MIXER: A Principled Framework for Multimodal, Multiway Data Association

Parker C. Lusk*, Ronak Roy*, Kaveh Fathian*, Jonathan P. How

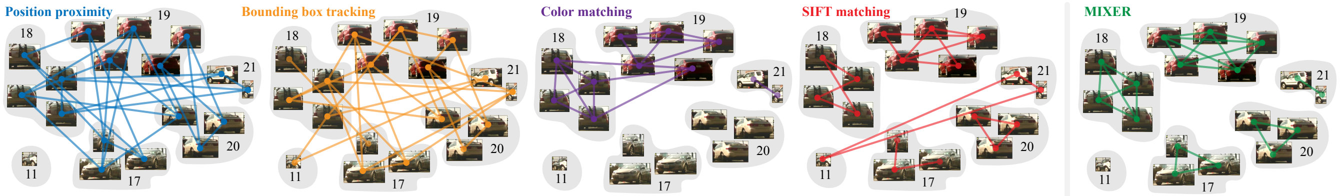

Fig. 1. Small-scale demonstration of MIXER associating identical cars. Each labeled cluster (grey) represents multiple observations of the same car. Based on four distinct matching modalities of position proximity, bounding box tracking, color, and SIFT matching, observations are associated if they are sufficiently similar. While each modality has erroneous associations (i.e., observations of different cars are thought to be of the same car), MIXER combines these modalities to yield a fully-accurate set of associations, correctly grouping all observations of the same car.

***Abstract*— A fundamental problem in robotic perception is matching identical objects or data, with applications such as loop closure detection, place recognition, object tracking, and map fusion. While the problem becomes considerably more challenging when matching should be done jointly across multiple, multimodal sets of data, the robustness and accuracy of matching in the presence of noise and outliers can be greatly improved in this setting. At present, multimodal techniques do not leverage multiway information, and multiway techniques do not incorporate different modalities, leading to inferior results. In contrast, we present a principled mixed-integer quadratic framework to address this issue. We use a novel continuous relaxation in a projected gradient descent algorithm that guarantees feasible solutions of the integer program are obtained efficiently. We demonstrate experimentally that correspondences obtained from our approach are more stable to noise and errors than state-of-the-art techniques. Tested on a robotics dataset, our algorithm resulted in a 35% increase in F_1 score when compared to the best alternative.**

I. Introduction

Identifying correspondences across sets of data, or across sources with different modalities, is a fundamental problem in robotics and computer vision. In practice, data points are noisy and contain outliers that should not be matched. These challenges make classical assignment techniques such as the Hungarian [1] or auction [2] algorithms ineffective, as they cannot reject outliers or produce consistent results across sets. State-of-the-art multiway data association algorithms [3], [4], [5], [6], [7], [8], [9] can remove outliers and produce consistent associations by matching data jointly across all sets, however, these techniques cannot fuse associations that come from different modalities. Furthermore, these schemes operate based on binary associations, i.e., data points should either be matched or not. Considering lack of information (i.e., when correspondences cannot be established and the decision should be delayed) is an important feature that is not currently present. In this work, we present the **M**ultimodality association matr**IX** fus**ER** (MIXER) algorithm. MIXER is a principled framework for associating data that contains outliers, comes from different modalities, and is observed across multiple instances. The MIXER formulation fuses modalities and allows incorporating uncertain or missing information in the decision making process. This is crucial if a modality is not applicable for establishing correspondences at a certain time. The associations returned by MIXER are guaranteed to be consistent across all sets and respect additional constraints imposed on the problem. A small-scale demonstration is shown in Fig. 1, where MIXER recovers 100% precision with 100% recall from the noisy input.

We formulate the problem as a mixed-integer quadratic program (MIQP). Since this MIQP is not scalable to large-sized problems, we present a continuous relaxation to efficiently obtain approximate solutions. The main contribution of our approach over similar relaxation techniques used in the literature is that solutions of the relaxed problem are guaranteed to converge to feasible, binary solutions of the original problem. Thus, rounding results to binary values, which is required when using other techniques and may lead to infeasible solutions, is avoided. To solve the relaxed problem efficiently, we present a projected gradient descent algorithm with backtracking line search based on the Armijo procedure. This polynomial-time algorithm has worst case cubic complexity in problem size (from matrix-vector multiplications) at each iteration, and is guaranteed to converge to stationary points. Proofs are omitted due to space limitations, but will be provided in future work.

