Temporal Model Adaptation for Person Re-Identification: Supplementary Material

Niki Martinel[1,3], Abir Das[2],
Christian Micheloni[1], and Amit K. Roy-Chowdhury[3]

[1] University of Udine, 33100 Udine, Italy
[2] University of Massatchussets Lowell, 01852 Lowell, MA, USA
[3] University of California Riverside, 92507 Riverside, CA, USA

Abstract. This supplementary material accompanies the paper entitled "Temporal Model Adaptation for Person Re-Identification", accepted for publication in ECCV 2016. It introduces an additional analysis of our approach and experiments on the 3DPeS and the CUHK03 datasets which, due to page limit constraints, could have not been included in the main paper. It also provides a complete derivation of both the standard deterministic as well as the proposed stochastic ADMM optimization solutions.

1 Additional Experimental Results

In the following, experimental results on two additional datasets as well as an analysis of the rank 1 re-identification performance achieved by using a fixed percentage of manually labeled pairs in each batch are presented.

1.1 Additional Datasets

3DPeS dataset [1] contains 1,012 images of 191 persons taken from a multi-camera distributed surveillance system. Each one of the 8 cameras has footages acquired with different light conditions. Persons images have been captured from different viewpoints at different time instants. This results in a very challenging dataset (see Fig. 1 for a few samples). Following the same protocol proposed in [2], the dataset has been randomly split into two sets (one for training and one for testing) containing 95 and 96 persons each, respectively. We report on the average results obtained from 10 independent random trials.
CUHK03 [3] is one of the largest and most challenging datasets. It contains 13,164 images of 1,360 pedestrians acquired by six disjoint cameras. Each person has been observed by two disjoint camera views and has an average of about 5 images in each view. This dataset also contains samples detected with a state-of-the-art person detector, hence it provides a more realistic setting (see Fig. 2). To run the experiments, we followed the same procedure as in [3] and used the 20 provided trials. Each of these splits the data into a training set and a test set containing 1,160 and 100 persons, respectively.

Fig. 1: 15 Image samples from 3DPeS dataset. Each column corresponds to a pair of images of the same person captured by the two different cameras.

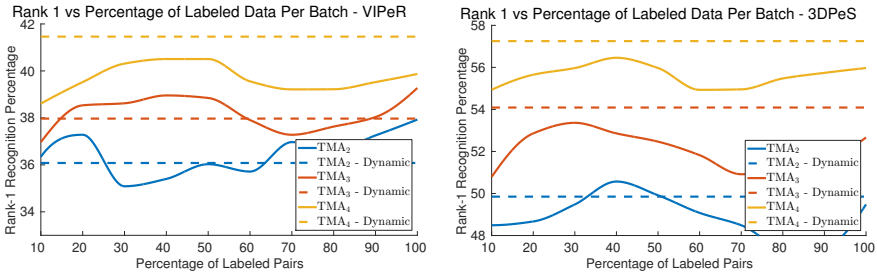
Fig. 2: 15 Image samples from CUHK03 dataset. Each column corresponds to a pair of images of the same person captured by the two different cameras.

1.2 Influence of the Temporal Model Adaptation Components

In the main paper we have conducted an analysis of the single components of the proposed approach by separately considering the similarity-dissimilarity metric learning and the probe relevant set selection criterion. To have additional insights on the achieved performance and to verify if the proposed dynamic probe relevant set selection criterion yields better performance and less manual labor than using a fixed percentage of pairs to label in each batch, we have computed the results in Fig. 3. The plots show the rank 1 recognition percentage achieved as a function of the percentage of labeled pairs in each batch (solid lines) and compare the results to ones obtained using the proposed graph-based dynamic solution (dashed lines). Precisely, to compute the results shown through the solid lines, hence to select the pairs to be labeled, for each person p and percentage s, we first sorted the elements of $\hat{\mathbf{h}}$ using a descending order. Then, we selected the first $s|\mathcal{D}_p|$ gallery persons in $\mathcal{D}_p$.

Results are shown for the VIPeR (Fig.3(a)) and 3DPeS (Fig.3(b)) datasets considering the 3 incremental updates. Both the plots show a similar trend: using few labeled pairs better performance than using completely labeled data is obtained. With increasing percentages of labeled pairs, rank 1 recognition rates tend to decrease –up to a certain point– then, they go up again.

Specifically, results in Fig.3(a) show that for the first two batch updates (i.e., TMA_2 and TMA_3), using a fixed percentage achieves better rank 1 performance than the proposed graph based solution (i.e., TMA_2-Dynamic and TMA_3-Dynamic). However, with the third batch incremental step, our TMA_4-Dynamic solution outperforms the other one used for comparison regardless of the percentage of manually labeled pairs. We speculate that such a behavior



(a) **(b)**

Fig. 3: Rank 1 correct recognition rate performance obtained using the 3 incremental batch updates is shown as a function of the percentage of labeled pairs in each batch. In (a) single-shot results on the VIPeR dataset. In (b) multiple-shot results on the 3DPeS dataset.

is due to the fact that, using a single shot strategy and a fixed set of gallery persons to be labeled for each probe person, we have that either many informative gallery samples are excluded or lots of uninformative and negative ones are included in the probe relevant set. This may cause the optimization parameters to be updated with too few informative or too many uninformative samples. As a result, the incremental steps are no longer significant in the long term.

As shown in Fig.3(b), similar results are obtained using the multiple-shot strategy. However, in such a case, using all the labeled data is not achieving anymore better rank 1 performance than using only few labeled pairs. In addition, our approach performs better than other solutions already after considering the 3^{rd} batch. This result may indicate that if the different correct match pairs are already included in the first ranks, i.e., considered for labeling even with a small percentage of labeled pairs, adding additional negative samples does not yield to increasing performance.

