Tracklet and Signature Representation for Multi-Shot Person Re-Identification

Salwa BAABOU*‡, Furqan M. KHAN†, Francois BREMOND†, Awatef BEN FRADJ‡,
Mohamed Amine FARAH‡ and Abdennaceur KACHOURI‡

*University of Gabes
National Engineering School of Gabes, Tunisia
Email: baabousalwa@gmail.com
†INRIA Sophia Antipolis,
2004 Route des Lucioles -BP93 Sophia Antipolis Cedex, 06902, France
Email: (furqan.khan | francois.bremond)@inria.fr
‡University of Sfax
National Engineering School of Sfax, Laboratory of Electronics and Information Technology (LETI),
B.P.W.3038, Sfax, Tunisia
Email: benfradj_awatef@yahoo.fr
med.farah@yahoo.fr
abdennaceur.kachouri@enis.rnu.tn

***Abstract*—Video surveillance has become more and more important in many domains for their security and safety. Person Re-Identification (Re-ID) is one of the most interesting subjects in this area. 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, appearance based person Re-ID is a challenging task due to similarity of human's appearance and visual ambiguities across different cameras. This paper provides a representation of the appearance descriptors, called *signatures*, for multi-shot Re-ID. First, we will present the tracklets, *i.e* trajectories of persons. Then, we compute the signature and represent it based on the approach of Part Appearance Mixture (PAM). 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.**

***Index Terms*—Person Re-Identification (Re-ID), Part Appearance Mixture (PAM), tracklet, signature representation**

I. Introduction

Person Re-Identification (Re-ID) aims to match individuals appearing across non-overlapping camera networks at distinct times and locations. A person's appearance across different cameras is very variable, which makes the recognition of individuals more and more challenging.

A typical Re-ID system may have an image (single shot) or a video (multi-shot) as input for feature extraction and signature generation. Thus, the first step in Re-ID is to learn a person's visual signature or model and then compare the two models to get either a match or a non-match. Extracting a reliable signature depends on the availability of good observations. Besides, faulty trajectory estimation and incorrect detections introduce errors in signature generation and extraction that affect the Re-ID quality. The most obvious and simplest signature of a person is characterized by low-level features like color, texture and shape. However, these features are hardly unique, not descriptive enough and prone to variations. Color/texture varies due to cross view illumination variations, pose variations, view angle or scale changes in multi-camera settings.

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)[20] 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. All these issues may affect the performance of person Re-ID which is still not robust enough to guarantee high accuracy in practice.

To sum up, the contribution of this paper is: how to pre-process the tracklets to make them good for computing the signature and then represent it for multi-shot Re-ID based on PAM approach[1]. 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.

The paper is organized as follows: The Re-ID process which contains person detection and tracking is

978-1-5386-5305-0/18/$31.00 ©2018 IEEE

described in the following section. Section III is the core of the paper: it introduces 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 quality of our signature representation based on the realized experiments and results before concluding.

II. Re-ID Process

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 [2, 3, 4, 5, 6, 7, 8] have presented them in detail.

Fig.1 illustrates the Re-ID process, *i.e* the diagram of person Re-Identification system, 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. 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.

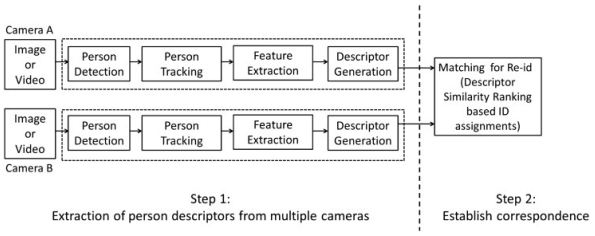

Fig. 1: Person Re-ID diagram

A. 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.

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 [21] 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[22] and SSD[9].

In this paper, we will use the SSD detector[9]. Fig.2 shows a visualization of a sample from CHU Nice dataset of the detection results. In fact, the SSD detector differs from other single shot detectors 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.