We evaluate MIXER on both synthetic and real-world datasets and compare the results with state-of-the-art multiway data association algorithms. Our synthetic analysis demonstrates a small optimality gap between MIXER's solution and the global minimum of the MIQP (max 3.3%), while achieving an average runtime of $252\,\mathrm{ms}$—an average speedup of $685\times$ over the MIQP solver. Our real-world evaluation considers associating identical cars observed by a robot moving in a parking lot. We use the four distinct modal-

P. C. Lusk, R. Roy, K. Fathian and J. P. How are with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology. {plusk, ronakroy, kavehf, jhow}@mit.edu. *Authors contributed equally.

This work is supported by ARL DCIST under Cooperative Agreement Number W911NF-17-2-0181.

ities of proximity, color, image features, and bounding box tracks for associating cars. Benchmarking MIXER against the state-of-the-art shows superior accuracy on individual modalities, and further shows that MIXER is effectively able to combine all similarity scores, improving the F_1 score by 35% compared to the best competing algorithm. In summary, the main contributions of this work include:

- A novel and principled MIQP framework for multi-modal, multiway data association
- A continuous relaxation of the MIQP leading to feasible, binary solutions that can be computed efficiently.
- A polynomial-time algorithm for solving the relaxed problem based on projected gradient descent and Armijo procedure with convergence guarantees.
- Improvements over state-of-the-art multiway data association algorithms showcased on a realworld dataset with four distinct modalities.

We expect MIXER to significantly improve the accuracy and robustness of existing data association techniques used in robotics and computer vision applications such as feature matching [4], multiple object tracking [10], person reidentification (ReID) [11], [12], place recognition [13], and loop closure detection [14].

A. Related Work

Associating elements from two sets is traditionally formulated as a linear assignment problem [15], [16], which can be solved in polynomial time [1], [2]. If elements have underlying structure (e.g., geometry) that should be considered, the problem can be formulated as a quadratic assignment program [17], [18], [19]. Unlike linear assignment, quadratic assignment (or its MIQP graph matching formulation) is, in general, NP-hard [20]. Exact methods for solving quadratic assignment use expensive branch and bound techniques [21], [22]. More efficient, but approximate solutions can be obtained, for example, from spectral relaxations [23], linear relaxations [24], and convex relaxations [25], [26], [27], [28].

Multiway association. Multiway data association frameworks jointly associate elements across multiple sets to ensure (cycle) consistency of associations. This can be formulated as permutation synchronization [3], which is computationally challenging due to binary constraints. With the exception of (expensive) combinatorial methods [29], [30], existing works focus on relaxations to obtain approximate answers; for example, spectral relaxation [3], [4], convex relaxation [31], [32], [33], [5], [34], matrix factorization [6], [7], [35], and graph clustering [36], [8], [37], [9]. While some of these methods accept weighted inputs, they are formulated for synchronizing unimodal, binary associations. Lastly, when data has underlying structure, the formulation becomes a multi-graph matching problem [38], [39], [36], [40], which is considerably more computationally demanding.

Multimodal association. Fusing multiple modalities for data association [41], [42], [43], [44] can be seen as a classifier combination problem [45], [46]. Methods of combining classifiers include evidential reasoning [47], [48], [49], Bayesian methods [50], and deep learning approaches [51]. A popular and foundational approach to combine classifiers is AdaBoost [52], which has been used to combine different types of LiDAR features for loop closure detection [53].