1.3 State-of-the-art Comparisons

3DPeS: Table 1 shows the performance comparison between our approach and existing ones on the 3DPeS dataset. Results reflect the ones provided in the main paper with our method performing better than recent ones using only about 5% of the data (i.e., TMA_1). Incremental updates yields performance improvement and after the 4^{th} batch is processed, the proposed solution reaches the 4^{th} highest rank 1 exploiting only 22.31% of labeled data. It is worth noticing that the three works that have better performance on such dataset, namely CSL [4], kLFDA [5] and KEPLER [6] exploits either a different dataset setup, learn a metric on a kernel space or use a multiple shot-strategy. More precisely, in CSL [4], authors create 3 groups of cameras having similar views, then train their approach considering pairs of camera groups and multiple-shot of the same person –this last is also exploited in KEPLER [6]. This, simplifies the re-identification performance since intra-camera variations are significantly reduced. To provide a more fair comparison with such approaches we also applied a multiple-shot strategy TMA_4^{MS}. As expected, results show significant improvements and our approach performs better than KEPLER and almost reaches the performances of CSL. In kLFDA [5], a kernel space is computed before training the LFDA [2] method.

Table 1: Comparison with state-of-the-art methods on the 3DPeS dataset. Best results for each rank are in boldface font.

Rank →	1	5	10	20	Labeled [%]	Reference
CSL	**57.9**	**81.1**	**89.5**	93.7	100	ICCV 2015 [4]
TMA_4^{MS}	57.25	76.75	86.45	**95.83**	19.06	Proposed
kLFDA	54.0	77.7	85.9	92.4	100	ECCV 2014 [5]
KEPLER	51.37	76.18	84.32	92.13	100	TIP 2015 [6]
TMA_0	47.91	76.03	85.54	92.51	100	Proposed
TMA_4	47.52	75.67	84.96	92.14	22.31	Proposed
rPCCA	47.3	75.0	84.5	91.9	100	ECCV 2014 [5]
TMA_3	44.75	69.76	82.50	93.75	16.18	Proposed
TMA_2	41.67	65.62	80.21	85.42	10.48	Proposed
IMS-LFDA	38.71	61.64	72.22	82.54	100	CVPR 2015 [7]
TMA_1	38.02	53.54	62.08	75.62	5.25	Proposed
LFDA	34.24	59.31	70.13	81.12	100	CVPR 2013 [2]
KISSME	32.76	57.65	68.51	79.71	100	CVPR 2012 [8]
PRDC	34.04	57.74	64.37	73.84	100	TPAMI 2013 [9]
eSDC	27.25	56.16	66.82	74.88	100	CVPR 2013 [10]

Table 2: Rank 1 performance comparison with state-of-the-art methods on the CUHK03 dataset. Best results are in boldface font.

	Labeled	Detected	Labeled [%]	Reference
MLAPG	**57.96**	**51.15**	100	ICCV 2015 [11]
TMA_4	54.78	48.13	19.09/21.41	Proposed
IDLA	54.74	44.96	100	CVPR 2015 [12]
XQDA	52.20	46.25	100	CVPR 2015 [13]
TMA_3	50.64	44.23	13.59/17.12	Proposed
TMA_2	44.12	37.74	10.43/11.21	Proposed
TMA_1	35.98	32.51	6.11/6.11	Proposed
DRSCH-128	21.96	–	100	TIP 2015 [14]
DeepReID	20.65	19.89	100	CVPR 2014 [3]
DRSCH-64	18.74	–	100	TIP 2015 [14]
KISSME	14.17	11.70	100	CVPR 2012 [8]

This allows to capture the non-linearities in the feature space. Indeed, as shown in Table 1, this improves LFDA [2] rank 1 performance by about 20%. Thus, we hypothesize that our approach can benefit of similar improvements if a kernel space is adopted before learning the similarity-dissimilarity metric.

CUHK03: To validate our method in a more realistic scenario we have computed the results on the CUHK03 dataset. Results in Table 2 are consistent with the ones obtained for the VIPeR and PRID450S datasets. Our approach has significantly better performance than the recent deep architectures even by using 6% of labeled data. Incremental updates bring in relevant improvements and we achieve the second best rank 1 recognition rate using both the labeled (54.78%) and detected (48.13%) person images with only 19.09% and 21.41% of manually labeled samples, respectively. This demonstrates that our approach can scale to a real scenario and achieve competitive performance with significantly less manual labor.

2 Alternating Direction Method of Multipliers Derivation for Our Approach

Following the explanation in the main paper, we propose to use the ADMM optimization method to solve the similarity-dissimilarity learning problem –in

eq.(5) of the main paper– by means of the corresponding augmented Lagrangian

$$L_{\mathbf{K},\mathbf{P},\mathbf{U},\mathbf{V},\mathbf{\Lambda},\mathbf{\Psi}} = \mathcal{J}_{\mathbf{K},\mathbf{P}} + \Omega_{\mathbf{U},\mathbf{V}} + \langle \mathbf{\Lambda}, \mathbf{K}-\mathbf{U}\rangle + \langle \mathbf{\Psi}, \mathbf{P}-\mathbf{V}\rangle + \frac{\rho}{2}\left(\|\mathbf{K}-\mathbf{U}\|_F^2 + \|\mathbf{P}-\mathbf{V}\|_F^2\right) \tag{1}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{r\times d}$ and $\mathbf{\Psi} \in \mathbb{R}^{r\times d}$ are two Lagrangian multipliers, $\langle \cdot,\cdot \rangle$ denote the inner product, $\|\cdot\|_F$ is the Frobenius norm and, $\rho > 0$ is a penalty parameter.

