Person Re-identification Using Partial Least Squares Appearance Modeling

Gabriel Lorencetti Prado¹, William Robson Schwartz², and Helio Pedrini¹

¹ Institute of Computing,
University of Campinas, Campinas, Brazil, 13083-852
`gabriel.prado@students.ic.unicamp.br`, `helio@ic.unicamp.br`

² Department of Computer Science,
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
`william@dcc.ufmg.br`

Abstract. Due to the large areas covered by surveillance systems, employed cameras usually lack intersection of field of view, refraining us from mapping the location of a person in a camera to another one. Therefore, when a subject appears in a camera, a person re-identification method is required to discover whether the subject has been previously identified in a different camera. Even though several approaches have been proposed in the literature, person re-identification is still a challenging problem due to appearance variation between cameras, changes in illumination, pose variation, and low quality data, among others. To reduce the effect of the aforementioned difficulties, we propose a person re-identification approach that models the appearance of the subjects based on multiple samples collected from multiple cameras and employs person detection and tracking to enhance the robustness of the method. Experiments conducted on three public available data sets demonstrate improvements over existing methods.

Keywords: person re-identification, partial least squares, appearance-based modeling, person detection, object tracking.

1 Introduction

Person re-identification consists on tracking multiple people in a camera network, registering their trajectories on each camera and assigning them consistent identifiers across the network. It is still an open problem in the Computer Vision area, as many challenges are faced when designing a robust re-identification system.

In the single camera case, issues include occlusion, pose variation and lighting conditions. Besides these difficulties, the major problem in a multi-camera system is to maintain a correspondence of the people tracked across cameras and people re-entering the camera field of view (FOV). When the camera calibration is not given or there is no overlapping field of views, the solution is even harder to be found, since homography-based methods such as [1] do not apply.

Several approaches have been proposed to address the re-identification problem and can be labeled in terms of different features. The classification under

single or multiple shot methods is the most common and refers to whether the method considers only one image to create a person representation or a group of images to perform it. In the following paragraphs, we present a brief review of works more related to the proposed approach. An extended description and discussion of person re-identification methods can be found in [2,3].

Some approaches based on a single image are [4,5,6]. In all these methods, the problem is treated as a classification task. Considering the multiple shot case, many distinct methods have been proposed. Interest point based approaches are presented in [7,8]. Some methods using person appearance modeling are given in [7,9,10]. Other strategies include the use of body regions [11], particle filtering [12], global descriptors [13], a master-slave approach [14], and finally a transfer approach to the scenario where only a subset of people is considered for [15]. Such methods present environmental constraints that limit their usage in a general condition. Unlike these works, our method introduces an appearance-based modeling that can be applied in a wider variety of real world scenarios.

This paper proposes a method to tackle the re-identification problem under multiple shot context built upon the work proposed in [5,6]. Similarly to these methods, appearance models based on Partial Least Squares (PLS) regressions [16] are used to grant the system a powerful discriminative characteristic. However, in our method, these models are associated to the information obtained from person tracking and detection to enhance the robustness of the original method. Moreover, we do not require that a person gallery is given, that is, classes in which people will be classified are not known in advance. Hence, our approach aims at the general scenario, in which cameras do not need to share FOV, camera calibration is ignored, and synchronized frames are not required.

The proposed technique is evaluated on three video data sets, which were also used for the ICPR 2012 Contest - *People tracking in wide baseline camera networks* [17]. The approach is compared with the method proposed in [5] to demonstrate the necessity of a tracking module to achieve better results for the person re-identification problem.

2 Proposed Method

Our approach to the re-identification problem is based on a full body appearance-based modeling via PLS regression [16]. The proposed method is composed of five stages, which are explained in details after a brief description of the Partial Least Squares regression technique. Figure 1 illustrates an overview of the method.

2.1 Partial Least Squares

In PLS regression, the input consists of a collection of classes in which the samples are classified and a collection of high dimensional feature vectors for each class. The method then significantly reduces the dimensionality of the feature vectors by creating variables (latent variables), which are obtained as linear combination of the original ones.

