PERSON RE-IDENTIFICATION WITH A PTZ CAMERA: AN INTRODUCTORY STUDY

Pietro Salvagnini[†], *Loris Bazzani*[†]

[†]Istituto Italiano di Tecnologia
Pattern Analysis & Computer Vision
Via Morego 30, 16163 Genova - Italy

Marco Cristani[†,‡], *Vittorio Murino*[†]

[‡]Università degli Studi di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona - Italy

ABSTRACT

We present an introductory study that paves the way for a new kind of person re-identification, by exploiting a single Pan-Tilt-Zoom (PTZ) camera. PTZ devices allow to zoom on body regions, acquiring discriminative visual patterns that enrich the appearance description of an individual. This intuition has been translated into a statistical direct re-identification scheme, which collects two images for each probe subject: the first image captures the probe individual, focusing on the whole body; the second can be a zoomed body part (head, torso or legs) or another whole body image, and is the outcome of an action-selection mechanism, driven by feature selection principles. The validation of this technique is also explored: in order to allow repeatability, two novel multi-resolution benchmarks have been created. On these data, we demonstrate that our approach selects effective actions, by focusing on body portions which discriminate each subject. Moreover, we show that the proposed compound of two images overwhelms standard multi-shot descriptions, composed by many more pictures.

Index Terms— Person Re-identification, Pan-Tilt-Zoom camera

1. INTRODUCTION

People re-identification (re-id) has definitely become a primary module for multi-camera video surveillance systems, allowing to recognize individuals across different locations and times. Now mature, the re-id literature is partitioned in direct VS learning-based and single-shot VS multi-shot methods. Direct approaches [1, 2, 3] are on-line feature extractors, while learning-based techniques [4, 5, 6, 7, 8, 9, 10] require a training step prior to work. Single-shot [1, 2, 3, 5, 6, 7, 8] and multi-shot [1, 2, 4, 8, 9] approaches differ for the number of images exploited to describe each probe or gallery subject.

This work adds a novel point of view to this taxonomy, introducing the usage of a PTZ camera for the re-id problem. PTZ cameras are nowadays widespread in many different environments (stadiums, banks, crossroads) and many tracking algorithms have been developed to automatically zoom on particular areas of interest [11, 12, 13].

Fig. 1. PTZ re-id datasets: each acquisition of a subject is characterized by four images: the whole-body image (left) and the three zoomed images portraying body parts at high resolution (right). Please note, for each individual we have multiple acquisitions.

Assuming this technology available, we present a PTZ protocol composed by a set of actions (*zoom on the head, torso, legs, do not zoom*), which produces a two-image description of each probe subject. The first image focuses on the whole body, and it is supposed to be acquired during an initialization phase, where the PTZ camera looks for individuals (as in a usual pedestrian detection scenario). The second image is collected after performing an action selected from the set of possible actions. To this end, we propose a max-variance action selection algorithm that is built upon the feature selection principles [14, 15]. The idea is that the method chooses the action that exhibits the highest variance in the feature space, because it is supposed to better discriminate between different individuals. In addition, we show how the two images are properly integrated to create a composite description of the individual for re-id.

The proposed method has been tested with the aim of simulating the behavior of a PTZ camera and of providing repeatable benchmarks often hard when using PTZ cameras. To this end, we propose a novel evaluation protocol that is based on building a dataset with the whole-body image at low resolution and the three body parts at high resolution for each subject (some examples are shown in Fig. 1).

The results highlight different facets of our proposal, promoting the use of PTZ for boosting the re-id performances. Moreover, we demonstrate that our two-image description for

PTZ cameras is far better than exploiting ten whole-body images in a fixed-camera multi-shot setup.

2. PTZ RE-IDENTIFICATION

In general, re-id aims at finding in the *gallery* set $\mathcal{G}$ the description $\mathbf{g}_j,\ j \in \{1,...,J\}$ related to the individual with identity label $l(j)$, such that it corresponds to the description $\mathbf{q}_i,\ i \in \{1,...,I\}$ in the *probe* set $\mathcal{Q}$, that is, $l(i) = l(j)$. In the direct approaches, this amounts to minimize an opportune distance $d(i,j)$. Single or multiple-shot re-id indicate the need of having a single or multiple images for building a description, respectively.

The proposed re-id scheme with PTZ cameras assumes that the camera may capture two kinds of images: whole body (with no zoom) and zoomed (where a single part of the body is recorded using the zoom). In the following, we refer to them as *whole* and *zoomed* images, respectively. Therefore, the gallery set for PTZ re-id consists of a whole image $\mathbf{g}_j^{a=0}$ and the set of zoomed images $\{\mathbf{g}_j^a\}_{a=\{1,2,3\}}$ for each single individual, where a indexes the set $\mathcal{A}$ of *actions* that the camera can perform, *i.e.*, getting another whole image (a=0) or zooming on the face (a=1), torso (a=2), legs (a=3).