To solve the problem in eq.(1) at each each epoch s, ADMM updates the variables in an alternating fashion as

$$\mathbf{K}^{(s+1)} = \arg\min_{\mathbf{K}} L_{\mathbf{K},\mathbf{P}^{(s)},\mathbf{U}^{(s)},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \tag{2}$$

$$\mathbf{P}^{(s+1)} = \arg\min_{\mathbf{P}} L_{\mathbf{K}^{(s+1)},\mathbf{P},\mathbf{U}^{(s)},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \tag{3}$$

$$\mathbf{U}^{(s+1)} = \arg\min_{\mathbf{U}} L_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)},\mathbf{U},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \tag{4}$$

$$\mathbf{V}^{(s+1)} = \arg\min_{\mathbf{V}} L_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)},\mathbf{U}^{(s+1)},\mathbf{V},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \tag{5}$$

$$\mathbf{\Lambda}^{(s+1)} = \mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}^{(s+1)} - \mathbf{U}^{(s+1)}\right) \tag{6}$$

$$\mathbf{\Psi}^{(s+1)} = \mathbf{\Psi}^{(s)} + \rho\left(\mathbf{P}^{(s+1)} - \mathbf{V}^{(s+1)}\right) \tag{7}$$

In the following we present the specific update rules for the standard *deterministic solution* and our stochastic version. Since the updates for eq.(4-7) do not depend on the number of training samples, those are identical for both approaches, thus we provide their derivation only once.

2.1 Deterministic ADMM

Update K: To update the similarity low-rank projection matrix we have to find the derivative of our objective with respect to such a parameter and solve for it at a stationary point. So, we can write our problem as

$$\begin{aligned}\mathbf{K}^{(s+1)} &= \arg\min_{\mathbf{K}} L_{\mathbf{K},\mathbf{P}^{(s)},\mathbf{U}^{(s)},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \\ &\equiv \frac{\partial}{\partial \mathbf{K}}\Bigg[\mathcal{J}_{\mathbf{K},\mathbf{P}^{(s)}} + \Omega_{\mathbf{U}^{(s)},\mathbf{V}^{(s)}} + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}-\mathbf{U}^{(s)}\rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}^{(s)}-\mathbf{V}^{(s)}\rangle \\ &\qquad + \frac{\rho}{2}\left(\left\|\mathbf{K}-\mathbf{U}^{(s)}\right\|_F^2 + \left\|\mathbf{P}^{(s)}-\mathbf{V}^{(s)}\right\|_F^2\right)\Bigg] = 0\end{aligned} \tag{8}$$

Directly removing the terms that do not depend on $\mathbf{K}$ we have to solve

$$\frac{\partial}{\partial \mathbf{K}}\left[\mathcal{J}_{\mathbf{K},\mathbf{P}^{(s)}} + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}-\mathbf{U}^{(s)}\rangle + \frac{\rho}{2}\left(\left\|\mathbf{K}-\mathbf{U}^{(s)}\right\|_F^2\right)\right] = 0 \tag{9}$$

Now, let first compute the derivatives for each of the terms separately.

– **Loss function:**

$$\frac{\partial}{\partial \mathbf{K}} \mathcal{J}_{\mathbf{K},\mathbf{P}^{(s)}} \quad \text{By substitution} \tag{10}$$

$$\frac{\partial}{\partial \mathbf{K}} \frac{1}{n} \sum_{i}^{n} \max\left(0, 1 - S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})\right) \quad \text{By the sum rule} \tag{11}$$

$$\frac{1}{n} \sum_{i}^{n} \frac{\partial}{\partial \mathbf{K}} \max\left(0, 1 - S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})\right) \tag{12}$$

Now, by computing the sub-gradient of the max function we obtain

$$\frac{1}{n} \sum_{i}^{n} \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)}) \geq 1 \\ \frac{\partial}{\partial \mathbf{K}} \left(1 - S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})\right) & \text{otherwise} \end{cases} \tag{13}$$

Substituting $S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})$ with its definition given in the main paper eq.(3), we have that the non-trivial case in eq.(13) can be written as

$$\frac{1}{n} \sum_{i}^{n} \frac{\partial}{\partial \mathbf{K}} \left[1 - y_{p^{(i)},g^{(i)}} \left(\sigma_{\mathbf{K}}(\mathbf{x}_{p^{(i)}}, \mathbf{x}_{g^{(i)}}) - \frac{1}{2}\delta_{\mathbf{P}^{(s)}}(\mathbf{x}_{p^{(i)}}, \mathbf{x}_{g^{(i)}})\right)\right] \tag{14}$$

which, by removing the terms that do not depend on $\mathbf{K}$ and substituting with the definition of $\sigma_{\mathbf{K}}$ –in eq.(1) of the main paper– corresponds to

$$\frac{1}{n} \sum_{i}^{n} \frac{\partial}{\partial \mathbf{K}} \left(-y_{p^{(i)},g^{(i)}} \mathbf{x}_{p^{(i)}}^{T} \mathbf{K}^{T} \mathbf{K} \mathbf{x}_{g^{(i)}}\right) \tag{15}$$

Solving the matrix derivative using the identity (82) in [15] we obtain

$$\frac{1}{n} \sum_{i}^{n} -y_{p^{(i)},g^{(i)}} \mathbf{K}(\mathbf{x}_{p^{(i)}} \mathbf{x}_{g^{(i)}}^{T} + \mathbf{x}_{g^{(i)}} \mathbf{x}_{p^{(i)}}^{T}) \tag{16}$$

So, we have that

$$\frac{\partial}{\partial \mathbf{K}} \mathcal{J}_{\mathbf{K},\mathbf{P}^{(s)}} = \frac{1}{n} \sum_{i}^{n} \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)}) \geq 1 \\ -y_{p^{(i)},g^{(i)}} \mathbf{K}(\mathbf{x}_{p^{(i)}} \mathbf{x}_{g^{(i)}}^{T} + \mathbf{x}_{g^{(i)}} \mathbf{x}_{p^{(i)}}^{T}) & \text{otherwise} \end{cases} \tag{17}$$

