LOW-COST MULTI-CAMERA OBJECT MATCHING

Syed Fahad Tahir and Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London,
{s.fahad.tahir, andrea.cavallaro}@eecs.qmul.ac.uk

ABSTRACT

We propose an object matching approach aimed at smartphone cameras that exploits the well-known concept of local sets of features for object representation. We also enable the temporal alignment of cameras by exploiting the frames of detected objects to group objects appeared in the same time interval for the assignment within each camera. The proposed approach does not need training thus making it suitable for matching during short temporal intervals. We use both outdoor and indoor datasets for the evaluation, and show that the proposed method reduces up to 95% the amount of information to be stored and communicated.

Index Terms— Smart cameras, object matching, data reduction, cost of features, temporal grouping

1. INTRODUCTION

Object matching is a fundamental task in multi-view and multi-camera scene observation [1, 2, 3, 4, 5]. Existing approaches generally focus on object matching accuracy without considering constraints on the available resources, such as storage, battery life, communication and computational capabilities [6], which are important for battery-powered devices such as smartphones and wireless smart cameras. In this paper we consider storage and communication constraints, typical of handled devices, while exploiting a minimum amount of prior information on the environment for object association.

Multi-view object matching may be performed by estimating the object position in the scene using camera projection matrices [7, 8, 9]. However in the case of hand-held cameras, because of camera motion, a continuous re-calibration and re-estimation is required. In the case of cameras with disjoint field-of-views, other features such as appearance information and inter-camera transition times can be exploited [3, 10, 11, 12, 13]. Methods extract from single [6, 10, 14] or multi-shot object views [15] different feature types, such as appearance information encoded in color histograms and texture descriptors [6, 11, 12, 14, 16], their relative positioning [17], and high-dimensional feature-point descriptors such as SIFT and HoG [15]. These features used for matching need to be communicated over a network. Learning-based approaches using AdaBoost [14] and rankSVM [10] for object association may only be applied when sufficient training data is available. In the case of insufficient training data, Direct Distance Minimization (DDM) approaches such as those based on the Kullback-Leibler [2], Bhattacharyya [1] or Euclidean distance for object matching are applied. However, DDM approaches are generally less robust to illumination changes, which can be compensated by learning inter-camera color transformations [1, 2, 3].

Matching moving objects may also exploit the cross camera spatio-temporal information along with the appearance [3, 11]. This requires camera synchronization as the occurrence of the same event has happened simultaneously (i.e. at the same time-stamp) in the camera views. Most existing synchronization approaches are applied to fixed cameras [18, 19] with knowledge of prior information of the scene under observation. Approaches also exist for moving cameras [20], which exploit known object matching information.

In this paper, we propose a simple yet effective object matching approach that minimizes the amount of data to be shared among cameras and that uses limited prior scene information for the matching. We define a compact object representation and a temporal grouping of objects within each camera to restrain the assignments in the defined temporal boundaries. We evaluate the approach on three datasets, namely an in-house dataset recorded with three hand-held cameras and two publicly available datasets from iLIDS. The proposed approach reduces the amount of data needed to be transferred while improving the matching rates with respect to existing approaches. The software implementing the proposed approach and the in-house dataset used in the evaluation are available at *http://www.eecs.qmul.ac.uk/~andrea/matching.html*.

The paper is organized as follows. Sec. 2 discusses the proposed compact object representation. In Sec. 3, we discuss the object matching in the defined temporal groups using the representation models. In Sec. 4, we evaluate and compare the proposed approach with existing methods. Finally, Sec. 5 draws conclusions and discusses the future work.

2. OBJECT REPRESENTATION

Let $\mathbf{C} = \{C_n\}_{n=1}^{N}$ be a set of N hand-held smart cameras with partially overlapping field-of-views. We assume that object detection and tracking have been solved [21, 22] within each camera independently. Let $\mathbf{P}_n = \{P_{mn}\}_{m=1}^{M}$ be the set of M objects detected in C_n and represented by the extracted images. The features extracted from P_{mn} are used for the matching (Fig. 1).

In order to minimize the effect of illumination variations and contrast adjustment for each P_{mn}, we perform histogram equalization within each camera [23]:

$$hist_{eq}^{(i)} = \left\lfloor \frac{\left(L \times hist_{cf}^{(i)}\right) - (h \times w)}{(h \times w)} \right\rfloor, \qquad (1)$$

where L is the number of intensity levels, h and w are the height and the width of the object image P_{mn} in pixels, and $hist_{cf}^{(i)}$ is the cumulative sum of the histogram until the bin with intensity value i

S.F. Tahir was supported by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the Education, Audiovisual & Culture Executive Agency (FPA n° 2010-0012). A. Cavallaro acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/K007491/1.