The process of PTZ re-id is carried out following a procedure composed by five steps, sketched in Fig. 2.

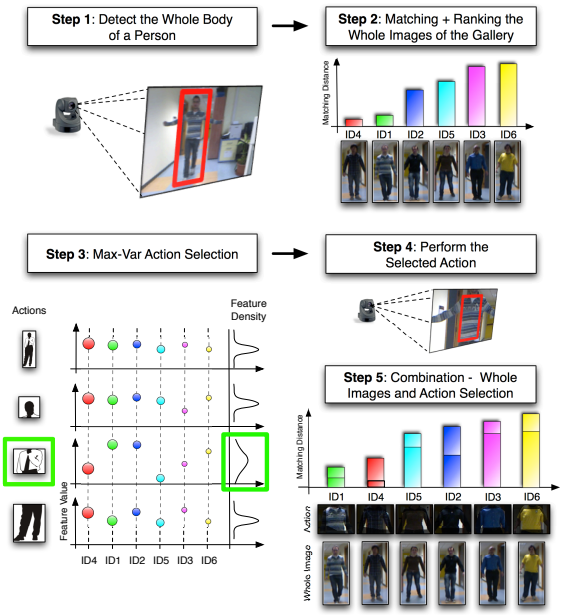

Fig. 2. The proposed method first detects the whole body of the probe and a first round of re-id is performed. The max-var action selection algorithm exploits this information to select the most informative part and performing a second attempt of re-id that combines the two images.

The first step detects the whole probe image $\mathbf{q}_i^0$.[1] The second step performs re-id by considering the entire gallery of whole images $\{\mathbf{g}_j^0\}_{j=1}^J$, thus obtaining a preliminary re-id ranking defined by the distance $d(\mathbf{q}_i^0, \mathbf{g}_j^0)$ (distances will be detailed later). The idea is that images with higher ranking (lower distance value) are more similar with the probe, and thus more important.

In the third step, the *max-var action selection* algorithm chooses an action $a \in \mathcal{A}$, deciding whether to zoom on a single part ($a = \{1,2,3\}$) or to keep the focus on the entire body ($a = 0$). To this end, a weighted variance on the features that characterizes the different parts or the whole body is calculated, across all the corresponding elements of the gallery. The weight is related to the distance between the features, the lower the distance the higher the weight. Features which exhibit high weighted variance indicate parts that are highly discriminant for the different identities, considering standard principles of feature selection. The details of the max-var action selection are reported in the next section. In the fourth step, the action is performed by the PTZ camera, obtaining the second probe image $\mathbf{r}_i^a$, $a = \{0,1,2,3\}$. The final step combines the images of the probe $\{\mathbf{q}_i^0, \mathbf{r}_i^a\}$ with the gallery images $\{\mathbf{g}_j^0, \mathbf{g}_j^a\}$, employing a proper distance $d_{joint}(\cdot,\cdot)$ and obtaining the final re-id ranking[2].

2.1. Max-Var Action Selection

Let us describe the re-id task in a probabilistic way, defining x as the random variable representing the unknown probe identity label $l(i)$ that we want to associate to the j-th gallery description with label $l(j)$.

As discussed above, the distance $d(\mathbf{q}_i^0, \mathbf{g}_j^0)$ provides an initial guess of the probe identity. This is formulated as a probability density for x, defined as:

$$p(x = l(j) \,|\, \mathbf{q}_i^0) = \frac{e^{-\lambda\, d(\mathbf{q}_i^0, \mathbf{g}_j^0)}}{\sum_j e^{-\lambda\, d(\mathbf{q}_i^0, \mathbf{g}_j^0)}} \equiv w_{i,j} \qquad (1)$$

where λ regulates how peaked is the density (high λ for peaked distributions), and it has been optimized in our experiments. We remark that this distance is defined over whole images, as it is reasonable that the first action performed by the PTZ camera is to detect a probe individual.

The action a defines the second probe image $\mathbf{r}_i^a$, and it should maximize the information available in the gallery, given the probe image $\mathbf{q}_i^0$. Taking inspiration from widely-known principles of feature selection based on maximum variance [14, 15], the best action $a^\star$ is defined as follows:

$$a^\star = \underset{a}{\text{argmax}}\ \text{var}\left[\mathbf{r}_i^a | \mathbf{q}_i^0, a\right] \qquad (2)$$

[1] From now on we shorten the apex a=y with y, unless the context requires otherwise.