– **Lagrange Components:**

$$\frac{\partial}{\partial \mathbf{K}} \left[\langle \mathbf{\Lambda}^{(s)}, \mathbf{K} - \mathbf{U}^{(s)} \rangle + \frac{\rho}{2} \left(\left\| \mathbf{K} - \mathbf{U}^{(s)} \right\|_{F}^{2}\right)\right] \quad \text{By definition} \tag{18}$$

$$\frac{\partial}{\partial \mathbf{K}} \left[\mathbf{Tr}\left((\mathbf{K} - \mathbf{U}^{(s)})^{T} \mathbf{\Lambda}^{(s)}\right) + \frac{\rho}{2} \left(\left\| \mathbf{K} - \mathbf{U}^{(s)} \right\|_{F}^{2}\right)\right] \quad \text{By trace property} \tag{19}$$

$$\frac{\partial}{\partial \mathbf{K}} \left[\mathbf{Tr}\left(\mathbf{K}^{T} \mathbf{\Lambda}^{(s)}\right) - \mathbf{Tr}\left((\mathbf{U}^{(s)})^{T} \mathbf{\Lambda}^{(s)}\right) + \frac{\rho}{2} \left(\left\| \mathbf{K} - \mathbf{U}^{(s)} \right\|_{F}^{2}\right)\right] \tag{20}$$

Discarding the terms that do not depend on $\mathbf{K}$, then solving the derivative for the trace and the Frobenius norm we obtain

$$\frac{\partial}{\partial \mathbf{K}}\left[\langle \mathbf{\Lambda}^{(s)}, \mathbf{K}-\mathbf{U}^{(s)}\rangle + \frac{\rho}{2}\left(\left\|\mathbf{K}-\mathbf{U}^{(s)}\right\|_F^2\right)\right] \tag{21}$$
$$= \mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}-\mathbf{U}^{(s)}\right)$$

Now, putting all together we have to solve

$$\mathbf{K}^{(s+1)} \equiv \mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}-\mathbf{U}^{(s)}\right) + \tag{22}$$
$$\frac{1}{n}\sum_i^n \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)}) \geq 1 \\ -y_{p^{(i)},g^{(i)}}\mathbf{K}(\mathbf{x}_{p^{(i)}}\mathbf{x}_{g^{(i)}}^T + \mathbf{x}_{g^{(i)}}\mathbf{x}_{p^{(i)}}^T) & \text{otherwise} \end{cases} = 0$$

Using linearity properties we can write

$$\mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}-\mathbf{U}^{(s)}\right) + \frac{\mathbf{K}}{n}\sum_i^n \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)}) \geq 1 \\ -y_{p^{(i)},g^{(i)}}(\mathbf{x}_{p^{(i)}}\mathbf{x}_{g^{(i)}}^T + \mathbf{x}_{g^{(i)}}\mathbf{x}_{p^{(i)}}^T) & \text{otherwise} \end{cases} \tag{23}$$

Then, introducing an additional variable

$$\mathbf{W} = \sum_i^n \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)}) \geq 1 \\ -y_{p^{(i)},g^{(i)}}(\mathbf{x}_{p^{(i)}}\mathbf{x}_{g^{(i)}}^T + \mathbf{x}_{g^{(i)}}\mathbf{x}_{p^{(i)}}^T) & \text{otherwise} \end{cases} \tag{24}$$

we have that

$$\mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}-\mathbf{U}^{(s)}\right) + \mathbf{K}\frac{\mathbf{W}}{n} = 0 \tag{25}$$

Solving for $\mathbf{K}$ yields

$$\mathbf{K}^{(s+1)} = \left(\rho\mathbf{U}^{(s)} - \mathbf{\Lambda}^{(s)}\right)\left(\rho + \frac{\mathbf{W}}{n}\right)^{-1} \tag{26}$$

Update P: To update the dissimilarity low-rank projection matrix, let us proceed as above and write our problem as

$$\mathbf{P}^{(s+1)} = \arg\min_{\mathbf{P}} L_{\mathbf{K}^{(s+1)},\mathbf{P},\mathbf{U}^{(s)},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \tag{27}$$
$$\equiv \frac{\partial}{\partial \mathbf{P}}\Bigg[\mathcal{J}_{\mathbf{K}^{(s+1)},\mathbf{P}} + \Omega_{\mathbf{U}^{(s)},\mathbf{V}^{(s)}} + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}^{(s+1)} - \mathbf{U}^{(s)}\rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}-\mathbf{V}^{(s)}\rangle$$
$$+ \frac{\rho}{2}\left(\left\|\mathbf{K}^{(s+1)} - \mathbf{U}^{(s)}\right\|_F^2 + \left\|\mathbf{P}-\mathbf{V}^{(s)}\right\|_F^2\right)\Bigg] = 0$$

Directly removing the terms that do not depend on $\mathbf{K}$ we have to solve

$$\frac{\partial}{\partial \mathbf{P}}\left[\mathcal{J}_{\mathbf{K}^{(s+1)},\mathbf{P}} + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}-\mathbf{V}^{(s)}\rangle + \frac{\rho}{2}\left(\left\|\mathbf{P}-\mathbf{V}^{(s)}\right\|_F^2\right)\right] = 0 \tag{28}$$

Now, let us again compute the derivatives for each of the terms separately.