Fig. 1. Visual description of the proposed method. People are first detected by a pedestrian detector based on Partial Least Squares (PLS), as shown in the top images. People detected in continuous frames are then grouped into tracklets. These tracklets are put into partitions, where tracklets known to belong to different people are put in the same partition. A PLS model is then created for each tracklet in a one-against-all approach using the remaining tracklets from its partition as counter-examples. Afterwards, a score matrix is built by evaluating the matching between each pair of models. Finally, the score matrix is used to discover which models (therefore, tracklets) belong to the same person by removing the pairs with higher matching scores and assigning them to the same person.

The dimension reduction is performed in such a way that the covariance between the classes (subject's identifiers) and their feature vectors is maximized. The result is a collection of weight vectors that can be used to reduce the dimensionality of new feature vectors and classify them in the low dimensionality space using a regression-based approach [18].

2.2 Person Re-identification Based on Partial Least Squares

The goal of the first step is to obtain the detection windows for people contained in each frame of the video sequence. To do so, we first train a person detector based on a PLS appearance model that works with two classes (positive and negative person samples) [19]. The training is achieved by cropping the samples into overlapping blocks and extracting low-level features from each of them.

Detection windows are found by decomposing each frame into samples of increasing sizes and then extracting the same features from each of them. A non-maximum suppression is applied in the results to clean up redundant detection windows in multiple scales. A more detailed description can be found in [19].

The next step is to group detection windows from sequential frames of the same camera into tracklets. This is performed by tracking each detection window with a Kalman filter [20]. This approach also allows the tracking to fill in missing detections when they occur during a small number of frames.

Tracklets from different detections of the same frame naturally correspond to different people. This information is maintained throughout the method by keeping a set of tracklet partitions, where each partition contains tracklets that correspond to different people. A new partition is created whenever a tracklet is lost due to missing detections or a tracked person walks off the camera range.

For each tracklet partition, a new PLS appearance model is created for each tracklet in a one-against-all approach [5], using the remaining tracklets in the same partition as counter-examples. Then, all tracklets are pairwise matched against each other using the generated models. The matching results are used to build a matrix of matching scores between each tracklet pair.

The next step consists on successive removals of the maximum value from the matrix, marking the corresponding tracklets as belonging to the same person. Auxiliary maps are used in this stage and are described as follows.

For each tracklet, a similarity map (S) and a distinctivity map (D) are used. As soon as the tracklet is created, it is inserted into its own similarity map (Equation 1a). All other tracklets in the same partition (P) are inserted into its distinctivity map (Equation 1b). When a matching is found in the score matrix, both maps for the two involved tracklets are merged (Equations 2a) unless one tracklet is found in the other distinctivity map. This new similarity map is used to update each of the distinctivity maps for the tracklets found in the new distinctivity map (Equation 2b).

Initialization equations:

$$S_i \leftarrow i, \quad \forall i \in P_p \quad (1a) \quad D_i \leftarrow j, \quad \forall i, j \in P_p \mid i \neq j \quad (1b)$$

Update equations:

$$S_{\text{new}} \leftarrow S_i \cup S_j \quad D_{\text{new}} \leftarrow D_i \cup D_j \quad (2a)$$

$$S_k \leftarrow S_{\text{new}}, \quad \forall k \in S_{\text{new}} \quad D_k \leftarrow D_{\text{new}}, \quad \forall k \in S_{\text{new}} \\ D_k \leftarrow D_k \cup S_{\text{new}}, \quad \forall k \in D_{\text{new}} \quad (2b)$$

Finally, when no more matrix removals are possible, unique identifiers are assigned to each tracklet according to the information held in the similarity maps, in order to obtain the final tracklet-person matching.

3 Experimental Results

In this section, we present the evaluation of our method for the task of person re-identification using three data sets and its comparison to a baseline method [5]. Experiments were run offline, so the method is not ready for real time usage.

