Use of Scene Geometry Priors for Data Association in Egocentric Views

Huiqin Chen, Emanuel Aldea, Sylvie Le Hégarat-Mascle, Vincent Despiegel

Use of Scene Geometry Priors for Data Association in Egocentric Views

Huiqin Chen
SATIE UMR 8029
Paris-Saclay University
huiqin.chen@u-psud.fr

Emanuel Aldea
SATIE UMR 8029
Paris-Saclay University
emanuel.aldea@u-psud.fr

Sylvie Le Hégarat-Mascle
SATIE UMR 8029
Paris-Saclay University
sylvie.le-hegarat@u-psud.fr

Vincent Despiegel
IDEMIA
vincent.despiegel@idemia.com

***Abstract*—The joint use of dynamic, egocentric view cameras and of traditional overview surveillance cameras in high-risk contexts has become a promising avenue for advancing public safety and security applications, as it provides more accurate localization and finer analysis of individual interactions. However, the strong scene scale changes, occlusions and appearance variations make the egocentric data association more difficult than the standard across-views data association. To address this issue, we propose to use two independent geometric priors and integrate them with the classic appearance cues into the objection function of data association algorithm. Our results show that the proposed method achieves significant improvement in terms of the association accuracy. We highlight the attractive use of geometric priors in across-views data association and its potential for supporting pedestrian tracking in this context.**

***Index Terms*—Across-views data association, Geometry priors, Egocentric views**

I. Introduction

Tracking pedestrians is a paramount task in computer vision, which enables us to identify, analyze and predict human interactions and activities in various contexts. In a multi-person tracking context, data association (DA) becomes a fundamental step, which refers to matching observations related to the same object. A large body of literature is devoted to DA (see for example [1] for a comprehensive survey), since this term encompasses actually multiple procedures. Most of them focus on across-time DA which matches a temporal sequence of observations for single-camera systems. One of the common difficulties for across-time DA is handling the occlusions which occur frequently in single-camera systems. An alternative approach is to employ multi-view systems and to perform across-views DA by matching observations in the views of different cameras.

Recently, egocentric (first-view) visual data became a rich source of information for safety and security applications, ranging from public photos in urban settings to wearable cameras used by law enforcement agents (LEA). The joint use with an overview/hovering camera has the potential to support a more accurate pedestrian localization, as well as a finer analysis of interactions occurring among the visible participants. Although multi-views DA may be performed among three or more cameras, we restrict our work to deal with the problem of data association between an egocentric camera and an overview/hovering camera with overlapping fields of view, as this is the standard scenario occurring in realistic situations.

Despite the challenges raised by across-views association due to strong visual appearance variations, the detection errors are better coped with in the case of *overlapping* fields of view, in which the additional geometric constraint can be provided. Geometry and appearance wise, egocentric DA is more difficult than standard across-views DA due to scene scale changes. However, this context still allows for relating directly visual information between the views. Recently proposed top-to-egocentric view DA strategies are strictly based on the consistency of higher level scene content extracted in individual views [2], [3], but they require video recordings in order to perform the across-views association in the temporal domain, while we can rely only on image pairs.

There are multiple avenues for enforcing the additional geometry consistency on the appearance similarity in DA. Typically, one may extend an across-time global optimization to multiple views [4], [5], or cast the DA as a global energy minimization [6]. The basic idea is to combine into a single objective function two penalty terms, one derived from appearance and one from geometry. However, the quality of the two terms usually suffers from unreliability due to different sources of errors. A fundamental challenge is then represented by how to combine two different terms in presence of imprecision.

In this work, we explore different approaches to use the geometric priors for DA in egocentric views. Specifically, we study how the geometric consistency cues can help improve the classical appearance cues. The benefit of this study is twofold. First, we derive a reliable multi-views DA by estimating 3D locations and spatial relationships among the observed pedestrians. This study is motivated by the S^{2}UCRE project, aiming to highlight perpetrators in body-camera images acquired by LEA, with the support of security camera footage from the same area. Beyond this task, an improvement in DA performance may be further integrated into tracking algorithms for initializing 3D tracks and then generating accurate 3D observations for track extension.