– **Loss function:** Since the first steps of the derivation are already given in eq.(10-14), hereby we only proceed with the derivation of the relevant terms for $\mathbf{P}$ only. Thus, by starting with eq.(14), removing terms that do not depend on $\mathbf{P}$ and substituting with the definition of $\delta_{\mathbf{P}}$ –eq.(2) in the main paper– we can write

$$\frac{1}{n}\sum_{i}^{n}\frac{\partial}{\partial \mathbf{P}}\left[y_{p^{(i)},g^{(i)}}\frac{1}{2}(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}\mathbf{P}^{T}\mathbf{P}(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})\right] \tag{29}$$

Recognizing that the problem has the same form as the one in eq.(15), we can obtain the following solution using the same identity (82) in [15]

$$\frac{1}{n}\sum_{i}^{n}y_{p^{(i)},g^{(i)}}\frac{1}{2}\mathbf{P}\Big((\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}+(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}\Big) \tag{30}$$

$$=\frac{1}{n}\sum_{i}^{n}y_{p^{(i)},g^{(i)}}\mathbf{P}\Big((\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}\Big) \tag{31}$$

So, we have that

$$\frac{\partial}{\partial \mathbf{P}}\mathcal{J}_{\mathbf{K}^{(s+1)},\mathbf{P}}=\frac{1}{n}\sum_{i}^{n}\begin{cases}0 & \text{if} \quad S_{\mathbf{K}^{(s+1)},\mathbf{P}}(p^{(i)},g^{(i)})\geq 1\\ y_{p^{(i)},g^{(i)}}\mathbf{P}\Big((\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}\Big) & \text{otherwise}\end{cases} \tag{32}$$

– **Lagrange Components:** The solution to the Lagrange components is very similar to the one above but with different variables. In particular, the solution requires that the derivation is performed using the updated similarity matrix $\mathbf{K}^{(s+1)}$. Other than that, all the steps remain the same so are not given here.

After solving the Lagrangian components for $\mathbf{P}$ and putting all together we have

$$\mathbf{P}^{(s+1)}\equiv\mathbf{\Psi}^{(s)}+\rho\left(\mathbf{P}-\mathbf{V}^{(s)}\right) \tag{33}$$

$$\frac{1}{n}\sum_{i}^{n}\begin{cases}0 & \text{if} \quad S_{\mathbf{K}^{(s+1)},\mathbf{P}}(p^{(i)},g^{(i)})\geq 1\\ y_{p^{(i)},g^{(i)}}\mathbf{P}\Big((\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T}\Big) & \text{otherwise}\end{cases}=0$$

Then, using linearity properties and introducing an additional variable

$$\mathbf{Z}=\sum_{i}^{n}\begin{cases}0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}^{(s)}}(p^{(i)},g^{(i)})\geq 1\\ -y_{p^{(i)},g^{(i)}}(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}}-\mathbf{x}_{g^{(i)}})^{T} & \text{otherwise}\end{cases} \tag{34}$$

we have that

$$\mathbf{\Psi}^{(s)}+\rho\left(\mathbf{P}-\mathbf{V}^{(s)}\right)+\mathbf{P}\frac{\mathbf{Z}}{n}=0 \tag{35}$$

which, by solving for $\mathbf{P}$, results in

$$\mathbf{P}^{(s+1)}=\left(\rho\mathbf{V}^{(s)}-\mathbf{\Psi}^{(s)}\right)\left(\rho+\frac{\mathbf{Z}}{n}\right)^{-1} \tag{36}$$

Remark 1. Using the standard deterministic ADMM solution we have that the update rules for $\mathbf{K}$ and $\mathbf{P}$ are very similar to each other. They require to compute an inverse of a matrix obtained by performing the computation of specific outer products between the considered feature vectors. Such an operation is clearly expensive in terms of computational resources, especially if we consider 26,960-D feature vectors as in our scenario.

Update U:

$$\begin{aligned}\mathbf{U}^{(s+1)} &= \arg\min_{\mathbf{U}} L_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)},\mathbf{U},\mathbf{V}^{(s)},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \\ &\equiv \frac{\partial}{\partial \mathbf{U}}\Bigg[\mathcal{J}_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)}} + \Omega_{\mathbf{U},\mathbf{V}^{(s)}} \\ &\quad + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}^{(s+1)} - \mathbf{U}\rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}^{(s+1)} - \mathbf{V}^{(s)}\rangle \\ &\quad + \frac{\rho}{2}\left(\left\|\mathbf{K}^{(s+1)} - \mathbf{U}\right\|_F^2 + \left\|\mathbf{P}^{(s+1)} - \mathbf{V}^{(s)}\right\|_F^2\right)\Bigg] = 0\end{aligned} \tag{37}$$

Removing terms that do not involve $\mathbf{U}$ we have to solve

$$\frac{\partial}{\partial \mathbf{U}}\left[\alpha \left\|\mathbf{U}\right\|_{2,1} + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}^{(s+1)} - \mathbf{U}\rangle + \frac{\rho}{2}\left(\left\|\mathbf{K}^{(s+1)} - \mathbf{U}\right\|_F^2\right)\right] = 0 \tag{38}$$

Following [16] (Section 3.1.1), by combining the linear and quadratic terms in the augmented Lagrangian and scaling the dual variable, we can write

$$\frac{\partial}{\partial \mathbf{U}}\left[\Omega_{\mathbf{U},\mathbf{V}^{(s)}} + \frac{\rho}{2}\left(\left\|\mathbf{K}^{(s+1)} - \mathbf{U} + \mathbf{\Lambda}^{(s)}/\rho\right\|_F^2\right)\right] = 0 \tag{39}$$

which, by deriving a similar explanation to the one for the "group-LASSO" given in [17] can be solved using the soft-thresholding operator, hence yielding to

$$\mathbf{U}^{(s+1)} = \left(\mathbf{K}_{i,:}^{(s+1)} + \mathbf{\Lambda}_{i,:}^{(s)}/\rho\right)\max\left(0, 1 - \frac{\alpha}{\rho\left\|\mathbf{K}_{i,:}^{(s+1)} + \mathbf{\Lambda}_{i,:}^{(s)}/\rho\right\|_2}\right) \tag{40}$$

where $i = 1, \cdots, r$ denotes the i-th row of a parameter matrix.