For the detection stage, a full body person detector was trained using cropped person images from the INRIA Person Dataset [21]. Only person images larger than 100 pixels were considered. The adopted features for both detection and modeling stages were histogram of oriented gradients [21] and gray-level co-occurrence matrices [22] to obtain edge and texture information, respectively.

Our re-identification approach was evaluated on three public video data sets, described as follows and summarized in Table 1. The first data set was CAT data set [23], in which people are mainly detected from only one of the cameras at a time. The second data set was sequence S7 of PETS2006 [24], where there is a larger number of frames with multiple person detections. The last data set considered was the Techgate data set. Unlike the previous data sets, cameras on this one are very close to the subjects.

Table 1. Summary of the data sets. Frames per camera denotes the number of frames available per camera; Camera sync indicates if the camera frames are synchronized; FOV Overlap indicates the amount of overlapping in the camera field of view.

	CAT	PETS2006	Techgate
Number of cameras	4	4	6
Number of people	4	56	4
Frames per second	15	25	15-30
Frames per camera	710	3401	3037-4415
Camera sync	Yes	Yes	No
FOV Overlap	Low	High	High

For the single camera case, selected metrics were Correct Detected Track (CDT), Track Detection Failure (TDF), False Alarm Track (FAT), and Track Fragmentation (TF) [25]. Best results are achieved when CDT is high and TDF, FAT and TF are low. Results for these metrics are shown in Tables 2-4.

For the multi-camera case, Crossing Fragments (X-Frag), Crossing ID Switches (X-IDS), Returning Fragments (R-Frag) and Returning ID Switches (R-IDS) [9] were selected. For all these metrics, lower values indicate better results. Results for the multi-camera metrics are shown in Table 5.

In Tables 2-5, metrics are displayed in percentages instead of absolute values and TF is shown as an average measure over all tracks in order to have a more general view of the results. Tracks shorter than 1 second were discarded, as suggested in the ICPR 2012 Contest [17] to avoid influence of short tracks.

The method described in [5] is used as baseline. We evaluate both methods with the same detections described above. Unlike ours, the baseline method requires a person gallery to classify the detected people. It is also used by the method to build person models in a one-against-all schema. Therefore, a gallery was created for each data set using person detections from the respective ground truth. For each person, 5 samples were randomly selected. To avoid influence from the training, results are an average of 10 executions with different galleries.

For the CAT data set, the proposed approach achieved accurate results for the single camera tracking, as it can be seen in Table 2. The CDT metric is quite high and the others present low values as expected for a good performance.

Table 2. Single camera metrics for each camera C_i of CAT data set

CAT	C_1	C_2	C_3	C_4	CAT	C_1	C_2	C_3	C_4
CDT	45.6%	27.2%	10.8%	26.7%	CDT	100%	94.4%	76.9%	83.3%
TDF	54.4%	72.8%	89.2%	73.3%	TDF	0.0%	5.6%	23.1%	16.7%
FAT	0.0%	0.0%	0.0%	0.5%	FAT	0.0%	0.6%	7.7%	6.3%
TF	0.3	0.2	0.0	0.0	TF	0.5	1.3	0.2	1.2

(a) Baseline Results

(b) Proposed Method Results

On the PETS2006 data set results, presented in Table 3, it can also be noted that the proposed method performance is higher than the baseline method. However, it was lower than the CAT data set due to the large number of people in the scene. In the second camera, performance is much lower since this camera is positioned at a quite higher place. Detections are therefore smaller and, consequently, fewer features can be extracted to perform the appearance modeling.