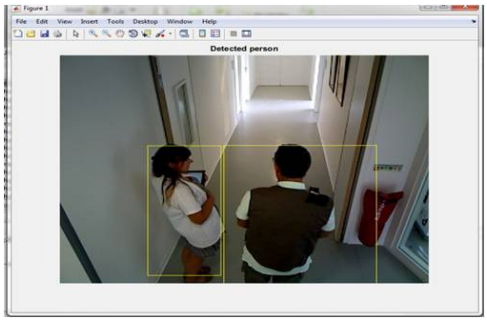

Fig. 2: A visualization of a sample from CHU Nice dataset of the Detection Results

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.

B. Multi-Object Tracking

Insprired by Multi-object tracking approach in [10], the tracking process is 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[11] 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 on line tracklet matching. Association and matching process happens based on Mahalanobis metric among representations of tracklets stacked in two previous video segments ($2\Delta t$).

Fig.3 presents a visualization of a sample of consecutive frames from CHU Nice dataset of the tracking results.

III. Part Appearance Mixture (PAM) Approach

A. Signature Representation

We define a tracklet Tr_i between two frames m and n as a sequence of tracked subject's bounding-boxes as follows:

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

Where N represents the subject bounding-box and i is the subject ID.

Fig 4 shows a representation of some samples of tracklets of a person from CHU Nice dataset.

Inspired by Part Appearance Mixture PAM approach in [1], 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.

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= $\{\mathrm{M}_q^p \ |\mathrm{p}\in \ \{full, upper, lower\}\}$; 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 an object, 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[12], LOMO[13], HSCD[14] 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.

B. Similarity metric for signature representation

As mentioned before, a signature is a set of part appearance mixtures. Similarity between two signatures G_i and G_j is defined as the sum of similarities between

Fig. 3: A visualization of a sample of frames from CHU Nice dataset of the Tracking Results

Fig. 4: A sample of tracklets representation from CHU Nice Dataset

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:

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

where $\mathrm{P} = \{full, upper, lower\}$, $\overline{d(M_p^p, M_p^g)}$ 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:

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

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

C. Distance computation between signatures

To define similarity between two signatures, 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:

$$JDiv(G_i, G_j) = \frac{1}{2}(\mu_i - \mu_j)^T \psi (\mu_i - \mu_j) + \frac{1}{2}$$

$$\mathrm{tr}\ \{\textstyle\sum_i^{-1} \sum_j + \sum_j^{-1} \sum_i - 2I\} \quad (4)$$

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:

$$d^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \quad (5)$$

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:

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

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

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[18] in similar spirit.

TABLE I: Top ranked Recognition rates (%) on *PRID2011*

Methods	Rank-1	Rank-5	Rank-10	Rank-20
LOMO+XQDA	-	-	-	-
PAM-HOG+KISSME	55.3	80.7	90.2	95.6
PAM-LOMO+XQDA	-	-	-	-
PAM-LOMO+KISSME	92.5	99.3	100.0	100.0

TABLE II: Top ranked Recognition rates (%) on *iLIDS-VID*

Methods	Rank-1	Rank-5	Rank-10	Rank-20
LOMO+XQDA	53.0	78.5	86.0	93.4
PAM-HOG+KISSME	33.9	60.0	70.2	79.1
PAM-LOMO+XQDA	-	-	-	-
PAM-LOMO+KISSME	79.5	95.1	97.6	99.1

IV. Experiments and results

A. 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*[15]: 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*[16]: 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.

B. 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.

TABLE I, TABLE II and TABLE III show the recognition rate (%) at different ranks (rank-1, 5, 10, 20) of a baseline method LOMO+XQDA[17] and PAM-LOMO+XQDA on PRID2011, iLIDS-VID and CHU Nice datasets, respectively.