II. MIXER Formulation

Consider n sets of data $\mathcal{A}_i$, $i = 1, \dots, n$, with cardinality $|\mathcal{A}_i| = m_i$ and $m = \sum_{i=1}^{n} m_i$. We define the *universe* as $\mathcal{U} \overset{\text{def}}{=} \cup_i \mathcal{A}_i$, with $|\mathcal{U}| = m_u \leq m$ the number of distinct elements across all sets. For each of the l modalities which associate identical elements across sets, a scalar *similarity score* $s \in [0, 1]$ is produced. Scores of 1, 0.5, and 0 correspond to maximum similarity, lack of information/preference, and maximum dissimilarity, respectively. We arrange these scores in diagonal *score matrices* defined as $S = \mathrm{diag}(s_1, \dots, s_l) \in [0, 1]^{l \times l}$.

Given two sets $\mathcal{A}_i$ and $\mathcal{A}_j$, we define the *association matrix* between the elements of these sets as

$$A_{ij} \overset{\text{def}}{=} \begin{bmatrix} S_{11} & \cdots & S_{1m_j} \\ \vdots & \ddots & \vdots \\ S_{m_i1} & \cdots & S_{m_im_j} \end{bmatrix} \in [0,1]^{m_i l \times m_j l}, \tag{1}$$

where S_{ab} denotes the score matrix between elements $a \in \mathcal{A}_i$ and $b \in \mathcal{A}_j$. These pairwise A_{ij} are used to create the symmetric *aggregate association matrix* between all sets as

$$A \overset{\text{def}}{=} \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix} \in [0,1]^{ml \times ml}. \tag{2}$$

Ground-truth association. In the ideal setting, score matrices S are either identity or zero and can be compactly represented using Kronecker products as $1 \otimes I_l$ or $0 \otimes I_l$, where I_l is an $l \times l$ identity matrix. Furthermore, matrix A can be factorized as $A = U U^\top \otimes I_l$, where

$$U^\top \overset{\text{def}}{=} \begin{bmatrix} U_1^\top & \dots & U_n^\top \end{bmatrix} \in \{0,1\}^{m_u \times m}. \tag{3}$$

Matrices $U_i \in \{0, 1\}^{m_i \times m_u}$ represent associations between elements of A_i and the universe $\mathcal{U}$.

Constraints. Often, data association algorithms must meet certain constraints imposed by the high-level task. The *one-to-one* constraint states that an object cannot be associated with other objects and is satisfied if each row of U has a single 1 entry. The *distinctness* constraint states that objects within a set are distinct and therefore should not be associated. This is satisfied if there is at most a single 1 entry in each column of U_i. When the association problem is solved across more than two sets, it is important to ensure associations are *cycle consistent*, which states that if $a \sim b$ and $b \sim c$, then $a \sim c$ and is satisfied if the association matrix can be factorized as $A = U U^\top \otimes I_l$. This fact is proven in [8] for associations of single modality ($l = 1$), which generalizes similarly to the multimodal case ($l \geq 2$). These three constraints are crucial for detecting and correcting erroneous similarity scores and associations, but increase the difficulty and complexity of the data association procedure.

Optimization problem. In practice, similarity scores are noisy and can lead to incorrect associations. Therefore, our objective is to map these scores to their closest binary value while respecting/leveraging the one-to-one, distinctness, and cycle consistency constraints to correct potential scoring mistakes. This goal can be formally stated as finding U that solves the MIQP with Frobenius objective

$$\begin{aligned} \underset{U\in\{0,1\}^{m\times m}}{\text{minimize}} \quad & \left\|U\,U^\top \otimes I_l - A\right\|_F^2 \quad \text{(cycle consistency)} \\ \text{subject to} \quad & U\,\mathbf{1}_m = \mathbf{1}_m \qquad \text{(one-to-one constraint)} \\ & U_i^\top \mathbf{1}_{m_i} \leq \mathbf{1}_{m_i} \quad \text{(distinctness constraint)} \end{aligned} \tag{4}$$

where $\mathbf{1}$ denotes a vector of ones. We note that the size of U in (4) is $m \times m$ as opposed to $m \times m_u$ defined in (3). This is because the number of unique elements across all sets (size of universe, m_u) is unknown a priori. A solution U of (4) has only m_u nonzero columns.

