. The factor $\frac{1}{3}$ in formula makes Gaussian similarity kernel goes to zero for q that has maximum normalized distance from g. We define the distance between two GMMs as he distance between their components weighted by their prior probabilities:

$$d(M_1, M_2) = \sum_{i=1:K, j=1:K} \pi_{1i} \pi_{2j} d(G_{1i}, G_{2j}) \tag{3}$$

where G_{nk} is the component k of GMM $M_{n \in \{1,2\}}$ with corresponding prior π_{nk} and $d(G_i, G_j) = JDiv(G_i, G_j)$.

F-divergence based distances; in particular Jeffry's Divergence (JDiv) is used, and since we restrict covariance matrices to be diagonal, it can be computed as follows:

$$JDiv(G_i, G_j) = \frac{1}{2}(\mu_i - \mu_j)^T \psi (\mu_i - \mu_j) + \frac{1}{2} tr\{\sum_i^{-1} \sum_j + \sum_j^{-1} \sum_i - 2I\} \tag{4}$$

where $\psi = \sum_i^{-1} + \sum_j^{-1}$

The distance between two GMMs is computed based on the Mahalanobis distance, squared Mahalanobis distance of a pair of vectors is defined as follows:

$$d^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \tag{5}$$

where M is a positive semi-definite matrix. The parameters of Matrix M are estimated using KISSME[12]. i.e $M = \sum_+^{-1} - \sum_-^{-1}$, where $\sum_+$ and $\sum_-$ are feature-difference covariance matrices of positive and negative classes, respectively. Given the mean of two Gaussian distributions, μ_i and μ_j, the positive and negative class covariance matrices are defined as:

$$\sum_+ = \sum_{y_{ij}=+} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{6}$$

$$\sum_- = \sum_{y_{ij}=-} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{7}$$

where $y_{ij} \in \{+, -\}$ is the ground truth similarity label between pairs of Gaussian distributions (G_i, G_j). Alternatively, matrix M can be estimated using XQDA[7] in similar spirit.

3.2 Tracklet Association

Similarity matrix S=$m_{i,j}$ is constructed with calculated scores between all of condidates, where i,j=1...n and n is the number of tracklets in the time interval [t-2 $\Delta t, t$].

If a tracklet j is in the condidate list of tracklet i, the similarity of the pair is calculated using Mahalanobis metric $m_{i,j} = Sim(G_i, G_j)$, otherwise it is set to zero in the similarity matrix. Once the cost matrix is computed, the optimal association pairs, which minimize the data association cost in S, are determined using Hungarian algorithm.

4 Experiments and results

All parameters have been found experimentally, and remained unchanged for benchmark datasets. The same threshold θ= 0.3 is used for all of the data association process. The size of a video segment is fixed to 15 frames. The minimum size of a tracklet is set to 3.

4.1 Datasets

We have evaluated our work on two challenging benchmark datasets: *PRID2011* and *iLids-VID* and on our own dataset: *CHU Nice dataset*. These datasets were chosen because they provide multiple images per individual (*i.e* Multi-shot datasets) collected in realistic visual surveillance settings using two cameras.

- *PRID2011*[27]: This dataset consists of image frames extracted from two static camera recordings, depicting people walking in different directions. Images from both cameras contain variations in viewpoint, illumination, background and camera characteristics. 475 and 856 person trajectories were recorded via individual cameras, with 245 persons appearing in both views/cameras.
- *iLIDS-VID*[28]: it contains 300 identities captured in indoor scenes. It is an extended version of iLIDS dataset. It is generally believed that iLIDS-VID is more challenging than *PRID2011* due to extremely heavy occlusion.
- *CHU Nice*: Collected in the hospital of Nice (CHU) in Nice, France. It's related to INRIA Sophia Antipolis. Most of the people recruited for this dataset were elderly people, aged 65 and above, of both genders. It contains 615 videos with 149365 frames. It's also an RGB-D dataset, *i.e* it provides RGB+Depth images.