This study was supported by the S^{2}UCRE project, co-funded by the German BMBF grant 13N14463 and by the French ANR grant ANR-16-SEBM-0001

978-1-7281-6232-4/20/$31.00 ©2020 IEEE

II. Data Association

A. Sources of error

Let us summarize below the main sources of error that impact directly the association, in order to understand better how they may be addressed.

Appearance In terms of visual cues, a pedestrian association between an egocentric and an overview camera is hampered by multiple factors which affect also the standard multi-views association, such as the strong pose and illumination variations. Two factors however increase the difficulty of the task, namely the scale variations and the stronger occlusions occurring in the first-view camera. The latter requires that DA be flexible in terms of not associating a potentially high number of targets which are present only in the overview camera.

Geometry The relative pose estimation between the cameras may potentially assist significantly the association by restricting the matching candidates to a small subset of detections in the other view. The challenges of this step are twofold. Firstly, depending on the alignment of the pedestrians with respect to the cameras, the potential match subset may still contain multiple detections - especially in crowded areas. The geometry constraints are thus not meant to solve entirely the DA, but rather to complement an appearance based constraint. Secondly, the pose estimation is built up on identifying invariant low-level visual cues, and this process suffers occasionally from the same limitations mentioned in the previous paragraph. This implies that the relevance of the geometrical constraints depends on the spatial configuration of the cameras and of the dynamic objects in the scene.

B. Performing the optimization

The optimization process may be generally modeled in a similar manner to the temporal data association as a Set Partition Problem, which is NP-complete [7]. However, the two frame assignment problem considered here may be solved exactly in polynomial time by the Kuhn-Munkres, or Hungarian algorithm [8], [9], [10], that we recall in the following paragraphs.

Given two sets of detections from two views, denoted by $P_1 = \{p_i\}_{1 \leqslant i \leqslant M}$ and $P_2 = \{p_j\}_{1 \leqslant j \leqslant N}$, the association is represented by a $M \times N$ matrix of binary elements, denoted as $\{x_{ij}, i \in \{1, \ldots, M\}, j \in \{1, \ldots, N\}\}$, subject to some constraints. As some individuals may be present only in one view, the choice of non association should be taken into account. The association solutions are thus designed as one to at most one, meaning that an individual in a view is associated to maximum one individual in the other view. This leads to the association matrix containing at maximum one non-null value per row and per column. Let c_{ij} denote the association costs (i.e. the measure of dissimilarity) between the i-th individual in the first view and j-th individual in the second view. The optimization problem is then defined as

$$\begin{aligned} &\min \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} x_{ij} \\ &\text{subject to } \sum_{i=1}^{M} x_{ij} \leqslant 1, \forall j \quad \text{and} \quad \sum_{j=1}^{N} x_{ij} \leqslant 1, \forall i. \end{aligned} \tag{1}$$

In order to quantify the benefit of a non association with respect to a costly association, a cost of non association is defined so that whenever an association cost is larger it, the optimization process is encouraged to reject it and to highlight a false association. Consequently, the dimension of the association matrix is extended to $(M+N) \times (M+N)$ with $x_{ij} = 1$, representing a non-association for object i if $j > N$ or for object j if $i > M$. The cost of non association is thus incorporated into the cost matrix $A_{(M+N)\times(M+N)}$ used in the Hungarian algorithm as follows:

$$A_{ij} = \begin{cases} c_{ij} & \text{if } 1 \leqslant i \leqslant M,\ 1 \leqslant j \leq N \\ \gamma & \text{otherwise,} \end{cases} \tag{2}$$

where γ is a threshold representing the non association cost.