III. Continuous Relaxation and Algorithm

Solving (4) to global optimality becomes impractical as the problem size grows; hence, we propose a relaxation. Existing relaxation techniques require rounding, which may produce infeasible solutions. A key contribution of our approach is that solutions are guaranteed to converge to feasible, binary solutions of the original problem without rounding.

Manipulation of the objective of (4) and the use of penalty functions to incorporate the constraints into the objective yields the continuous relaxation over the nonnegative reals

$$\begin{aligned} \underset{U\in R_+^{m\times m}}{\text{minimize}} \quad & \langle UU^\top, \bar{A}\rangle + d\left(\langle U^\top U, P_o\rangle + \langle UU^\top, P_d\rangle\right. \\ & \left. + \langle U^\top U - U^\top - U,\ \mathbf{1}_{m\times m}\rangle\right) \\ \text{subject to} \quad & U\mathbf{1}_m - \mathbf{1}_m \leq \mathbf{0}_m \end{aligned} \tag{5}$$

where penalty matrices P_o and P_d correspond to orthogonality and distinctness constraints, and $d \geq 0$ is a scalar parameter. As d increases, the positive penalty value pushes the solution U toward having orthogonal columns, satisfying the distinctness constraint, and having row-sum equal to 1. Once d is large enough, solutions of (5) become binary and satisfy the constraints of the original problem (4). Note that $\bar{A}$, defined in the expansion of objective of (4), can be indefinite, making (5) non-convex in general.

Problem (5) is efficiently solved using a first-order projected gradient descent scheme with greedy steps and backtracking Armijo line search [54]. The worst case complexity of this local search is bounded by $\mathcal{O}(m^3)$ per iteration, corresponding to matrix multiplications.

IV. Experiments

We evaluate MIXER on two datasets. First, we use synthetic data and study the optimality gap of MIXER, finding that over the problem sizes considered, MIXER can achieve near-optimal performance with significantly improved runtime. Then, we demonstrate the ability of MIXER to combine sensing modalities on a challenging robotics dataset collected as part of this work. We compare the performance of MIXER on this dataset with other multiway matching algorithms.

Synthetic Dataset. We use synthetically generated data to empirically analyze MIXER's optimality gap and runtime. Data is generated using partial views of $m_u = 10$ objects with randomly added outliers and noise, resulting in a problem sizes of $10 \leq m \leq 50$. We use Gurobi 9.1.1 [55] to solve problem (4) to optimality and refer to this implementation as MIXER*. MIXER is executed in MATLAB on an i7-6700.

TABLE I. MIXER optimality gap and runtime speedup relative to MIXER*, determined over 50 Monte Carlo trials. MIXER obtains considerable speedup while maintaining a low optimality gap, indicating a large basin of attraction.

n_o	gap (%)	p (%)	r (%)	runtime (ms)	speedup
2	0.8 ± 2.5	0.9 ± 7.3	-1.5 ± 6.2	142 ± 7	$0.1\times$
4	0.5 ± 0.7	5.6 ± 10.8	-0.6 ± 5.6	149 ± 7	$1.1\times$
6	0.4 ± 0.5	5.8 ± 12.5	-2.0 ± 8.5	161 ± 11	$6.5\times$
8	0.4 ± 0.3	7.8 ± 10.5	-0.7 ± 8.9	201 ± 30	$29.1\times$
10	0.4 ± 0.2	5.3 ± 11.1	-2.4 ± 10.4	218 ± 32	$109.9\times$
12	0.4 ± 0.2	9.5 ± 26.7	-3.4 ± 9.3	252 ± 39	$685.2\times$

Table I shows the percent change of the objective value, precision, and recall relative to MIXER*, absolute runtime of MIXER in milliseconds, and the relative speedup of MIXER. MIXER is able to achieve a strikingly small optimality gap, while gaining a considerable speedup, with absolute runtimes less than 300 ms. Interestingly, we observe that MIXER converges to local minima that on average lead to better precision and lower recall. This is likely due to these local minima corresponding to associations that are more conservative, i.e., smaller clusters.