4.2 Performance Evaluation

We use the Part Appearance Mixture approach with two different image descriptors: HOG and LOMO. The image descriptors are computed just from the full body, then we extract the upper and lower descriptors from the full body descriptors. For HOG, upper and lower-body descriptors correspond to 3x6 grids aligned with top and bottom of the bounding box of the person, respectively. A full-body HOG descriptor has 792 dimensions, whereas upper-and lower-body descriptors have 432 dimensions. For full-body LOMO descriptor with 26960 dimensions is computed over 3 scales, by dividing an image in 24, 11 and 5 horizontal bands. To extract upper- and lower-body LOMO descriptors, we aggregate information over all 3 scales from 12, 6 and 3 horizontal bands aligned respectively with top or bottom of bounding box of the person.

Before we clean the dataset CHU-Nice we get as a result mAP= 58 % with the PAM approach and after we select the performant and good tracklets we improve the results to 60% as mAP for PAM and still we can improve this approach to get an operational performant Re-ID system.

Table 1 Comparison of rank-1 recognition rate (%) of our approach PAM to various Re-ID methods on PRID and iLIDS-VID. Best results are highlighted in bold.

Methods	PRID2011	iLIDS-VID	CHU-Nice
HOG3D+RankSVM[29]	19.4	12.1	-
Color+RankSVM[29]	29.7	16.4	-
ColorLBP+RankSVM[30]	34.3	23.2	-
STFV3D+KISSME[31]	64.1	43.8	-
LOMO+XQDA[32]	-	53.0	30.7
LOMO+SBSR+XQDA[33]	-	68.5	-
RFA-Net+RankSVM[34]	58.2	49.3	-
CNN+KISSME[35]	69.9	48.8	-
CNN+XQDA[35]	77.3	53.0	-
PAM-HOG+KISSME	55.3	33.9	-
PAM-LOMO+XQDA	-	-	38.5
PAM-LOMO+KISSME	**92.5**	**79.5**	**81.8**

4.3 Evaluation of signature representation quality

As shown in Fig.9, a visualization of a selected sample from CHU Nice dataset of PAM signature representation is presented. Indeed, the first image corresponds to one of the input images used to learn appearance model. Its followed by the composite images, one for each component of the GMM mixture. Optimal number of GMM components for each appearance model varies between persons. GMM components focus on different pose and orientation of the person. Moreover, We visualize each GMM component by constructing a composite image. In fact, given appearance descriptor, we compute the likelihood of an image belonging to a model component and then by summing

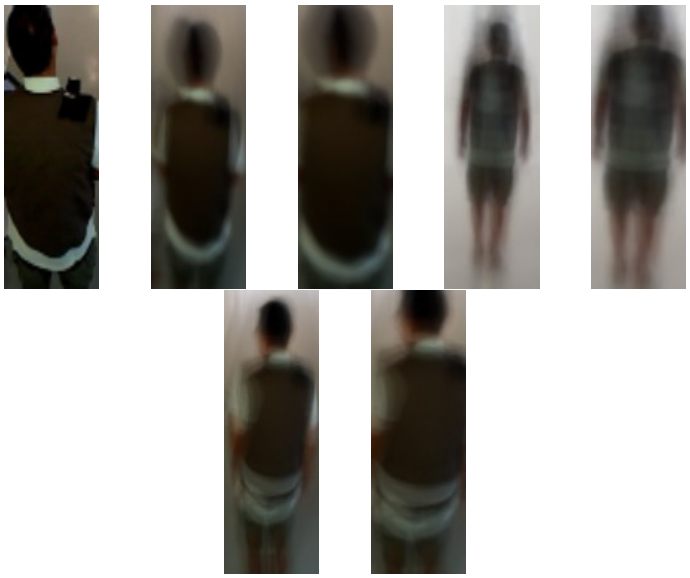


Fig. 9 A visualization of selected samples of signature representation from CHU Nice Dataset

Fig. 10 A sample of tracklets with bad/miss detection

images of corresponding person weighted by their likelihood we generate the composite image. We can say that our signature representation is able to cater variance in person's pose and orientation as well as illumination, it deals also with occlusions and is able to reduce effect of background. However, we can notice that this PAM signature present some limitations, specially on our own dataset CHU Nice, which can affect the quality of our signature representation (see Fig.10, 11, 12, 13). Among these challenging problems, we can cite:

- Occlusions
- illumination changes
- scale-adaptation and alignment problem
- Bad/miss detection
- Number of frames by pose
- Number of GMM components not adequate with the number of person's pose/orientation and depends of the low-level features used.