Update V:

$$\begin{aligned}\mathbf{V}^{(s+1)} &= \arg\min_{\mathbf{V}} L_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)},\mathbf{U}^{(s+1)},\mathbf{V},\mathbf{\Lambda}^{(s)},\mathbf{\Psi}^{(s)}} \\ &\equiv \frac{\partial}{\partial \mathbf{V}}\Bigg[\mathcal{J}_{\mathbf{K}^{(s+1)},\mathbf{P}^{(s+1)}} + \Omega_{\mathbf{U}^{(s+1)},\mathbf{V}} \\ &\quad + \langle \mathbf{\Lambda}^{(s)}, \mathbf{K}^{(s+1)} - \mathbf{U}^{(s+1)}\rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}^{(s+1)} - \mathbf{V}\rangle \\ &\quad + \frac{\rho}{2}\left(\left\|\mathbf{K}^{(s+1)} - \mathbf{U}^{(s+1)}\right\|_F^2 + \left\|\mathbf{P}^{(s+1)} - \mathbf{V}\right\|_F^2\right)\Bigg] = 0\end{aligned} \tag{41}$$

Since the steps for solving the above problem are very similar to the ones used to solve eq.(37), we omit them and provide only the final solution. This corresponds

to

$$\mathbf{V}^{(s+1)} = \left(\mathbf{P}_{i,:}^{(s+1)} + \mathbf{\Psi}_{i,:}^{(s)}/\rho\right) \max\left(0, 1 - \frac{\beta}{\rho\left\|\mathbf{P}_{i,:}^{(s+1)} + \mathbf{\Psi}_{i,:}^{(s)}/\rho\right\|_2}\right) \tag{42}$$

Update $\mathbf{\Lambda}$ and $\mathbf{\Psi}$: The updates for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are straightforward and can be easily obtained by defining the dual of our augmented Lagrangian objective. See [16] for more details. In our case these correspond to

$$\mathbf{\Lambda}^{(s+1)} = \mathbf{\Lambda}^{(s)} + \rho\left(\mathbf{K}^{(s+1)} - \mathbf{U}^{(s+1)}\right) \tag{43}$$

and to

$$\mathbf{\Psi}^{(s+1)} = \mathbf{\Psi}^{(s)} + \rho\left(\mathbf{P}^{(s+1)} - \mathbf{V}^{(s+1)}\right) \tag{44}$$

respectively.

Remark 2. As stated above, the preceding updates (corresponding to eq.(4-7))do not depend on the number of training samples, hence are identical for both the deterministic ADMM as well as for the proposed stochastic ADMM. Thus, the derivation is not given for the proposed stochastic ADMM.

2.2 Stochastic ADMM

Following [18], to obtain the stochastic updates for $\mathbf{K}$ and $\mathbf{P}$ using SCAS-ADMM we have to compute the gradients of the augmented Lagrangian with respect to different parameters. To simplify the steps required in the process, we derive the two updates separately.

Before deriving the specific update rule, let us state the gradients for the loss computed with respect to generic $\mathbf{K}$ and $\mathbf{P}$. To do this, we take the help of previous derivatives computed in eq.(10-17) and eq.(29-32) such that, for a given random sample pair $(p^{(i)}, g^{(i)})$ we can write

$$\frac{\partial}{\partial \mathbf{K}}\ell_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) = \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) \geq 1 \\ -y_{p^{(i)},g^{(i)}}\mathbf{K}(\mathbf{x}_{p^{(i)}}\mathbf{x}_{g^{(i)}}^T + \mathbf{x}_{g^{(i)}}\mathbf{x}_{p^{(i)}}^T) & \text{otherwise} \end{cases} \tag{45}$$

and

$$\frac{\partial}{\partial \mathbf{P}}\ell_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) = \begin{cases} 0 & \text{if} \quad S_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) \geq 1 \\ y_{p^{(i)},g^{(i)}}\mathbf{P}\left((\mathbf{x}_{p^{(i)}} - \mathbf{x}_{g^{(i)}})(\mathbf{x}_{p^{(i)}} - \mathbf{x}_{g^{(i)}})^T\right) & \text{otherwise} \end{cases} \tag{46}$$

Finally, the derivatives of the loss computed using all pairs in the training set are written as

$$\frac{\partial}{\partial \mathbf{K}}\mathcal{J}_{\mathbf{K},\mathbf{P}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \mathbf{K}}\ell_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) \tag{47}$$

$$\frac{\partial}{\partial \mathbf{P}}\mathcal{J}_{\mathbf{K},\mathbf{P}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \mathbf{P}}\ell_{\mathbf{K},\mathbf{P}}(p^{(i)}, g^{(i)}) \tag{48}$$

Update K: At iteration t, the update rule for the similarity low-rank projection matrix with the proposed stochastic ADMM solution is given by

$$\tilde{\mathbf{K}}^{(t+1)} = \tilde{\mathbf{K}}^{(t)} - \eta \Bigg\{ \frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \Bigg[\Omega_{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}} + \langle \mathbf{\Lambda}^{(s)}, \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}^{(s)} - \mathbf{V}^{(s)} \rangle + \frac{\rho}{2} \left(\left\| \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \right\|_F^2 + \left\| \mathbf{P}^{(s)} - \mathbf{V}^{(s)} \right\|_F^2 \right) \Bigg] \Bigg\} \tag{49}$$

Removing the terms that do not depend on the differentiation variables yields

$$\tilde{\mathbf{K}}^{(t+1)} = \tilde{\mathbf{K}}^{(t)} - \eta \Bigg[\frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \left(\langle \mathbf{\Lambda}^{(s)}, \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \rangle + \frac{\rho}{2} \left\| \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \right\|_F^2 \right) \Bigg] \tag{50}$$

Using the same scaling method to get from eq. (38) to eq.(39) we obtain

$$\tilde{\mathbf{K}}^{(t+1)} = \tilde{\mathbf{K}}^{(t)} - \eta \Bigg[\frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \left(\frac{\rho}{2} \left\| \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} + \mathbf{\Lambda}^{(s)}/\rho \right\|_F^2 \right) \Bigg] \tag{51}$$

from which solving only the last term yields to

$$\tilde{\mathbf{K}}^{(t+1)} = \tilde{\mathbf{K}}^{(t)} - \eta \Bigg(\frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \rho \left(\tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} + \mathbf{\Lambda}^{(s)}/\rho \right) \Bigg) \tag{52}$$

Results for the partial derivatives of the loss computed with respect to a single random sample or with respect of all training ones can be easily obtained from eq.(45-46) and eq.(47-48), respectively.