Parking Lot Dataset. Experimental data is created by driving a Clearpath Jackal fitted with a Velodyne VLP-32 LiDAR and an Intel RealSense D435i around a parking lot containing 29 cars as shown in Fig. 2. The resulting dataset contains RGB frames with time-synchronized robot pose and LiDAR, along with the pre-determined extrinsic calibrations. Using this dataset, the objective is to associate cars across 100 frames where each view contains noisy, incomplete, and partial detections (i.e., not every car is seen in each frame, and some cars extend out of frame).

1) Feature Extraction: For each RGB frame in the dataset, we use YOLOv3 [56], [57] to detect *2D bounding boxes* associated with the 29 cars. Ground truth associations are generated by manual annotation (see Fig. 2). Three additional simple and complementary features of each car are extracted using this bounding box. A car's *3D centroid* is reconstructed using the median of corresponding LiDAR points that can be reprojected into the 2D bounding box. The dominant *semantic color* (i.e., red, blue, orange, etc.) is also extracted, with grey representing an inconclusive color. Finally, the visual appearance of each car is captured using *SIFT keypoints and*

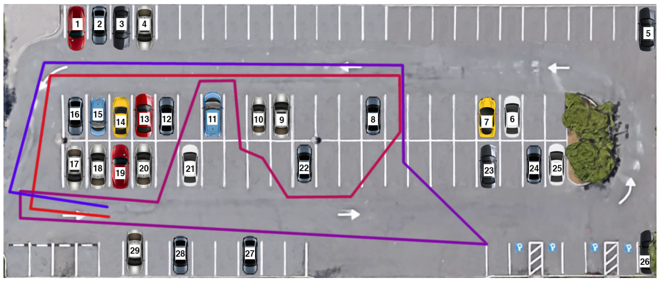

Fig. 2. Illustration of the parking lot environment. Each of the 29 parked cars are located in the parking spots denoted by the numbered car glyph, with approximate color and location. The approximate robot path is shown by the red-to-blue gradient path.

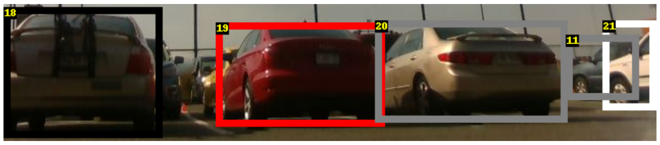

Fig. 3. Each car's bounding box is colored with its extracted semantic color. Harsh lighting in this dataset makes color a weak sensing modality—of the five cars, car 18 is misidentified as black, and cars 20 and 11 are deemed inconclusive (grey). The bounding box number corresponds to the ground truth label.

TABLE II. Precision and recall (reported as P/R) of association algorithms for differing modalities. In each modality column, the algorithm score with the highest F_1 score is bolded. The highest F_1 score overall is underlined.

	bbox	proximity	color	SIFT	combined
All-pairs	0.47/0.41	**0.75/0.50**	0.11/0.71	0.08/1.00	0.08/1.00
Consecutive	0.47/0.41	0.93/0.29	**0.27/0.45**	0.57/0.43	0.68/0.43
Spectral [3]	0.13/0.42	0.15/0.51	0.14/0.45	0.29/0.63	0.35/0.62
MatchLift [31]	0.83/0.24	0.83/0.34	0.34/0.17	0.51/0.83	0.51/0.72
QuickMatch [8]	0.92/0.20	0.97/0.24	0.18/0.32	0.20/0.36	0.29/0.34
CLEAR [9]	0.56/0.21	0.14/0.58	0.13/0.41	0.30/0.75	0.28/0.77
MIXER	**0.88/0.35**	0.85/0.45	0.13/0.35	**0.85/0.62**	**0.88/0.82**

descriptors [58], [59] from each car's bounding box.