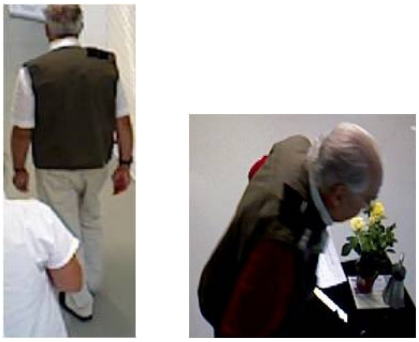


Fig. 11 A sample of tracklets with occlusions

Fig. 12 A sample of tracklets with illumination changes

Fig. 13 A sample of tracklets with scale-adaptation and alignment problem

5 Conclusion

Person Re-ID is a challenging task with three aspects: First, it is important to determine which parts should be segmented and compared. Second, there is a need to generate invariant signatures for comparing the corresponding parts. Third, an appropriate matching function (i.e similarity metric or a ranking function) must be applied to compare these signatures.

In most studies, the Re-ID process/system is designed under the assumption of perfect detection and tracking and with the idea that the appearance of persons is unchanged and which don't seem reasonable in practise. Therefore, we present in this paper the tracklets pre-processing beginning from the detection step, tracking and then the person Re-Identification step in order to compute the signature of persons and represent it properly based on PAM approach which uses multiple appearance models based on GMM model. Each

appearance is described as a probability distribution of some low-level features for a certain part of person's body. Indeed, this improves the appearance descriptors and deals with occlusions and variance in pose/orientation of individuals. The robustness of the quality of this signature representation is verified by extensive experiments. Moreover, we are trying to study the whole Re-ID surveillance system in order to present a good/autonomous/operationnel one. We discuss how different components of the surveillance system affect the overall performance and we proposed a method to improve it.

As future work, we are trying to improve the proposed approach (PAM) by using the skeleton and extracting the pose machines from our dataset, *i.e* CHU Nice dataset, which will be soon introduced as a new public dataset in the field of person Re-ID.

References

1. Zheng, L., Yang, Y., Hauptmann, A. G. (2016) Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984.
2. Mazzon, R., Tahir, S. F., Cavallaro, A. Person re-identification in crowd. Pattern Recognition Letters, 33(14), 1828-1837, 2012.
3. Vezzani, R., Baltieri, D., Cucchiara, R. (2013) People reidentification in surveillance and forensics: A survey. ACM Computing Surveys (CSUR), 46(2), 29.
4. Saghafi, M. A., Hussain, A., Zaman, H. B., Saad, M. H. M. (2014) Review of person re-identification techniques. IET Computer Vision,8(6), 455-474.
5. Bedagkar-Gala, A., Shah, S. K. (2014) A survey of approaches and trends in person re-identification. Image and Vision Computing, 32(4), 270-286.
6. Jasher Nisa, A.J., Sumithra, M.D. (2016) A Review on Different Methods of Person Re-identification. JETIR, Volume 3, Issue 6.
7. T.L.A. Nguyen, P. Chau and F. Bremond, *Robust Global Tracker based on an Online Estimation of Tracklet Descriptor Reliability*, In Proceedings of the 17th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Part of AVSS 2015, Karlsruhe, Germany, 25 August 2015.
8. T.L.A. Nguyen, F. Bremond and J. Trojanova, *Multi-Object Tracking of Pedestrian Driven by Context*, In Proceedings of the 13th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, in Colorado Springs, Colorado, USA, 24-26 August 2016.
9. F. M. Khan and F. Bremond, *Multi-shot Person Re-Identification using part apperance mixture*, In Proceedings of the Winter Conference on Applications of Computer Vision, WACV, 27-29th March 2017.
10. F. Pala, R. Satta, G. Fumera, and F. Roli, *Multimodal person Re-Identification using RGB-D cameras*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no 4, p.788-799, 2016.
11. S. Ren, K. He, R. Girshick, J. Sun, *Faster R-CNN: Towards real-time object detection with region proposal networks*. In Advances in neural information processing systems, pp. 91-99, 2015.
12. J. Redmon, S. Divvala, R.Girshick, A. Farhadi, *You only look once: Unified, real-time object detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.
13. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, *SSD: Single Shot multibox Detector*, In European conference on computer vision,pp. 21-37, Springer, October 2016.
14. F. POIESI, R. MAZZON, A. CAVALLARO. Multi-target tracking on confidence maps: An application to people tracking, in "Computer Vision and Image Understanding", 2013, vol. 117, no 10, pp.1257 - 1272.