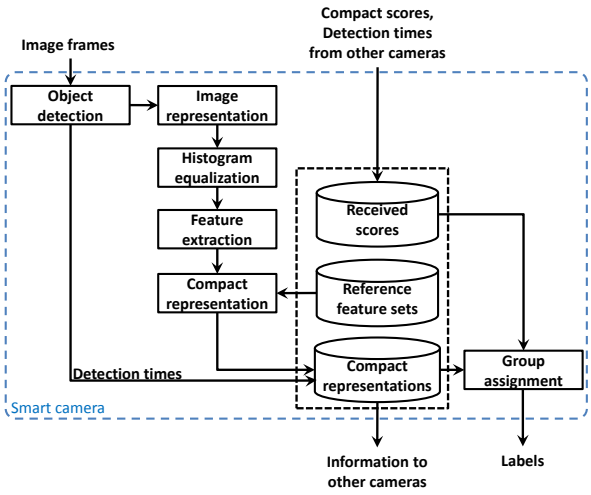


Fig. 1. Block diagram of the proposed object matching approach.

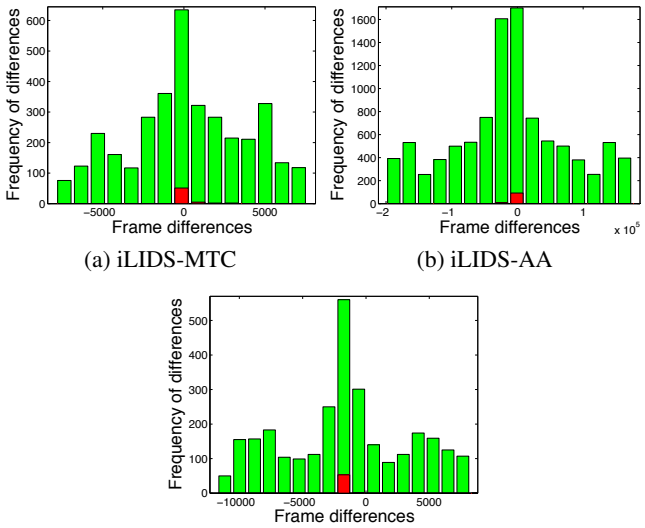


Fig. 2. Histograms of differences of the detection frame-numbers. The green bars show all the possible differences between detection pairs across two cameras. The red bars show the differences in detections of the same object in two cameras.

in P_{mn}. For each R, G, and B color plane, the intensity value i of the image is replaced with $hist_{eq}^{(i)}$ and a potentially narrow band of colors is spread over the whole available intensity range.

From each histogram-equalized image, we extract a set of R appearance features $\mathbf{F}_{mn} = \{f_{mn}^r\}_{r=1}^R$ as in [14, 10, 12]. To reduce the cost of storage and transfer of $\mathbf{F}_{mn}$, we generate a compact representation $\mathbf{\Omega}_{mn} = \{\Omega_{mn}^k\}_{k=1}^K$ for each P_{mn} by measuring the difference between the extracted feature set $\mathbf{F}_{mn}$ and K reference feature sets $\{\mathbf{\Gamma}_k\}_{k=1}^K$ within each camera as

$$\mathbf{\Omega}_{mn} = \{||\mathbf{\Gamma}_k, \mathbf{F}_{mn}||\}_{k=1}^K, \tag{2}$$

where $||.||$ is the Euclidean norm. In order to obtain $\mathbf{\Gamma}_k$, we use an image dataset [14] for reference images. Unlike other methods for image retrieval and classification [24], we have no scene dependency requirements. The only requirement is that the set of features extracted from the detected object is the same as that of extracted from the reference images. The extracted feature sets from the reference dataset are clustered using the Lloyd's algorithm [25]. The clustering returns K clusters of feature sets, where K is fixed to the number of features, i.e. R, and the mean of each cluster represents one reference feature set $\mathbf{\Gamma}_k$. Similarly to the bag-of-words model, each camera locally stores $\{\mathbf{\Gamma}_k\}_{k=1}^K$. The compact representation $\mathbf{\Omega}_{mn}$ (Eq. 2) we use for matching the objects across cameras, reduces the amount of data for local storage and communication.