2) Computing Association Scores: Each modality's noisy association matrix $A_{ij} \in [0,1]^{m_i \times m_j}$ (i.e., $l = 1$) is constructed for each pair of car detections. This scoring leverages knowledge about the modality; for example, *bounding box intersection-over-union* can give an indication of similarity for temporally consecutive frames, but for non-consecutive frames this method is inconclusive—we set the similarity score to 0.5 in this case. Similarly, similarity based on *spatial proximity* using 3D centroid degrades between temporally distant frames due to odometric drift; therefore, we discount scores such that temporally distant frames yield an inconclusive 0.5 score. On the other hand, *color similarity* between two cars is time-independent and takes on 0 or 1 depending on semantic color matching; if either car color is grey, the score is set to 0.5. Finally, SIFT matching with Lowe's ratio test [59] is used to score *visual similarity* of pairs of cars. Like other modalities, scoring is defined such that weak matching is mapped to the inconclusive 0.5 score.

Each of the aforementioned similarity scores arise from distinct modalities with their own strengths and weaknesses. We create a combined modality by formulating diagonal score matrices S_{ij} with $l = 4$ (see Section II), allowing MIXER to perform multimodal, multiway matching.

3) Evaluation: We use precision (P), recall (R), and F_1-score to compare MIXER against state-of-the-art algorithms. As a baseline, we use an all-pairs association strategy as a naïve multiway matcher, which considers any object pair with an association score greater than 0.5 a match. Similarly, we use a consecutive association strategy which performs this simple thresholding, but does not form associations between temporally non-consecutive detections—a commonly used paradigm in e.g., object tracking. We include the following multiway matching algorithms in our comparison: the Spectral [3] algorithm, extended by Zhou et al. [6]; MatchLift [31] which is based on a convex relaxation with a similar Frobenius objective to MIXER; recent graph clustering algorithms QuickMatch [8] and CLEAR [9].

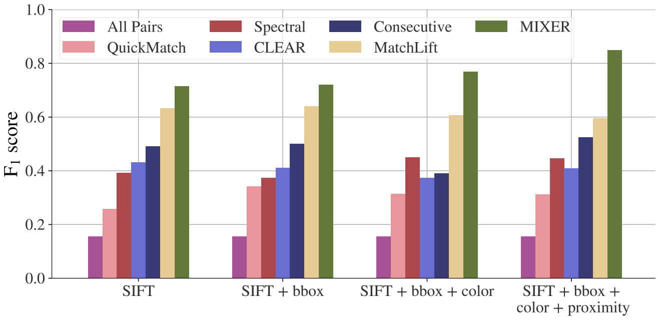

Fig. 4. Algorithm results starting with visual similarity, the strongest modality, and incrementally combining with others. Leveraging the unique information in each modality, MIXER is able to improve upon visual similarity F_1 score by nearly 20%.

Table II lists the P/R results for each association algorithm operating on a given modality. As expected, the color matching modality is the weakest. The bounding box modality also obtains low scores, specifically in R. This highlights the complementary nature of these two modalities—bounding box similarity is inherently frame-to-frame and thus performs well at identifying matches in consecutive frames (↑P), but is unable to cluster associations in a multiway matching sense, resulting in fragmented tracks [60] (↓R). On the other hand, color similarity is time-independent, allowing more non-consecutively seen cars to be matched (↑R), although semantic colors may not be rich enough to distinguish cars of similar color (↓P). This reasoning also explains why the consecutive algorithm scores the highest F_1 score with the color modality—cars of the same semantic color are more likely to be the same car when the observations are made consecutively. Visual similarity and proximity produce the best scores, owing to SIFT's robustness even across non-consecutive frames and the local consistency of consecutive position estimates, even in the presence of global drift. These four modalities are combined and we see that MIXER is capable of recovering associations with high, well-balanced P/R, resulting in an F_1 score of 0.85. Second-best performance is achieved by MatchLift with an F_1 score of 0.60.

The high P/R of MIXER using multiple modalities is enabled by the 0.5 "inconclusive" score. By appropriately designing the scoring function of each modality, MIXER is able to delay association decisions that a single modality may have deemed inconclusive. The benefit is that MIXER *improves* the combined modality associations, whereas other algorithms struggle to appropriately combine information. To show that MIXER improves the results of any one modality, we apply the association algorithms to an aggregate association matrix of incrementally combined modalities and show the resulting F_1 scores in Fig. 4. These results reveal that although the visual similarity modality is strong on its own, MIXER is able to improve it by nearly 20% because of its ability to mix multiple modalities of varying strengths.