The cost function c_{ij} plays a core role on the performance of association. The basic approach of relying on single cues such as the 2D appearance features may be not robust due to the visual ambiguity between individuals. Although it requires a spatial calibration of cameras, the 3D geometric priors relating two views helps overcome visual ambiguity. In the following, we first introduce the costs based on the appearance cues. Different geometric priors are then explored and proposed to be integrated with the appearance cues into the cost functions in order to improve the performance of data association.

III. Association cost functions

A. Appearance based costs

The association costs based on appearance rely on the feature descriptors which are used as a representation for individuals. The cost values are then derived as distances between feature descriptors. These descriptors can be divided into two groups: traditional ones based on color, texture or shape information and learning based representations extracted by a deep neural network.

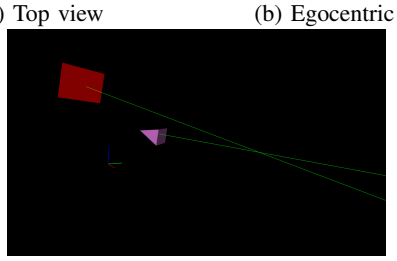
1) Baseline representation: First, let us consider a traditional descriptor based on the color histogram, that we couple with the χ^2 Bin Ratio-Based Distance [11] for the distance measure. Let $\mathbf{h^i} \in \Re^n$ denote a L_2 normalized histogram with n bins used as descriptor of the object i, and $\mathbf{h^j} \in \Re^n$ describing an object j. Then, the χ^2 Bin Ratio-Based Distance between $\mathbf{h^i}$ and $\mathbf{h^j}$ is defined as [11]:

$$d_{hist} = d_{\chi^2} - \frac{1}{2} \left\| \mathbf{h^i} + \mathbf{h^j} \right\|_2^2 \sum_{k=1}^{n} \frac{(h_k^i - h_k^j)^2 h_k^i h_k^j}{(h_k^i + h_k^j)^3} \tag{3}$$

where $d_{\chi^2} = \frac{1}{2} \sum_{k=1}^{n} \frac{(h_k^i - h_k^j)^2}{h_k^j + h_k^j}$.

(a) Top view

(b) Egocentric view



(c) 3D visualization of projection rays.

Fig. 1: Illustration of geometric distance prior. (a) and (b) show a synchronised pair of frames from the static overview camera and the egocentric camera. (c) visualizes the cameras (two polygons) and the projection rays (green lines) which intersect in 3D for a correct association across two views, marked by red rectangle on (a) and (b).

2) Learning approach: Secondly, let us consider a deep convolutional neural network based pedestrian reidentification algorithm. Specifically, similarly to face recognition algorithms as [12], a feature extractor is trained as a classifier on internal pedestrian datasets, the last layer is extracted and pedestrian recognition can be performed on those features using cosine similarity.

B. Geometric costs

We present two different ways to apply the geometric priors for the considered data association problem. One of them is derived from the geometric distance and the other a basic geometric constraint which are presented in the following.

Geometric distance As the rays from two different views for a same object intersect in the 3D location of object, it is possible to discriminate between true and false elementary association based on the geometric distance between rays which pass through the camera center and the object, as shown in Fig. 1. Assuming that the camera parameters are available, namely, the 3D rotation matrix $\mathbf{R} \in \Re^{3\times3}$, the camera center $\mathbf{c} \in \Re^3$ and the intrinsic matrix $\mathbf{K} \in \Re^{3\times3}$, we have the following relationship between image pixel (u, v) and 3D point $\mathbf{q} \in \Re^3$

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{KR}(\mathbf{q}-\mathbf{c}) = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \mathbf{R}(\mathbf{q}-\mathbf{c}). \quad (4)$$

Then, the ray which passes through the camera center $\mathbf{c}$ and object $\mathbf{q}$ can be expressed as

$$r = \mathbf{c} + t(\mathbf{q}-\mathbf{c}) = \mathbf{c} + \lambda t \mathbf{R}^T \begin{bmatrix} \frac{u-c_x}{f_x} \\ \frac{v-c_y}{f_y} \\ 1 \end{bmatrix} = \mathbf{c} + s\mathbf{n}, \quad (5)$$

where $s = \lambda t$ is constant on scale, and $\mathbf{n}$ is the direction vector.