3. OBJECT MATCHING

We perform inter-camera object matching by group assignment using $\mathbf{\Omega}_{mn}$ of the detected objects. Let us consider two (unsynchronized) cameras C_n and C_l, where $n, l \in N$ and $n \neq l$. Let P_{mn} be detected and tracked between frames $T_{mn}^{(s)}$ and $T_{mn}^{(e)}$ in C_n, where s and e indicate the start and the end frames of a tracked object. The number of frames ω_{mn} during which P_{mn} is tracked are $\omega_{mn} = T_{mn}^{(e)} - T_{mn}^{(s)}$.

For each P_{mn}, we define a temporal search window $\mathbf{\Phi}_{ml}$ in C_l representing the time interval in which P_{mn} is likely to be observed in C_l. In order to select $\mathbf{\Phi}_{ml}$, we apply a plesiochronous approach to perform the temporal alignment of the cameras. Let $\mathbf{P}_l = \{P_{ql}\}_{q=1}^Q$ be Q objects in C_l, first detected in frames $\{T_{ql}^{(s)}\}_{q=1}^Q$. For each $T_{mn}^{(s)}$, we obtain a set $\mathbf{\Lambda}_{nl}^m$ of Q differences from $\{T_{ql}^{(s)}\}_{q=1}^Q$ in C_l as

$$\mathbf{\Lambda}_{nl}^m = \{\alpha T_{mn}^{(s)} - T_{ql}^{(s)}\}_{q=1}^Q, \tag{3}$$

where α is the ratio of the frame rates of C_l and C_n. For M detected objects in C_n, we obtain an $M \times Q$ difference matrix $\mathbf{\Delta}_{nl} = \{\mathbf{\Lambda}_{nl}^m\}_{m=1}^M$. By analyzing the distribution of values in $\mathbf{\Delta}_{nl}$, we can observe that the difference of frame numbers of the first frames of two different tracked objects detected in C_n and C_l can vary significantly, while the difference between the first frames of the same objects detected in two cameras consistently remains within a narrow range $\mathbf{\Upsilon}_{nl}$. In order to identify $\mathbf{\Upsilon}_{nl}$, we take the histogram of values in $\mathbf{\Delta}_{nl}$ (see Fig. 2). The bin size of the histogram depends on the average number of frames during which an object remains visible in C_n, measured as $\tilde{\omega}_n = \frac{1}{M}\sum_{m=1}^{M}\omega_{mn}$. The bin with the most frequently occurring values, $\mathbf{\Upsilon}_{nl}$, is represented as

$$\mathbf{\Upsilon}_{nl} = [S_{nl} \pm \alpha\tilde{\omega}_n], \tag{4}$$

where S_{nl} is the time shift (in number of frames) between C_n and C_l, measured as the mean of the values in the bin $\mathbf{\Upsilon}_{nl}$. Using $\mathbf{\Upsilon}_{nl}$, we estimate the temporal search window $\mathbf{\Phi}_{ml}$ for P_{mn} as

$$\mathbf{\Phi}_{ml} = \left[\alpha T_{mn}^{(s)} + \mathbf{\Upsilon}_{nl}\right]. \tag{5}$$

The objects detected in C_l within $\mathbf{\Phi}_{ml}$ are the candidates for matching with P_{mn}.

In order to find the association between the objects, we measure the Bhattacharyya distance between $\mathbf{\Omega}_{mn}$ of P_{mn} and $\acute{Q}_m$ compact representations $\{\mathbf{\Omega}_{ql}\}_{q=1}^{\acute{Q}_m}$ of the objects detected within $\mathbf{\Phi}_{ml}$ in C_l

$$\mathbf{\Pi}_{mn} = \{\Pi_{mn}^q\}_{q=1}^{\acute{Q}_m} = \left\{-ln\left(\sum_{k=1}^{K}\sqrt{\Omega_{mn}^k.\Omega_{ql}^k}\right)\right\}_{q=1}^{\acute{Q}_m}, \tag{6}$$

(a) Torch-C3 (b) Torch-C2

Fig. 3. Two sample frames from (a) C_3 and (b) C_2 in the Torch dataset. These frames are captured almost at the same time instance and represent two very different views of the same scene.