Once the K iterations over the randomly selected training samples are completed, the low rank projection matrix for the next epoch $s+1$ is obtained as

$$\mathbf{K}^{(s+1)} = \frac{1}{K} \sum_{t=1}^{K} \tilde{\mathbf{K}}^{(t)} \tag{53}$$

Update P: At iteration t, the update rule for the dissimilarity low-rank projection matrix with the proposed stochastic ADMM solution is given by

$$\tilde{\mathbf{P}}^{(t+1)} = \tilde{\mathbf{P}}^{(t)} - \eta \Bigg\{ \frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \Bigg[\Omega_{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}} + \langle \mathbf{\Lambda}^{(s)}, \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \rangle + \langle \mathbf{\Psi}^{(s)}, \mathbf{P}^{(s)} - \mathbf{V}^{(s)} \rangle + \frac{\rho}{2} \left(\left\| \tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} \right\|_F^2 + \left\| \mathbf{P}^{(s)} - \mathbf{V}^{(s)} \right\|_F^2 \right) \Bigg] \Bigg\} \tag{54}$$

Removing the terms that do not depend on the differentiation variables yields

$$\tilde{\mathbf{P}}^{(t+1)} = \tilde{\mathbf{P}}^{(t)} - \eta \Bigg[\frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \left(\langle \mathbf{\Psi}^{(s)}, \tilde{\mathbf{P}}^{(t)} - \mathbf{V}^{(s)} \rangle + \frac{\rho}{2} \left\| \tilde{\mathbf{P}}^{(t)} - \mathbf{V}^{(s)} \right\|_F^2 \right) \Bigg] \tag{55}$$

As before, let us exploit the same scaling method to get from eq. (38) to eq.(39) to get

$$\tilde{\mathbf{P}}^{(t+1)} = \tilde{\mathbf{P}}^{(t)} - \eta \Bigg[\frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \left(\frac{\rho}{2} \left\| \tilde{\mathbf{P}}^{(t)} - \mathbf{V}^{(s)} + \mathbf{\Psi}^{(s)}/\rho \right\|_F^2 \right) \Bigg] \tag{56}$$

from which solving only the last term yields to

$$\tilde{\mathbf{P}}^{(t+1)} = \tilde{\mathbf{P}}^{(t)} - \eta \Bigg(\frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)}, \tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)}, \mathbf{P}^{(s)}} + \rho \left(\tilde{\mathbf{P}}^{(t)} - \mathbf{V}^{(s)} + \mathbf{\Psi}^{(s)}/\rho \right) \Bigg) \tag{57}$$

Again, results for the partial derivatives of the loss computed with respect to a single random sample or with respect of all training ones can be easily obtained from eq.(45-46) and eq.(47-48), respectively.

Once the K iterations over the randomly selected training samples are completed, the low rank projection matrix for the next epoch $s+1$ is obtained as

$$\mathbf{P}^{(s+1)} = \frac{1}{K} \sum_{t=1}^{K} \tilde{\mathbf{P}}^{(t)} \tag{58}$$

For clarification, all the steps required by the proposed stochastic ADMM solution are finally summarized in Algorithm 1.

Algorithm 1: Stochastic ADMM optimization for Low-Rank Sparse Similarity-Dissimilarity Learning

Input: $\eta > 0$, $\rho > 0$, $S > 0$ $\mathbf{K}^{(1)}$,$\mathbf{P}^{(1)}$,$\mathcal{X}$
Iterate $s = 1, \ldots, S$
1. Compute the full partial derivatives using all training samples

$$\frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})$$
$$\frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}}(p^{(i)}, g^{(i)})$$

2. Set

$$\tilde{\mathbf{K}}^{(1)} = \mathbf{K}^{(s)}, \quad \tilde{\mathbf{P}}^{(1)} = \mathbf{P}^{(s)}$$

Iterate $t = 1, \ldots, n$
1. Update $\tilde{\mathbf{K}}^{(t)}$

$$\tilde{\mathbf{K}}^{(t+1)} = \tilde{\mathbf{K}}^{(t)} - \eta \left(\frac{\partial}{\partial \tilde{\mathbf{K}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t)},\tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{K}^{(s)}} \ell_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{K}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}} + \rho \left(\tilde{\mathbf{K}}^{(t)} - \mathbf{U}^{(s)} + \mathbf{\Lambda}^{(s)}/\rho \right) \right)$$

2. Update $\tilde{\mathbf{P}}^{(t)}$

$$\tilde{\mathbf{P}}^{(t+1)} = \tilde{\mathbf{P}}^{(t)} - \eta \left(\frac{\partial}{\partial \tilde{\mathbf{P}}^{(t)}} \ell_{\tilde{\mathbf{K}}^{(t+1)},\tilde{\mathbf{P}}^{(t)}}(p^{(t)}, g^{(t)}) - \frac{\partial}{\partial \mathbf{P}^{(s)}} \ell_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}}(p^{(t)}, g^{(t)}) + \frac{\partial}{\partial \mathbf{P}^{(s)}} \mathcal{J}_{\mathbf{K}^{(s)},\mathbf{P}^{(s)}} + \rho \left(\tilde{\mathbf{P}}^{(t)} - \mathbf{V}^{(s)} + \mathbf{\Psi}^{(s)}/\rho \right) \right)$$