References

[1] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[2] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Annals of operations research*, vol. 14, no. 1, pp. 105–123, 1988.

[3] D. Pachauri, R. Kondor, and V. Singh, "Solving the multi-way matching problem by permutation synchronization," in *Advances in neural information processing systems*, 2013, pp. 1860–1868.

[4] E. Maset, F. Arrigoni, and A. Fusiello, "Practical and efficient multi-view matching," in *IEEE International Conference on Computer Vision*, 2017, pp. 4578–4586.

[5] S. Leonardos, X. Zhou, and K. Daniilidis, "Distributed consistent data association via permutation synchronization," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 2645–2652.

[6] X. Zhou, M. Zhu, and K. Daniilidis, "Multi-image matching via fast alternating minimization," in *IEEE International Conference on Computer Vision*, 2015, pp. 4032–4040.

[7] F. Bernard, J. Thunberg, J. Goncalves, and C. Theobalt, "Synchronisation of partial multi-matchings via non-negative factorisations," *arXiv preprint arXiv:1803.06320*, 2018.

[8] R. Tron, X. Zhou, C. Esteves, and K. Daniilidis, "Fast multi-image matching via density-based clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.

[9] K. Fathian, K. Khosoussi, Y. Tian, P. C. Lusk, and J. P. How, "CLEAR: A Consistent Lifting, Embedding, and Alignment Rectification Algorithm for Multiview Data Association," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1686–1703, 2020.

[10] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *Artificial Intelligence*, vol. 293, p. 103448, 2021.

[11] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 523–536, 2019.

[12] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, and S. Satoh, "Beyond intra-modality: A survey of heterogeneous person re-identification," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 4973–4980.

[13] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[14] C. Qin, Y. Zhang, Y. Liu, and G. Lv, "Semantic loop closure detection based on graph matching in multi-objects scenes," *Journal of Visual Communication and Image Representation*, p. 103072, 2021.

[15] R. Burkard, M. Dell'Amico, and S. Martello, "Assignment problems," *SIAM*, 2009.

[16] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[17] T. C. Koopmans and M. Beckmann, "Assignment problems and the location of economic activities," *Econometrica: journal of the Econometric Society*, pp. 53–76, 1957.

[18] E. L. Lawler, "The quadratic assignment problem," *Management science*, vol. 9, no. 4, pp. 586–599, 1963.

[19] E. M. Loiola, N. M. M. de Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido, "A survey for the quadratic assignment problem," *European journal of operational research*, vol. 176, no. 2, pp. 657–690, 2007.

[20] S. Sahni and T. Gonzalez, "P-complete approximation problems," *Journal of the ACM*, vol. 23, no. 3, pp. 555–565, 1976.

[21] M. S. Bazaraa and A. N. Elshafei, "An exact branch-and-bound procedure for the quadratic-assignment problem," *Naval Research Logistics Quarterly*, vol. 26, no. 1, pp. 109–121, 1979.

[22] M. Bazaraa and O. Kirca, "A branch-and-bound-based heuristic for solving the quadratic assignment problem," *Naval research logistics quarterly*, vol. 30, no. 2, pp. 287–304, 1983.

[23] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2005, pp. 1482–1489.

[24] P. Swoboda, C. Rother, H. Abu Alhaija, D. Kainmuller, and B. Savchynskyy, "A study of lagrangean decompositions and dual ascent solvers for graph matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1607–1616.

[25] Q. Zhao, S. E. Karisch, F. Rendl, and H. Wolkowicz, "Semidefinite programming relaxations for the quadratic assignment problem," *Journal of Combinatorial Optimization*, vol. 2, no. 1, pp. 71–109, 1998.