where $\mathbf{\Pi}_{mn}$ is the set of $\acute{Q}_m$ differences from P_{mn}. The assignment of P_{mn} to P_{ql} with the minimum distance from P_{mn} in $\mathbf{\Pi}_{mn}$ results in multiple assignments to a single object because P_{ql} can also have minimum distance in another search window. Unlike the distance minimization, we perform a group assignment and the correct match is selected by optimal assignment within the group using the Hungarian algorithm [26]. We find distances of the compact representations of M objects in C_n from the groups of objects detected within their corresponding temporal search windows in C_l while assigning large distances to the remaining $|Q| - |\acute{Q}_m|$ objects outside their temporal search windows $\mathbf{\Phi}_{ml}$. This results in an $M \times Q$ matrix $\mathbf{H} = \{\mathbf{\Pi}_{mn}\}_{m=1}^{M}$. The labels are assigned without repetition to the objects in two cameras such that the summation between the assigned pairs in the group remains minimum.

4. EVALUATION AND DISCUSSION

We compare the proposed approach with the following DDM, learning and probabilistic methods: the Bhattacharyya distance, RankSVM [10], Attribute-Sensitive Feature Importance (ASFI) [16], Probabilistic Relative Distance Comparison (PRDC) [12] and Landmark Based Model (LBM) [11], where RankSVM, ASFI and PRDC require training data and LBM utilizes the spatio-temporal information along with the appearance features within the fixed cameras. For the comparison, we use three *people datasets*, namely the outdoor dataset *Torch* and two publicly available indoor datasets from iLIDS; and we assume that the results generated by a person detector and tracking method are available as input to our pipeline. The Torch dataset contains videos with three partially overlapping views recorded during the Olympics 2012 torch relay passing through Mile End road in London, UK. The recordings contain a crowd scene captured with hand-held smartphones, thus leading to occasional jitters and blurring (Fig. 3) in addition to changes in illumination, size and pose of people, and occlusions. Single images of $|\mathbf{P}_n| = 50$ people common in the three cameras are manually extracted on their first appearance in each camera and their detection frame is stored. The proposed object matching approach is applied in a pairwise manner in the three cameras. As for the other two image datasets: iLIDS-AA [13] contains multiple images of 100 individuals automatically extracted using a HoG detection algorithm, and iLIDS-MTC [11] contains manually cropped multiple images of 60 pairs of persons in 2 cameras. Since we require a dataset with single images and the detection information (frame numbers) for evaluation, we select the single cropped images of each person in the two datasets along with their detection frame numbers.

We use the validation criteria that are based on the amount of data to be communicated among cameras and the matching rate using the Cumulative Matching Characteristics (CMC) curves. CMC curves show the true target rates for given false target rates.

Table 1. Comparison of the amount of data per person needed to be stored within the camera for object matching.

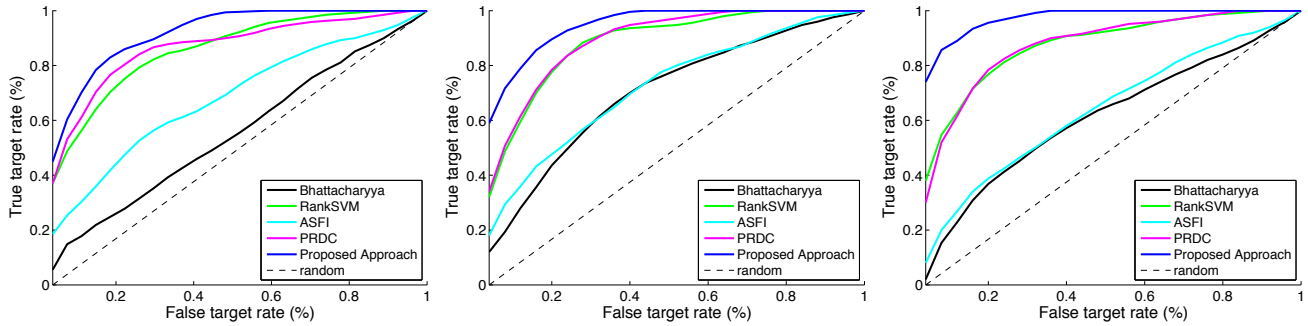
Dataset	Number of features	Number of people	Bytes per person	
			$\mathbf{F}_{mn}$	$\mathbf{\Omega}_{mn}$
Torch	29	54	7539	64
iLIDS-MTC [11]	29	60	7415	64
iLIDS-AA [13]	29	100	6422	62

Each image is equalized, and we extract color and texture features, where each feature is a 16-bin histogram of a color channel or a filtered image, extracted from each of the 6 horizontal stripes of the person image as in [10, 14, 12]. Although the proposed approach does not require a training for the matching, in order to compare with the existing learning approaches, we applied the 2-fold cross validation using half of the data for training and the remaining for testing of existing approaches. The data generated by the compact representations of all the detected persons in a camera is encoded using the lossless data compression algorithm deflate [27], which combines LZ77 and Huffman coding.