Let $p_i = (u_i, v_i)$ and $p_j = (u_j, v_j)$ denote the pixels on the image plane of two cameras. The distance between their corresponding projection rays r_i and r_j is

$$d_{geo}(r_i, r_j) = \frac{\left| \overrightarrow{\mathbf{c_1 c_2}} \cdot (\mathbf{n}^i \times \mathbf{n}^j) \right|}{\|\mathbf{n}^i \times \mathbf{n}^j\|}, \quad (6)$$

where $\overrightarrow{\mathbf{c_1 c_2}}$ represents the baseline connecting two camera centers, $\mathbf{n}^i$ and $\mathbf{n}^j$ are the direction vector in Eq. (5).

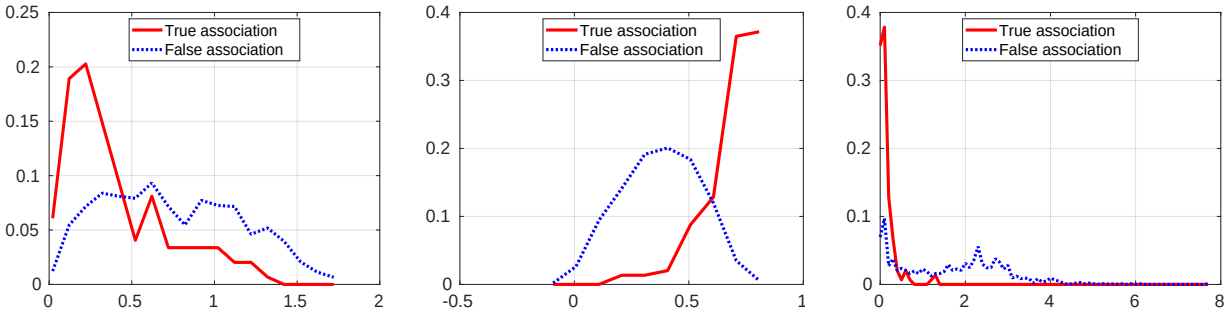
Cheirality constraint A basic constraint provided by the geometric priors is that the detected object should be located in front of both cameras if the association is correct. This property is denoted as *cheirality* [13] in the computer vision community. We note that the cheirality constraint is an additional piece of information, different from the distance between the 3D rays itself, as the latter does not provide any cues about the relative location of the pedestrian with respect to the cameras. In addition, it is independent on the metric scale factor (used to convert distances in metric units), that may be unavailable when the relative pose between the cameras is computed only up to scale. Given the pixels on the image plane of two cameras p_i and p_j, the 3D point $\mathbf{q}$ can be obtained by triangulation[14]. The constraint is then expressed as $z(\mathbf{q}, \mathbf{c_1}) > 0 \wedge z(\mathbf{q}, \mathbf{c_2}) > 0$ where $z(\mathbf{q}, \mathbf{c})$ denotes the depth value of the 3D point $\mathbf{q}$ in the reference system of the camera $\mathbf{c}$ and $\wedge$ stands for the AND logical operator. If the metric scale is available, and due to the uncertainty of the geometry estimations, we tighten the camera front constraint as follows, in order to avoid considering the first-view camera wearer as a potential association:

$$z(\mathbf{q}, \mathbf{c_1}) \geqslant 0.5 \wedge z(\mathbf{q}, \mathbf{c_2}) \geqslant 0.5. \quad (7)$$

This crisp constraint can be easily taken into account by setting the corresponding cost in the association matrix to a value $\beta \gg 1$ when the cheirality constraint does not hold.