3. Update $\mathbf{K}^{(s)}$ and $\mathbf{P}^{(s)}$

$$\mathbf{K}^{(s+1)} = \frac{1}{n} \sum_{t=1}^{n} \tilde{\mathbf{K}}^{(t)}$$
$$\mathbf{P}^{(s+1)} = \frac{1}{n} \sum_{t=1}^{n} \tilde{\mathbf{P}}^{(t)}$$

4. Update $\mathbf{U}^{(s)}$ and $\mathbf{V}^{(s)}$

$$\mathbf{U}^{(s+1)} = \left(\mathbf{K}_{i,:}^{(s+1)} + \mathbf{\Lambda}_{i,:}^{(s)}/\rho \right) \max \left(0, 1 - \frac{\alpha}{\rho \left\| \mathbf{K}_{i,:}^{(s+1)} + \mathbf{\Lambda}_{i,:}^{(s)}/\rho \right\|_2} \right)$$

$$\mathbf{V}^{(s+1)} = \left(\mathbf{P}_{i,:}^{(s+1)} + \mathbf{\Psi}_{i,:}^{(s)}/\rho \right) \max \left(0, 1 - \frac{\beta}{\rho \left\| \mathbf{P}_{i,:}^{(s+1)} + \mathbf{\Psi}_{i,:}^{(s)}/\rho \right\|_2} \right)$$

5. Update $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Psi}^{(s)}$

$$\mathbf{\Lambda}^{(s+1)} = \mathbf{\Lambda}^{(s)} + \rho \left(\mathbf{K}^{(s+1)} - \mathbf{U}^{(s+1)} \right)$$
$$\mathbf{\Psi}^{(s+1)} = \mathbf{\Psi}^{(s)} + \rho \left(\mathbf{P}^{(s+1)} - \mathbf{V}^{(s+1)} \right)$$

Output: Estimated optimal solutions $\mathbf{K} = \mathbf{K}^{(S)}$ and $\mathbf{P} = \mathbf{P}^{(S)}$

2.3 Discussion

In the preceding section we have derived the deterministic ADMM as well as the proposed stochastic ADMM solutions for the updates of the two low-rank projection matrices $\mathbf{K}$ and $\mathbf{P}$, respectively. In addition, we have given the solution for the update rules involving the remaining components, which, as stated, are shared by both.

While, such last updates are common between the two, the update rules for $\mathbf{K}$ and $\mathbf{P}$, deterministic ADMM implicitly implies that the considered data measurements, i.e., $\mathbf{x}$, are exact, hence there is no noise in any component [19]. Even though the method showed to be successfully applicable even in presence of such a strong assumption, we believed that considering the possibility of noisy data in the optimization would have enabled us to obtain a more robust estimator of the parameter to be optimized. Following such an idea, we have considered a stochastic solution that can handle noisy data. The provided derivations as well as the re-identification performance comparisons between the two solutions (see

Table 6 in the main paper) have shown that the stochastic approach produces better results at a lower computational cost.

References

1. Baltieri, D., Vezzani, R., Cucchiara, R.: 3DPeS: 3D people dataset for surveillance and forensics. In: Joint ACM workshop on Human gesture and behavior understanding, ACM Press (2011) 59–64
2. Pedagadi, S., Orwell, J., Velastin, S.: Local Fisher Discriminant Analysis for Pedestrian Re-identification. In: International Conference on Computer Vision and Pattern Recognition. (2013) 3318–3325
3. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. Conference on Computer Vision and Pattern Recognition (jun 2014) 152–159
4. Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., Wang, J.: Person Re-identification with Correspondence Structure Learning. In: International Conference on Computer Vision. (2015) 3200–3208
5. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Using Kernel-Based Metric Learning Methods. In: European Conference Computer Vision. (2014) 1–16
6. Martinel, N., Micheloni, C., Foresti, G.L.: Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning. IEEE Transactions on Image Processing **24**(12) (dec 2015) 5645–5658
7. Li, X., Zheng, W.s., Wang, X., Xiang, T., Gong, S.: Multi-scale Learning for Low-resolution Person Re-identification. In: International Conference on Computer Vision. (2015) 3765–3773
8. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: International Conference on Computer Vision and Pattern Recognition. (2012) 2288–2295
9. Zheng, W.S., Gong, S., Xiang, T.: Re-identification by Relative Distance Comparison. IEEE transactions on pattern analysis and machine intelligence **35**(3) (jun 2013) 653–668
10. Zhao, R., Ouyang, W., Wang, X.: Unsupervised Salience Learning for Person Re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3586–3593
11. Liao, S., Li, S.Z.: Efficient PSD Constrained Asymmetric Metric Learning for Person Re-identification. In: International Conference on Computer Vision. (2015) 3685–3693
12. Ahmed, E., Jones, M., Marks, T.K.: An Improved Deep Learning Architecture for Person Re-Identification. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2015)
13. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In: International Conference on Computer Vision and Pattern Recognition. (2015)
14. Zhang, R., Lin, L., Zhang, R., Zuo, W., Zhang, L.: Bit-Scalable Deep Hashing With Regularized Similarity Learning for Image Retrieval and Person Re-Identification. IEEE Transactions on Image Processing **24**(12) (dec 2015) 4766–4779
15. Petersen, K.B., Syskind, P.: The Matrix Cookbook. (2012)
16. Boyd, S., Parikh, N., E Chu, B.P., Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning **3**(1) (2010) 1–122
17. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Convex optimization with sparsity-inducing norms. In Sra, S., Nowozin, S., Wright, S.J., eds.: Optimization for Machine Learning. The MIT Press (2011) 1–35

18. Zhao, S.Y., Li, W.J., Zhou, Z.H.: Scalable Stochastic Alternating Direction Method of Multipliers. arXiv preprint arXiv:1502.03529 (2015)
19. Ouyang, H., He, N., Tran, L., Gray, A.: Stochastic alternating direction method of multipliers. International Conference on Machine Learning (2013) 80–88