[26] F. Fogel, R. Jenatton, F. Bach, and A. D' Aspremont, "Convex relaxations for permutation problems," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[27] I. Kezurer, S. Z. Kovalsky, R. Basri, and Y. Lipman, "Tight relaxation of quadratic matching," in *Computer Graphics Forum*, vol. 34, no. 5. Wiley Online Library, 2015, pp. 115–128.

[28] F. Bernard, C. Theobalt, and M. Moeller, "Ds*: Tighter lifting-free convex relaxations for quadratic matching problems," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4310–4319.

[29] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1426–1433.

[30] A. Nguyen, M. Ben-Chen, K. Welnicka, Y. Ye, and L. Guibas, "An optimization approach to improving collections of shape maps," in *Computer Graphics Forum*, vol. 30, no. 5. Wiley Online Library, 2011, pp. 1481–1491.

[31] Y. Chen, L. Guibas, and Q. Huang, "Near-optimal joint object matching via convex relaxation," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 32, no. 2, 22–24 Jun 2014, pp. 100–108.

[32] N. Hu, Q. Huang, B. Thibert, and L. J. Guibas, "Distributable consistent multi-object matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2463–2471.

[33] J.-G. Yu, G.-S. Xia, A. Samal, and J. Tian, "Globally consistent correspondence of multiple feature sets using proximal gauss–seidel relaxation," *Pattern Recognition*, vol. 51, pp. 255–267, 2016.

[34] S. Leonardos and K. Daniilidis, "A distributed optimization approach to consistent multiway matching," in *IEEE Conference on Decision and Control*, 2018, pp. 89–96.

[35] S. Leonardos, X. Zhou, and K. Daniilidis, "A low-rank matrix approximation approach to multiway matching with applications in multisensory data association," in *2020 IEEE International Conference on Robotics and Automation*, 2020, pp. 8665–8671.

[36] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 6, pp. 1228–1242, 2016.

[37] Z. Serlin, G. Yang, B. Sookraj, C. Belta, and R. Tron, "Distributed and consistent multi-image feature matching via quickmatch," *The International Journal of Robotics Research*, vol. 39, no. 10-11, pp. 1222–1238, 2020.

[38] J. Yan, Y. Tian, H. Zha, X. Yang, Y. Zhang, and S. M. Chu, "Joint optimization for consistent multiple graph matching," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1649–1656.

[39] X. Shi, H. Ling, W. Hu, J. Xing, and Y. Zhang, "Tensor power iteration for multi-graph matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5062–5070.

[40] P. Swoboda, D. Kainmuller, A. Mokarian, C. Theobalt, and F. Bernard, "A convex relaxation for multi-graph matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[41] H. F. Durrant-Whyte, *Sensor Models and Multisensor Integration*. New York, NY: Springer New York, 1990, pp. 73–89.

[42] H. Durrant-Whyte and T. C. Henderson, *Multisensor Data Fusion*. Cham: Springer International Publishing, 2016, pp. 867–896.

[43] F. Castanedo, "A review of data fusion techniques," *The scientific world journal*, vol. 2013, 2013.

[44] Y. Bar-Shalom and X.-R. Li, "Estimation and tracking- principles, techniques, and software," *Norwood, MA: Artech House, Inc, 1993.*, 1993.

[45] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[46] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[47] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[48] K. Sentz, S. Ferson *et al.*, *Combination of evidence in Dempster-Shafer theory*. Sandia National Laboratories Albuquerque, 2002, vol. 4015.

[49] R. R. Murphy, "Dempster-shafer theory for sensor fusion in autonomous mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 2, pp. 197–206, 1998.

[50] P. Cheeseman, "In defense of probability," in *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'85. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1985, p. 1002–1009.

[51] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[52] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ser. ICML'96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, p. 148–156.

[53] K. Granström, T. B. Schön, J. I. Nieto, and F. T. Ramos, "Learning to close loops from range data," *The international journal of robotics research*, vol. 30, no. 14, pp. 1728–1754, 2011.

[54] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[55] G. O. LLC, "Gurobi optimizer reference manual," 2021. [Online]. Available: http://www.gurobi.com

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[57] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 2018.

[58] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[59] ——, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[60] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.