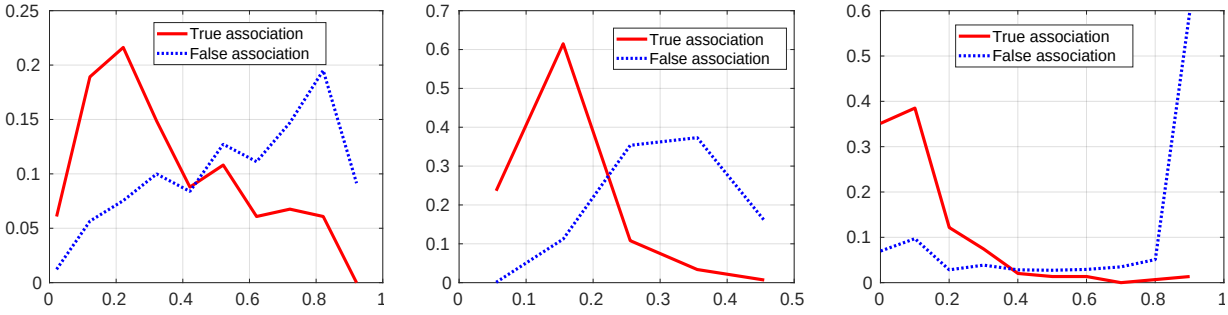
C. Fusion costs

Whenever moving cameras are considered, a data association strategy that relies on a single appearance based cost is often not robust. It is possible to improve the association accuracy by aggregating costs from different cues and relying both on appearance and geometry consistency. The general idea is to combine different costs into a single one that can be directly introduced in the cost matrix. For such a purpose, there exist a large number of combination rules, ranging from the generally applicable ones such as sum, weighted sum, product or min/max, up to domain specific strategies such as belief combination [15]. In multi-views DA as well, the existing literature has underlined the difficulty of combining appearance cues with very simple additional information, such as coarse spatio-temporal cues [16]. For data fusion, it is not straightforward to identify an effective combination without prior knowledge about the characteristics of individual sources. In order to explore an efficient combination of association



(a) Color histogram based distance (b) NN similarity score (c) Geometric distance

Fig. 2: Distribution of true association and false association for different costs before normalization: (a) appearance costs based on color histogram; (b) appearance costs based on learned approach; (c) geometric costs.



(a) Normalized histogram based distance (b) Normalized NN distance (c) Normalized geometric distance

Fig. 3: Distribution of true association and false association for different costs after normalization: (a) appearance costs based on color histogram; (b) appearance costs based on learned approach; (c) geometric costs.

costs, we investigate first the statistical distribution of individual costs for the true and false associations separately.

As the distribution of true associations and false associations for the appearance based costs and the geometric costs are heterogeneous as shown in Fig.2, it is necessary to normalize different costs before combining them. Depending on the considered costs, different normalization techniques are used. Specifically, the appearance based cost using the learning approach is derived from the cosine similarity, called s_{nn}, which ranges in $[-1, 1]$, Min-max normalization is thus applied as suggest in [17]. By subtracting the normalized similarity from 1, we obtain the normalized distance for the appearance based cost using learning approach, denoted by $d'_{nn} = \frac{1}{2}(1 - s_{nn})$. For the distance metric ranging in $[0, +\infty)$, such as the color histogram distance and the geometric distance, we transform them in the range of $[0, 1]$ with a $\tanh$ function, which allows to preserve the influence of small values and reduce that of large values. The normalized distance is then computed as $d'_{hist} = \tanh(d_{hist})$ for the color histogram distance and $d'_{geo} = \tanh(d_{geo})$ for the geometric distance.