[2] Please note that, when the selected action is $a = 0$ (*i.e.*, the whole body image) the gallery images $\mathbf{g}_i^0$ and $\mathbf{g}_i^a$ actually coincide.

Algorithm 1: The Max-Var action selection method chooses the action that fetches the gallery descriptions with the higher feature variance for each probe individual.

Input: Probe individual $i \in \mathcal{Q}$
Compute the weights $w_{i,j}$, $\forall j \in \mathcal{G}$ using Eq. 1
forall the $a \in \mathcal{A}$ **do**
  Compute the mean using Eq. 6
  Compute the variance using Eq. 7
end
Select $a^\star$ using Eq. 2
Output: Selected action $a^\star$

where $\mathrm{var}\left[\mathbf{r}_i^a|\mathbf{q}_i^0, a\right]$ is the variance of $\mathbf{r}_i^a$ under the probability density $p(\mathbf{r}_i^a|\mathbf{q}_i^0, a)$. To estimate the variance in Eq. 2, we decompose $p(\mathbf{r}_i^a|\mathbf{q}_i^0, a)$ as follows:

$$p(\mathbf{r}_i^a|\mathbf{q}_i^0, a) = \int p(\mathbf{r}_i^a|x, a)p(x|\mathbf{q}_i^0)dx \tag{3}$$

where $p(x|\mathbf{q}_i^0)$ is defined in Eq. 1. The first term in the integral of Eq. 3 is defined as:

$$p(\mathbf{r}_i^a|x, a) = \delta(\mathbf{r}_i^a - \mathbf{g}_j^a), \quad l(j) = x. \tag{4}$$

where $\delta(\cdot)$ selects the individual j in the gallery set that has label x. In practice, Eq. 4 selects the gallery description related to action a, assuming it corresponds to the probe description that we have not yet acquired.

The integral in Eq. 3 therefore becomes a weighted sum over the gallery set:

$$p(\mathbf{r}_i^a|\mathbf{q}_i^0, a) = \sum_{j \in \mathcal{G}} w_{i,j}\, \delta(\mathbf{r}_i^a - \mathbf{g}_j^a), \tag{5}$$

The mean and the variance of such discrete distribution can be easily derived as follows:

$$\mathbb{E}\left[\mathbf{r}_i^a|\mathbf{q}_i^0, a\right] = \sum_j w_{i,j}\, \mathbf{g}_j^a \equiv \mu_i^a. \tag{6}$$

$$\mathrm{var}\left[\mathbf{r}_i^a|\mathbf{q}_i^0, a\right] = \sum_j w_{i,j}\, (\mathbf{g}_j^a - \mu_i^a)^\top (\mathbf{g}_j^a - \mu_i^a). \tag{7}$$

The action selection strategy is summarized in Alg. 1.

2.2. Features

All the images, whole and zoomed, are described by the chromatic and structural features used in most of the re-id approaches. The chromatic description is inspired by [2], given by the concatenation of the normalized 2-dimensional Hue-Saturation (HS) histogram $\mathbf{r}_i^{HS}$ and the normalized value histogram $\mathbf{r}_i^V$ as: $\mathbf{r}_i^a = [\gamma\, \mathbf{r}_i^{HS},\ (1-\gamma)\, \mathbf{r}_i^V]$, where γ (set to 0.95 in our experiments) is the weight that balances the importance of each feature. Textural information is modeled by the Local Binary Pattern histogram descriptor [16]. We also tested the combination of the two descriptors, dubbed HSV+LBP, which is obtained simply concatenating the two normalized histograms. Before all the features were computed, the pedestrian images were equalized on each RGB channel independently. For the second probe image, where the whole pedestrian is not completely visible, we applied the equalization transformation computed in the first image.

2.3. Distances

Our approach deals with two kinds of distances: the $d(\cdot,\cdot)$, which is the standard Bhattacharyya distance between normalized histograms as in [17], and $d_{joint}(\cdot,\cdot)$ that compares the complete probe description $\{\mathbf{q}_i^0, \mathbf{r}_i^{a^\star}\}$ and the gallery description $\{\mathbf{g}_i^0, \mathbf{g}_i^{a^\star}\}$, and that consists in a linear combination of Bhattachryya distances:

$$d_{joint}(i,j) = (1-\alpha)\, d(\mathbf{q}_i^0, \mathbf{g}_j^0) + \alpha\, d(\mathbf{r}_i^{a^\star}, \mathbf{g}_j^{a^\star}) \tag{8}$$

where d applies indistinctively to whole or zoomed images depending on the best action selected $a^\star$; α is a parameter that weights the combination of the two Bhattachryya distances (set to 0.5 in the experiments).