Table 3. Single camera metrics for each camera C_i of PETS2006 data set

PETS	C_1	C_2	C_3	C_4	PETS	C_1	C_2	C_3	C_4
CDT	5.1%	1.5%	3.0%	0.6%	CDT	89.8%	37.3%	83.0%	85.2%
TDF	94.9%	98.1%	97.0%	99.4%	TDF	10.2%	62.7%	17.0%	14.8%
FAT	0.0%	0.1%	0.0%	0.0%	FAT	2.2%	3.3%	1.0%	1.8%
TF	0.2	0.0	0.1	0.0	TF	0.6	0.5	1.0	0.4

(a) Baseline Results

(b) Proposed Method Results

In Techgate data set, an issue opposite to the found in PETS2006 data set occurs at some detections. The proximity of some people with the camera does not allow the full body person detector to find them, also reducing the system performance. Nevertheless, our method still produces better results.

Table 4. Single camera metrics for each camera C_i of Techgate data set

Tech	C_1	C_2	C_3	C_4	C_5	C_6
CDT	12.5%	13.8%	22.9%	9.0%	22.0%	26.4%
TDF	87.5%	86.2%	77.1%	91.0%	78.0%	73.6%
FAT	0.9%	0.4%	0.1%	0.0%	2.3%	0.1%
TF	0.1	0.3	0.1	0.2	0.2	0.5

(a) Baseline Results

Tech	C_1	C_2	C_3	C_4	C_5	C_6
CDT	75.0%	92.3%	73.5%	50.0%	60.0%	71.8%
TDF	25.0%	7.7%	26.5%	50.0%	40.0%	28.2%
FAT	5.5%	5.5%	1.3%	1.6%	14.8%	1.5%
TF	1.2	1.3	0.8	0.4	0.0	1.0

(b) Proposed Method Results

Finally, for the multi-camera metric results shown in Table 5, it can be seen that our method demonstrates superior results for the X-Frag and R-Frag metrics. However, it is not as good as the baseline with metrics X-IDS and R-IDS.

Table 5. Multi-camera metrics for each data set

	CAT	PETS	Tech		CAT	PETS	Tech
X-Frag	95.7%	100%	98.3%	X-Frag	58.0%	85.4%	73.2%
X-IDS	0.0%	0.0%	0.1%	X-IDS	15.2%	10.1%	10.9%
R-Frag	94.9%	99.8%	96.9%	R-Frag	56.4%	80.8%	71.3%
R-IDS	0.1%	0.0%	0.1%	R-IDS	24.7%	12.5%	12.5%

(a) Baseline Results

(b) Proposed Method Results

From the results presented above, it is clear that results for the proposed approach overcomes the baseline ones in all the considered data sets.

For the CDT metric, results are highly superior due to the addition of tracking as shown in Tables 2-4. This feature makes our method more robust against change of ids in the same tracklet, which causes CDT metric to achieve a higher performance. The same applies for the TDF metric.

Results are also low for the other single camera metrics, pointing out that not only our method is able to discover more correct tracks, but it also does not

add false positive tracks to the results. Actually, for both methods, this metric depends on false positive detections, which are shared by them both.

In the results for the multi-camera tracking metrics X-Frag and R-Frag, it can also be observed that our method achieves better results when compared to the baseline, as depicted in Table 5. For the R-IDS and X-IDS metrics, our method is also able to keep them low, however, higher than the baseline. This fact can be explained because baseline results are low due to the small amount of detected tracks, reflected in the CDT and TDF metrics. Furthermore, these metrics are also highly dependent on the detection results, which were clearly higher for the CAT data set when compared to the others.

4 Conclusions and Future Work

In this work, we presented a novel appearance-based modeling method for person re-identification across multiple cameras. Our system does not have any restriction on the camera network configuration, such as calibration, FOV intersection or frame synchronization. Therefore, it is sufficiently general to be used in many data sets with different characteristics, as demonstrated in the experiments. Moreover, it can be easily modified to use other low level features.

The feasibility of the proposed approach was demonstrated by its execution on three different data sets and the obtained results are evaluated in terms of current state-of-art metrics for both single and multiple camera tracking. Experimental results demonstrated that it outperforms the baseline method regarding most of considered metrics for both single and multiple camera cases.

As future work, we plan to investigate how to update PLS models online, as new person associations are discovered during the matching stage.