The distribution of costs after normalization is shown in Fig. 3. We note that the appearance based costs follow a bell-shaped distribution, while the distribution of the normalized geometric cost exhibits different behaviour. The potential variation of the appearance based distributions requires that the weights in a sum combination $d_{+}^{geo,app} = w_{geo}d'_{geo} + w_{app}d'_{app}$ be tuned accordingly, along with the non-association cost γ. The product combination however is particularly adapted to the profile of the geometric distances. The fusion costs based on the production combination is defined as:

$$d_{*}^{geo,app} = d'_{geo}d'_{app} \tag{8}$$

D. Generality of combination rule

The considered combination rule is intended to be applied to an arbitrary pair of synchronized images, as for example a photo of an event taken by a participant at ground level, along with the corresponding photo from a static overview camera. One may wonder in this case whether the assumptions made to support the cost aggregation strategy in Eq. 8 would be generally valid. Although the distributions presented in Fig. 3 are generated from data acquired in a specific location with the same camera pair, we expect that for various scenes the distances for true/false associations will follow the same families of distributions. Moreover, the product rule does not involve any parameters that require tuning depending on the slightly varying parameters of the distance distributions. The only parameter of the method which depends on the experimental conditions is the non-association cost γ (Eq. 2). The choice of γ allows also for favoring pedestrian associations with the risk of generating false matches (high value of γ) versus enforcing a stricter matching policy with the risk of missing some harder associations (low value of γ).

IV. EXPERIMENTS

A. Testing scenario

We collected an outdoor scene dataset containing 143 pairs of synchronised images containing 1109 pedestrian occurrences, for which the number of the observed persons in each pair of images ranges from 5 to 13. These images are acquired using a smartphone and a GoPro camera in an open outdoor environment. We annotated the ground truth for pedestrian detections and association labels, in order to evaluate the performance of the different association strategies we considered*. The GoPro camera is fixed on the upper floor of a building, while the phone is held by a moving pedestrian. In our experiments, we evaluate the association performance for the geometric costs, appearance based costs, as well as for different combination rules. We also study the influence of the cheirality constraint on the association performance.

Geometric Priors. We compute the intrinsic parameters of the cameras, and evaluate the extrinsic parameters (rotation matrix and center) of the static camera using PnP [18]. To obtain the extrinsic parameters of the mobile camera in the fixed reference system, we first compute relative rotation and translation between the mobile and static cameras [19] and then combine the relative pose with the extrinsic parameters of the static camera. The scene scale is determined by the constraint of mobile camera height with respect to the ground. The height is set to $1.5m$ by considering the mobile camera is held at eye level by a person whose height is $1.60m$, and that the distance from eyes to the top of head is about $0.1m$.

Person detection and association. For each pair of image, we apply separately two different pedestrian detectors. The baseline we consider in this work for extracting the bounding boxes of pedestrians is the widely popular object detector YOLO [20]. The second detector is related to the representation extraction algorithm introduced in Section III-A2.

Regarding the appearance cost term, the baseline, denoted as **Hist**, is a color histogram distance computed as follows. For each bounding box, we consider 32 bins in the HSV color space, where we used 8 bins for partitioning the H channel values, and 2 bins for S and V channels respectively. The cost is computed using the χ^2-BRD histogram distance (Section III). The second appearance cost, denoted as **NN**, is related to the angle between vector representations of the detections computed by the learning algorithm (Section III-A2). For the geometric distance measure, we consider the top-center point of bounding box for projecting it in 3D.

The cost matrix is fed with different distance measures for all detected bounding boxes. Following our analysis of the distribution of correct and incorrect association distances, the non association cost is set to $\gamma = 0.25$ for the sum combination, and to $\gamma = 0.1$ for the product combination respectively. Finally the Hungarian algorithm [8] is applied to this computed cost matrix. The solution provides the optimal association solution.