3. EXPERIMENTS

The ideal experimental setup for benchmarking PTZ applications should be a repeatable scenario, in which the different actions can be tested with the identical input. In practice, this is often not possible, and most of the *repeatable* tests with PTZ cameras use synthetic data [18]. Since testing re-id using real data is very important, we proposed a validation protocol that exploits realistic, existing re-id datasets. Starting from the original images, they have been bi-linearly downsampled (factor 1/4) to simulate the whole images $\{\mathbf{q}_i^0\}$ and the correspondent gallery images $\{\mathbf{g}_j^0\}$. The whole person bounding boxes and the parts are obtained from the whole image by applying the pictorial structures detector [19].

To ensure that the produced images mimic genuine PTZ imagery (SONY SNC-RX550P camera), we analyze the difference between a downsampled image taken at zoom 4x and the corresponding area of an image at zoom 1x so that they both share the same field of view. We obtained an RMSE of 30 (summing over all the RGB channels), averaged over 10 image pairs. As comparison, the RMSE between two consecutive images framed at constant zoom is 10.

Following this line, we adapted two public datasets, namely RGB-Did [20] and iLIDS-MA [21], for our purpose. RGB-Did contains 93 people, where occlusions do not occur, but the pose of the individual and the illumination of the scene change. iLIDS-MA contains 40 individuals manually extracted from two cameras by [22], and brings occlusions into play. In both the datasets we select for each person 3 images, maximizing the variations in pose and light between the

Table 1. Re-id results on the RGB-Did dataset in terms of CMC(1) and nAUC (between brackets). In bold the best score, in italic the second best score.

	No Action	Fixed Action				Action Selection	
	Low	Low + Face	Low + Torso	Low + Legs	Low + Low	Low + Max-Var	Low + Oracle
HSV	62.37 (96.98)	*68.82* (97.75)	63.44 (98.64)	66.67 *(98.10)*	62.37 (96.84)	**68.82** **(98.74)**	78.49 (99.29)
LBP	11.83 (80.76)	21.51 (86.59)	*21.51* (86.22)	16.13 *(86.67)*	15.05 (81.43)	**22.58** **(87.86)**	31.18 (92.13)
HSV+ LBP	60.22 (97.49)	67.74 (98.36)	68.82 **(98.99)**	*72.04* (98.31)	59.14 (97.17)	**73.12** *(98.71)*	82.80 (99.40)

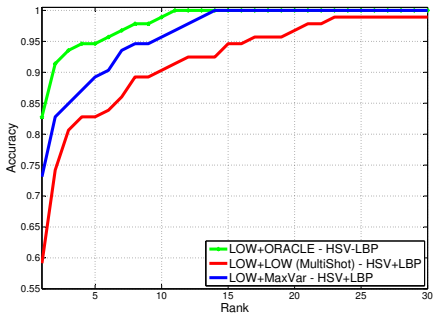

Fig. 3. CMC (first 30 ranks) on the RDG-Did dataset for different approaches.

probe and the gallery and preferring large images (averagely, 150×350). From them we obtain the complete descriptions (whole + zoomed) as described above. Two images form the probe images and one populates the gallery. The two probe images have been selected with an interval of few frames between them, to simulate the time a PTZ requires to perform an action. In general, the whole person bounding boxes are of size 32x64, while the single parts bounding boxes have average size 35x30 (head), 60x60 (torso), 100x50 (legs) (see Fig. 1).

Results. The standard metrics for re-id are used to evaluate the proposed method: the Cumulative Match Curve (CMC), the normalized Area Under the CMC (nAUC) and the first rank in the CMC (CMC(1)).

We first consider the RGB-Did dataset, performing three different tests, whose results are reported in Table 1. The first column (no action) reports the results when only one image at low resolution is used as description, representing thus the standard single-shot scenario. The second block of experiments test fixed-action policies (*i.e.*, after the whole image, takes always the head, or the torso, or the legs, or another whole image). The last two columns report the max-var action selection performance and the oracle results, where the oracle selects the action giving the highest re-id accuracy knowing the ground truth.

The considerations are: i) as expected, the single image approach (column 1) performs the worst; ii) the proposed approach is better for CMC(1) to all the fixed policies (columns 2-4) for all the features, reaching the best performances considering nAUC except in one case; iii) having a zoomed image is better than exploiting another whole image; iv) there is still room of improvement, looking at the results of the oracle. The CMC curves are reported in Fig. 3, showing the behavior of HSV+LBP which gives the best performance.