Acknowledgements. The authors are grateful to FAPESP, FAPEMIG, CNPq, and CAPES for the financial support and to the organizers of the ICPR 2012 Contest - *People tracking in wide baseline camera networks* [17] for providing the data sets.

References

1. Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., Maybank, S.: Principal Axis-Based Correspondence between Multiple Cameras for People Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(4), 663–671 (2006)
2. Mazzon, R., Tahir, S.F., Cavallaro, A.: Person re-identification in crowd. *Pattern Recognition Letters* 33(14), 1828–1837 (2012)
3. Satta, R.: Appearance Descriptors for Person Re-identification: a Comprehensive Review. *CoRR* abs/1307.5748 (2013)
4. Hirzer, M., Beleznai, C., Köstinger, M., Roth, P.M., Bischof, H.: Dense Appearance Modeling and Efficient Learning of Camera Transitions for Person Re-Identification. In: *International Conference on Image Processing* (2012)

5. Schwartz, W.R., Davis, L.S.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Brazilian Symposium on Computer Graphics and Image Processing (2009)
6. Schwartz, W.R.: Scalable People Re-Identification Based on a One-Against-Some Classification Scheme. In: International Conference on Image Processing (2012)
7. Gheissari, N., Sebastian, T.B., Hartley, R.: Person Reidentification Using Spatiotemporal Appearance. In: Computer Vision and Pattern Recognition (2006)
8. Hamdoun, O., Moutarde, F., Stanculescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: International Conference on Distributed Smart Cameras (2008)
9. Kuo, C.-H., Huang, C., Nevatia, R.: Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 383–396. Springer, Heidelberg (2010)
10. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-Shot Person Re-identification by HPE Signature. In: International Conference on Pattern Recognition (2010)
11. Bazzani, L., Cristani, M., Murino, V.: Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identification. *Computer Vision and Image Understanding* 117(2), 130–144 (2013)
12. Li, M., Chen, W., Huang, K., Tan, T.: Visual Tracking via Incremental Self-tuning Particle Filtering on the Affine Group. In: Computer Vision and Pattern Recognition (2010)
13. Cai, Y., Pietikäinen, M.: Person Re-identification Based on Global Color Context. In: Koch, R., Huang, F. (eds.) ACCV 2010 Workshops, Part I. LNCS, vol. 6468, pp. 205–215. Springer, Heidelberg (2011)
14. Alahi, A., Vanderghelynst, P., Bierlaire, M., Kunt, M.: Cascade of Descriptors to Detect and Track Objects Across Any Network of Cameras. *Computer Vision and Image Understanding* 114(6), 624–640 (2010)
15. Zheng, W.S., Gong, S., Xiang, T.: Transfer Re-identification: From Person to Set-based Verification. In: Computer Vision and Pattern Recognition (2012)
16. Wold, H.: Partial Least Squares. *Encyclopedia of Statistical Sciences* 6, 581–591 (1985)
17. International Conference on Pattern Recognition Contest - People tracking in wide baseline camera networks (2012),
<http://www.wide-baseline-camera-network-contest.org>
18. Schwartz, W.R., Guo, H., Choi, J., Davis, L.S.: Face Identification Using Large Feature Sets. *IEEE Transactions on Image Processing* 21(4), 2245–2255 (2012)
19. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human Detection Using Partial Least Squares Analysis. In: International Conference on Computer Vision (2009)
20. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. Technical report, Department of Computer Science, University of North Carolina, USA (1995)
21. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Computer Vision and Pattern Recognition (2005)
22. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics SMC* 3, 610–621 (1973)

23. CAT Project, <http://www.cat-project.at>
24. Thirde, D., Li, L., Ferryman, J.: Overview of the PETS2006 Challenge. In: International Workshop on Performance Evaluation of Tracking and Surveillance (2006)
25. Yin, F., Makris, D., Velastin, S.A.: Performance Evaluation of Object Tracking Algorithms. In: International Workshop on Performance Evaluation of Tracking and Surveillance (2007)