*Dataset contact form: http://hebergement.u-psud.fr/emi/S2UCRE/

B. Evaluation and results

To determine if the predicted bounding box is true or false, we set the threshold to 0.3 for Intersection over Union (IoU) between the predicted bounding box and the ground truth bounding box. The evaluation for the performance of association is based on the average accuracy over all dataset. For each image pair in dataset, the association accuracy is:

$$\text{Accuracy} = \frac{\text{\# true positives} + \text{\# true negatives}}{\text{\# of unique persons}} \quad (9)$$

In the reported data, the different methods being evaluated are denoted as:

- Geometry only: the association cost is only the normalized geometric distance d'_{geo}
- Appearance only: the association cost is only the appearance based distance, either Hist (d'_{hist}) or NN (d'_{nn})
- Sum combination: the association cost is a weighted sum of geometry and appearance costs ($w_i = 0.5$)
- Product combination: the association cost is a product of geometry and appearance costs

Additionally, for each method we perform the association with and without the cheirality constraint.

One important consideration is that the final result of the DA algorithm is influenced at the same time by the association step, but also by the detection algorithm. The coupling between the two steps may be quite complex: a high precision/low recall detector causes inevitably missed associations, while a high recall/low precision detector may allow for a good final result only if the non-association cost and a good distance matrix help in rejecting the false positive detections. In order to avoid the influence caused by the difference in performance between the detectors, we first evaluate the association accuracy without considering the performance of detection and then evaluate the global detection and association pipeline.

TABLE I: Association accuracy.

Cost Function		**Accuracy**		
Method	Cheirality constraint	YOLO + Hist	Ours + Hist	Ours + NN
Geometry only	No	0.610	0.654	0.654
	Yes	0.851	0.849	0.849
Appearance only	No	0.574	0.537	0.665
	Yes	0.628	0.579	0.820
Sum combination	No	0.724	0.782	0.709
	Yes	**0.864**	**0.895**	0.890
Product combination	No	0.614	0.660	0.707
	Yes	0.839	0.831	**0.896**

Association evaluation We focus first on evaluating the performance of the DA step specifically. To this aim, we consider the default detection thresholds for the two detectors, which allow for a reasonable compromise between precision and recall. For the used dataset, we report that $Pr_{YOLO} = 92.2\%$, $Re_{YOLO} = 72.9\%$, $Pr_{Ours} = 97.6\%$, $Re_{Ours} = 84.3\%$.

The number of unique persons in Eq. 9 is considered as the number of correct detections *provided by the detectors* in

the DA input. This allows for a maximum DA accuracy of 100% independently of the detector performance. The performance of association following the different cost functions is summarised in Table I.

TABLE II: Global method accuracy (the two detectors are tuned for the same recall level)

Cost Function		Accuracy*	
Method	Cheirality constraint	YOLO* + Hist	Ours + NN
Geometry only	No	0.571	0.599
	Yes	0.684	0.742
Appearance only	No	0.360	0.648
	Yes	0.482	0.751
Sum combination	No	0.611	0.709
	Yes	**0.714**	**0.784**
Product combination	No	0.550	0.639
	Yes	0.663	0.779

Global evaluation In the second part of the experiments, we compare the entire detection and association pipeline. By considering this time the number of unique persons in Eq. 9 as the number of correct *ground truth* pedestrians, we account for errors in detection and in association. Otherwise stated, the accuracy level is bounded by the detector recall in this case. Let us denote **Accuracy*** this stricter accuracy measure.

In order to provide a fair comparison, we tune YOLO to have a similar recall as the second detector (i.e. 85%); consequently, YOLO's precision falls down to 49.8% due to the higher rate of false positives. The performance of each of the two entire pipelines is presented in Table II.

C. Discussion of results

The overall results highlight clearly that geometric priors improve significantly the performance of the association based on appearance in two different ways. The first one is by performing the combination rules for geometric cost and appearance based cost, and the second one is the use of the cheirality constraint. Both priors contribute in independent ways to filtering the detection associations, while requiring different preliminary steps for their use (a relative camera pose estimation in the case of the cheirality constraint, and an additional PnP+scale estimation for the 3D distance prior).