Table 1 shows the amount of data that needs to be stored and communicated per person between the cameras. It can be observed that the storage size per person is reduced to 1% using the compact representation $\mathbf{\Omega}_{mn}$ of the proposed approach as compared to that of the initial feature set $\mathbf{F}_{mn}$, since $\mathbf{F}_{mn}$ for each person contains 2784 elements for 29 features extracted from 6 stripes of the image, whereas $\mathbf{\Omega}_{mn}$ contains only $K = 29 \times 6$ elements. In addition, we require 170 KB per camera for storing the reference feature sets $\{\mathbf{\Gamma}_k\}_{k=1}^{K}$. The size of the additional storage requirement is a constant that is not affected by the observed number of persons and can be pre-allocated.

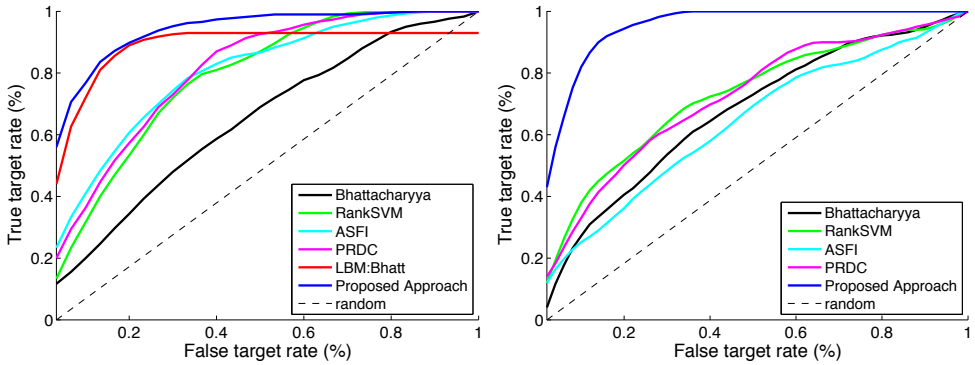
Figure 4 shows the matching rate of objects from three camera pairs in the Torch dataset. The proposed approach shows the highest matching results between 50% and 75% true target rate for zero false targets, as compared to the existing approaches showing a maximum of 40% true target rate for zero false targets in all three pairs of camera settings. In hand-held data gatherings using smart cameras, a sufficient data from the same scene may not always be possible that limits the training of the learning methods and their performance is compromised. The DDM approach shows the minimum performance in the absence of illumination and contrast handling. Additionally the proposed approach effectively reduces the search space for matching by locally estimating the inter-camera temporal shift, which results in a higher matching rate.

Figure 5 shows the evaluation results of the proposed approach on two datasets from the pair of cameras in iLIDS. The proposed approach shows the higher matching rates with 60 % and 45% true target rates at zero false target rate in iLIDS-MTC and iLIDS-AA respectively as compared to the existing approaches. In iLIDS-MTC, we also compare the proposed approach with LBM, a spatio-temporal and appearance approach requiring the actual map and the location of people in the scene along with the appearance information. Our approach outperforms LBM, without requiring the spatial information of the scene and only utilizes the detection frame numbers, thus allowing it to be applied in devices which vary their locations (Torch data). The performance of the learning methods is again affected by the amount of training data. In the iLIDS-AA dataset, since the objects are extracted after applying the HoG detection algorithm, even in the case of true detections the extracted



(a) Torch C21, $|\mathbf{P}_n| = 27$ (b) Torch C31, $|\mathbf{P}_n| = 25$ (c) Torch C32,$|\mathbf{P}_n| = 25$

Fig. 4. CMC curves obtained for matching using the existing approaches: PRDC [12], ASFI [16] and rankSVM [10] compared with the proposed approach on the new Torch dataset with 3 hand-held cameras. The matching is performed pairwise when an object is observed in (a) C2 and C1, (b) C3 and C1, and (c) C3 and C2.



(a) iLIDS-MTC, $|\mathbf{P}_n| = 30$ (b) iLIDS-AA, $|\mathbf{P}_n| = 50$

Fig. 5. CMC curves obtained for matching using existing approaches: PRDC [12], ASFI [16], rankSVM [10] and LBM [11] compared with the proposed approach in two existing datasets extracted from iLIDS: (a) iLIDS-MTC [11], (b) iLIDS-AA [13].

image may not have the complete representation of the object. In such scenarios the temporal grouping of the proposed approach improves the overall performance as compared to the approaches based only on appearance information.