From the above experiments, we notice that PAM-LOMO+KISSME achieves good performance on three datasets; it achieves 92.5%, 79.5% and 81.8% rank-1 recognition rates on PRID2011, iLIDS-VID and CHU Nice datasets, respectively. This shows that our adaptation of feature descriptor LOMO and metric learning KISSME to PAM representation is effective.

TABLE III: Top ranked Recognition rates (%) on *CHU Nice dataset*

Methods	Rank-1	Rank-5	Rank-10	Rank-20
LOMO+XQDA	30.7	64.6	80.3	90.3
PAM-HOG+KISSME	-	-	-	-
PAM-LOMO+XQDA	38.5	69.2	84.6	100.0
PAM-LOMO+KISSME	81.8	90.9	100	100

C. Evaluation of signature representation quality

As shown in Fig 5, 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 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 5). Among these challenging problems, we can cite:

- Bad 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.

V. Conclusion

Person Re-ID is a challenging task with three aspects: First, is 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 is designed under the assumption that the appearance of the person is unchanged which doesn't seem reasonable in practise.

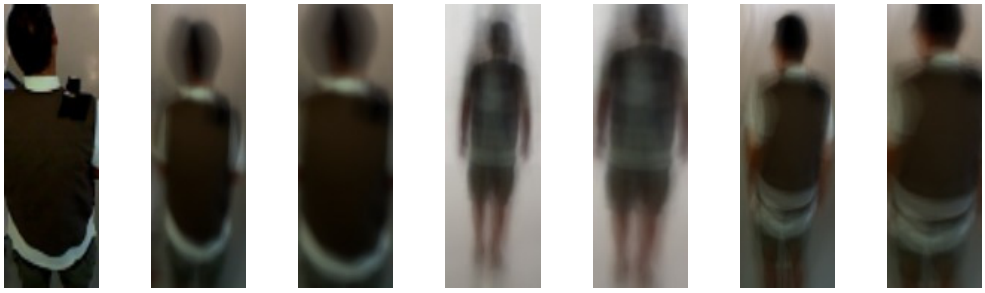

Fig. 5: A visualization of selected samples of signature representation from CHU Nice Dataset

Therefore, we present in this paper a signature representation of persons 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.

As future work, we are trying to improve the PAM signature representation, 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] 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.

[2] M. Paul, S. M. Haque, and S. Chakraborty, *Human detection in surveillance videos and its applications-a review*, EURASIP Journal on Advances in Signal Processing, no 1, p.176, 2013.

[3] T. B. Moeslund and E. Granum, *A survey of computer vision-based human motion capture*, Computer vision and image understanding, 81(3), p.231-268, 2001.

[4] A. Yilmaz, O. Javed and M. Shah, *Object tracking: A survey*, Acm computing surveys (CSUR), 38(4), 13, 2006.

[5] R. Vezzani, D.Baltieri and R. Cucchiara, *People reidentification in surveillance and forensics: A survey*, ACM Computing Surveys (CSUR), vol. 46, no 2, p.29, 2013.

[6] A. Bedagkar-Gala, and S. K. Shah, *A survey of approaches and trends in person re-identification*, Image and Vision Computing, vol.32, no 4, p.270-286, 2014.

[7] X. Wang, R. Zhao, *Person Re-Identification: system design and evaluation overview*, In Person Re-Identification, Springer London, p.351-370, 2014.

[8] A.J. Jasher Nisa, M.D. Sumithra, *A review on different Methods of person Re-Identification*, JETIR, v.3, Issue 6, June 2016.

[9] 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.

[10] 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.

[11] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, *Large scale metric learning from equivalence constraints*. In CVPR, pages 2288-2295, June 2012.

[12] 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.

[13]].S. Liao, Y. Hu, and S. Z. Li, *Joint dimension reduction and metric learning for person re-identification*. CoRR,abs/1406.4216, 2014.

[14] 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.

[15] 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.

[16] 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.

[17] S. Liao, Y. Hu, X. Zhu and S. Z. Li, *Person re-identification by local maximal occurrence representation and metric learning*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2197-2206, 2015.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.