15. G. SHU, A. DEHGHAN, O. OREIFEJ, E. HAND, M. SHAH. Part-based multiple-person tracking with partial occlusion handling, in "2012 IEEE Conference on Computer Vision and Pattern Recognition", June 2012, pp.1815-1821.
16. L. ZHANG, Y. LI, R. NEVATIA. Global data association for multi-object tracking using network flows, in "Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on", June 2008, pp. 1-8.
17. A. ROSHAN ZAMIR, A. DEHGHAN, M. SHAH. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs, in "Proceedings of the European Conference on Computer Vision (ECCV)", 2012.
18. S. H. BAE, K. J. YOON. Robust Online Multi-object Tracking Based on Tracklet Confidence and Online Discriminative Appearance Learning, in "2014 CVPR", June 2014, pp. 1218-1225.
19. N. THI LAN ANH, F. BREMOND, J. TROJANOVA. Multi-Object Tracking of Pedestrian Driven by Context, in "Advance Video and Signal-based Surveillance", Colorado Springs, United States, IEEE, August 2016.
20. S. LIAO, Y. HU, S. Z. LI. Joint Dimension Reduction and Metric Learning for Person Re-identification, in "CoRR", 2014, vol. abs/1406.4216.
21. M. ZENG, Z. WU, C. TIAN, L. ZHANG, L. HU. Efficient person re-identification by hybrid spatiogram and covariance descriptor, in "2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", June 2015, pp. 48-56.
22. M. KOSTINGER, M. HIRZER, P. WOHLHART, P. M. ROTH, H. BISCHOF. Large scale metric learning from equivalence constraints, in "2012 CVPR", June 2012, pp. 2288-2295.
23. T.L.A. Nguyen, F.M. Khan, F. Negin and F. Bremond, *Multi-Object tracking using Multi-Channel Part Appearance Representation*, In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, in Lecce, Italy, 29 August-1st September, 2017.
24. N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection*, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 886893, June 2005.
25.].S. Liao, Y. Hu, and S. Z. Li, *Joint dimension reduction and metric learning for person re-identification*. CoRR,abs/1406.4216, 2014.
26. M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu, *Efficient person re-identification by hybrid spatiogram and covariance descriptor*. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 4856, June 2015.
27. M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, *Person re-identification by descriptive and discriminative classification*. In Scandinavian conference on Image analysis, pp. 91-102, Springer, May 2011.
28. T. Wang, S. Gong, X. Zhu and S. Wang, *Person re-identification by video ranking*. In European Conference on Computer Vision, ECCV pp. 688-703, Springer, September 2014.
29. T. WANG, S. GONG, X. ZHU, S. WANG. Person Re-identification by Video Ranking, in "ECCV", 2014
30. M. HIRZER, P. ROTH, M. KSTINGER, H. BISCHOF. Relaxed pairwise learned metric for person reidentification, in "ECCV", 2012
31. K. LIU, W. ZHANG, R. HUANG. A Spatio-Temporal Appearance Representation for Video-Based Pedestrian Re-Identification, in "ICCV", 2015.
32. S. LIAO, Y. HU, X. ZHU, S. Z. LI. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning, in "CVPR", 2015.
33. S. CHAN-LANG, Q. PHAM, C. ACHARD. Bidirectional Sparse Representations for Multi-Shot Person Reidentification, in "AVSS", 2016
34. Y. YAN, B. NI, Z. SONG, C. MA, Y. YAN, X. YANG. Person Re-identification via Recurrent Feature Aggregation, in "ECCV", 2016
35. L. ZHENG, Z. BIE, Y. SUN, J. WANG, C. SU, S. WANG, Q. TIAN. MARS: A Video Benchmark for Large- Scale Person Re-identification, in "ECCV", 2016