5. CONCLUSIONS

We proposed a simple yet effective object matching approach, which significantly reduces the amount of data needed to be stored and shared among cameras for object representation. The approach maximizes the dependency on locally available information for object matching and achieves a higher matching rate compared to existing approaches. The amount of data needed to be communicated is less than 100 bytes per person and requires local storage for the reference feature sets. Using one class of features only, we achieved up to 75% matching accuracy in the Torch dataset recorded with hand-held cameras.

Our future work includes extending the evaluation to other feature types and we aim to port the algorithm into actual smartphones and smart camera platforms to analyze its performance beyond the current simulation environment.

6. REFERENCES

[1] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proc. BMVC*, Leeds, UK, Sep. 2008.

[2] K. Jeong and C. Jaynes, "Object matching in disjoint cameras using a colour transfer approach," *Springer J. of Mach. Vis. and Appl.*, vol. 19, no. 5, pp. 88–96, 2008.

[3] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views," *Comput. Vis. Image Understand.*, vol. 109, no. Issue 2, pp. 146 – 162, 2008.

[4] L.F. Teixeira and L.Corte-Real, "Video object matching across multiple independent views using local descriptors and adaptive learning," *Pattern Recognition Letters*, vol. 320, no. 2, pp. 157–167, 2009.

[5] V. Sulic, J. Pers, M. Kristan, and S. Kovacic, "Efficient feature distribution for object matching in visual-sensor networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 903 –916, 2011.

[6] S.F. Tahir and A. Cavallaro, "Cost-effective features for re-identification in camera networks," *IEEE Trans. Circuits Syst. Video Technol.*, (to appear).

[7] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *Int. J. Comput. Vision*, vol. 66, pp. 231–259, 2006.

[8] M. Ferecatu and H. Sahbi, "Multi-view object matching and tracking using canonical correlation analysis," in *Proc. IEEE ICIP*, Cairo, Egypt, Nov. 2009.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Number ISBN: 978-0-521-54051-3. Cambridge University Press, 4th edition, 2006.

[10] B. Prosser, W.S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMVC*, Aberystwyth, UK, Aug. 2010.

[11] R. Mazzon, S.F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1828–1837, 2012.

[12] W.S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, Colorado Springs, USA, Jun. 2011.

[13] S. Bak, E. Corvee, F. Bremond, and M.Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *Proc. IEEE AVSS*, Klagenfurt, Austria, Sep 2011.

[14] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, Marseille, France, Oct. 2008.

[15] N. Martinel, C. Micheloni, and C. Piciarelli, "Distributed signature fusion for person re-identification," in *Proc. ACM/IEEE ICDSC*, Oct. 2012.

[16] C. Liu, S. Gong, C.C. Loy, and X. Lin, "Person re-identification: What features are important?," in *Proc. ECCV*, Firenze, Italy, Oct. 2013.

[17] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, 2013.

[18] F.L.C. Padua, R.L. Carceroni, G.A.M.R. Santos, and K.N. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 304–320, 2010.

[19] Y. Caspi, D. Simakov, and M. Irani, "Feature-based sequence-to-sequence matching," *Int. J. Comput. Vision*, vol. 68, no. 1, pp. 53–64, 2006.

[20] L. Zini, A. Cavallaro, and F. Odone, "Action-based multi-camera synchronization," *IEEE J. on Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 2, pp. 165–174, 2013.

[21] S. Gong, T. Xiang, and S. Hongeng, "Learning human pose in crowd," in *Proc. ACM Multimedia*, Firenze, Italy, Oct. 2010.

[22] Y. Ishii, H. Hongo, K. Yamamoto, and Y. Niwa, "Face and head detection for a real-time surveillance system," in *Proc. Int. Conf. on Pattern Recognition*, Cambridge, UK, Aug. 2004.

[23] T. Acharya and A.K. Ray, *Image Processing - Principles and Applications*, Number ISBN: 0-47171-998-6. Wiley-Interscience, 2005.

[24] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Processing*, vol. 93, no. 8, pp. 2316 – 2327, 2013.

[25] S.P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[26] H.W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[27] D. Salomon, *Data Compression: The Complete Reference*, Number ISBN: 1-84628-602-6. Springer, 2007.