Table 2. Re-id results on the i-LIDS-MA dataset in terms of CMC(1) and nAUC (between brackets).

	No Action	Fixed Action				Action Selection	
	Low	Low + Face	Low + Torso	Low + Legs	Low + Low	Low + Max-Var	Low + Oracle
HSV	7.50 (72.25)	17.50 (75.69)	*17.50* *(77.63)*	15.00 (67.94)	10.00 (69.06)	**17.50** **(77.94)**	32.50 (86.50)
LBP	2.50 (65.56)	*7.50* *(70.31)*	5.00 (70.00)	5.00 (69.31)	5.00 (65.06)	**7.50** **(72.87)**	15.00 (82.19)
HSV+ LBP	17.50 (74.06)	25.00 (78.75)	*27.50* *(79.25)*	22.50 (72.63)	17.50 (71.00)	**27.50** **(80.12)**	45.00 (87.88)

Table 3. Comparing multishot performance at low resolution on the iLIDS-MA dataset with the proposed max-var method.

	Multi-shot Low Res.			Max-Var
	2 imgs	5 imgs	10 imgs	
HSV+ LBP	11.98 (70.99)	13.35 (71.57)	13.90 (71.09)	**27.50** **(80.12)**

Considering the iLIDS-MA dataset, analogue experiments have been carried out, whose results are reported on Table 2. In this case the max-var action selection algorithm reaches the best performance, demonstrating that such method works also when partial occlusions are present.

The last test on iLIDS-MA aims at comparing our PTZ re-id with a standard multi-shot strategy. Let us reasonably assume that the time needed by a standard PTZ camera to zoom at 4X is 1 second (as in the case of our SONY camera). Thus, the PTZ re-id method extracts an image at low resolution and one at high resolution in about 1 second, while a fixed camera can acquire reasonably between 7 to 10 frames per second at low resolution. We thus performed an experiment using up to 10 images at low resolution taken in an interval of about 1 second after the first probe image. Please note that in this case each probe image is compared with the gallery image, and a unique distance is obtained by extending Eq. 8 with uniform weights. Results are reported in Table 3, and shows that our framework is definitely superior.

4. CONCLUSIONS

This paper shows that a strong boost in the re-id literature can be given by the usage of a PTZ camera, overcoming multi-shot policies. Here we move the first step towards this direction, designing a protocol that selects the best action that a PTZ camera should perform, assuming it is able to zoom on single parts. In addition, we create a public benchmark that can be distributed to the community. Many are the desirable future works: first of all, the extension to deal with more actions performed sequentially; second the study of which features are better for a given action.

5. REFERENCES

[1] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130 – 144, 2013.

[2] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference (BMVC)*, 2011.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2011.

[4] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, 2011.

[5] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision (ECCV)*, 2008.

[6] W. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference (BMVC)*, 2009.

[7] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *European Conference on Computer Vision (ECCV)*, 2012.

[8] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *European Conference on Computer Vision (ECCV)*, 2012.

[9] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1, 2012.

[10] D. Figueira, L. Bazzani, H.Q. Minh, M. Cristani, A. Bernardino, and V. Murino, "Semi-supervised multi-feature learning for person re-identification," in *International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2013.

[11] S. Venugopalan and M. Savvides, "Unconstrained iris acquisition and recognition using cots ptz camera," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 38, 2010.

[12] H.C. Choi, U. Park, and A.K. Jain, "Ptz camera assisted face acquisition, tracking & recognition," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2010.

[13] K. Bernardin, F. Van De Camp, and R. Stiefelhagen, "Automatic person detection and tracking using fuzzy controlled active cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[14] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., pp. 507–514. MIT Press, Cambridge, MA, 2006.

[15] J. Denzler and C.M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 145–157, 2002.

[16] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Pattern Recognition, (ICPR). 12th IAPR International Conference on*. IEEE, 1994, vol. 1, pp. 582–585.

[17] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2360 –2367.

[18] P. Salvagnini, M. Cristani, A. Del Bue, and V. Murino, "An experimental framework for evaluating ptz tracking algorithms," in *8th International Conference on Computer Vision Systems (ICVS)*, Berlin, Heidelberg, 2011, ICVS'11, pp. 81–90, Springer-Verlag.

[19] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*. IEEE, 2011, pp. 1385–1392.

[20] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *First International Workshop on Re-Identification*, 2012.

[21] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image Vision Computing*, vol. 30, no. 6-7, pp. 443–452, June 2012.

[22] S. Bak, E. Corvee, F. Brémond, M. Thonnat, et al., "Person re-identification using spatial covariance regions of human body parts," in *International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2010.