Regarding the combination rules, the performance improvement varies depending on the specific feature extractors. If the appearance based distance provides a good separability, as in the case of the NN descriptor, then the use of the product aggregation is effective and avoids further parameter tuning. In the case of a less discriminative distance, such as the histogram-based distance in our case, the sum rule may provide a better performance while requiring at the same time a more careful tuning of the cost weights and non-association cost. Finally, Table II presents the global performance of a detection-association pipeline in which the accuracy measure is affected by the two steps in different ways. In this case too, the experiments underline the significant positive impact in all situations of the additional geometry information.

V. Conclusion

This work explores the integration of geometric priors in pedestrian localization problem formulated as multi-views data association. Our work evaluates the performance of DA independently, and also by considering the detection step performance as well in a complete system. We study the expected distribution of the different appearance-based distances and of the proposed geometric distance, which should serve as a guideline for adapting this algorithm in a specific context. Further works will consider more specific combination rules and the integration of additional sensor information (e.g. IMU, GPS), as well as the use of our DA method in a tracking algorithm in case of video data processing.

References

[1] M. Betke and Z. Wu, "Data association for multi-object visual tracking," Synthesis Lectures on Computer Vision, vol. 6, no. 2, pp. 1–120, 2016.

[2] S. Ardeshir and A. Borji, "Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment," in ECCV, 2018, pp. 285–300.

[3] ——, "Egocentric meets top-view," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 6, pp. 1353–1366, 2018.

[4] Z. Cai, S. Hu, Y. Shi, Q. Wang, and D. Zhang, "Multiple human tracking based on distributed collaborative cameras," Multimedia Tools and Applications, vol. 76, no. 2, pp. 1941–1957, 2017.

[5] L. Wen, Z. Lei, M.-C. Chang, H. Qi, and S. Lyu, "Multi-camera multi-target tracking with space-time-view hyper-graph," International Journal of Computer Vision, vol. 122, no. 2, pp. 313–333, 2017.

[6] N. Pellicanò, E. Aldea, and S. Le Hégarat-Mascle, "Geometry-based multiple camera head detection in dense crowds," in BMVC - 5th Activity Monitoring by Multiple Distributed Sensing Workshop, 2017.

[7] R. T. Collins, "Multitarget data association with higher-order motion models," in CVPR. IEEE, 2012, pp. 1744–1751.

[8] 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.

[9] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," Communications of the ACM, vol. 14, no. 12, pp. 802–804, 1971.

[10] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," Computing, vol. 38, no. 4, pp. 325–340, 1987.

[11] W. Hu, N. Xie, R. Hu, H. Ling, Q. Chen, S. Yan, and S. Maybank, "Bin ratio-based histogram distances and their application to image classification," TPAMI, vol. 36, no. 12, pp. 2338–2352, 2014.

[12] A. Hasnat, J. Bohne, J. Milgram, S. Gentric, and L. Chen, "Deepvisage: Making face recognition simple yet with powerful generalization skills," in ICCV Workshops, Oct 2017.

[13] R. I. Hartley, "Cheirality invariants," in Proc. DARPA Image Understanding Workshop, 1993, pp. 745–753.

[14] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge university press, 2003.

[15] T. Denoeux, N. El Zoghby, V. Cherfaoui, and A. Jouglet, "Optimal object association in the Dempster-Shafer framework," IEEE Trans. on Cybernetics, vol. 44, no. 22, pp. 2521–2531, 2014.

[16] W. Jiuqing, C. Xu, B. Shaocong, and L. Li, "Distributed data association in smart camera network via dual decomposition," Information Fusion, vol. 39, pp. 120–138, 2018.

[17] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," Pattern recognition, vol. 38, no. 12, pp. 2270–2285, 2005.

[18] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," TPAMI, vol. 25, no. 8, pp. 930–943, 2003.

[19] H. Chen, E. Aldea, and S. L. Hégarat-Mascle, "Determining epipole location integrity by multimodal sampling," in AVSS, 2019, pp. 1–8.

[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.